HIGH-DIMENSIONAL PROBLEMS IN PROBABILITY, OPTIMIZATION, AND LEARNING

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF MATHEMATICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Mark Sellke
August 2022

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Andrea Montanari, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Sebastien Bubeck, Co-Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Persi Diaconis

Approved for the Stanford University Committee on Graduate Studies.

Stacey F. Bent, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format.

Preface

This thesis concerns several problems in probability, optimization, and machine learning.

In the first part we study mixing and sampling. We begin by revisiting the riffle shuffle, and generalizing the classical "seven shuffles suffice" result of Bayer and Diaconis to shuffles with asymmetric cuts. Our guiding perspective is to pinpoint at what locations and scales the original deck ordering retains stability. We then turn to our first spin glass model, aiming to sample from the Sherrington-Kirkpatrick Gibbs measure in the *high-temperature* phase. We develop a new approach based not on a Markov Chain, but instead on Eldan's stochastic localization. Moreover we prove that no *stable* algorithm can satisfy the same properties once replica symmetry breaks due to the phenomenon of *disorder chaos*.

In the second part we turn to optimization, aiming to find approximate ground states in spin glass models. This problem is intimately related to their *low temperature* behavior, and the limiting ground state energy is given by the Parisi formula at zero temperature. We determine an exact algorithmic threshold for a natural class of stable algorithms, which is achieved by approximate message passing algorithms. The broader class of algorithms is defined by its Lipschitz dependence on the random coefficients of the function to be optimized; it includes general gradient-based algorithms and Langevin dynamics on dimension-free time scales. Our hardness results stem from a refined landscape property that we christen the branching overlap gap property.

The third part concerns two problems in high-dimensional machine learning. We first study the problem of chasing convex bodies, in which one aims to perform stable convex optimization to obtain robust performance guarantees in changing environments. The solution involves a generalization of the classical Steiner point in convex geometry and its connections to Lipschitz selection. Finally we establish the *law of robustness*, which states that a natural robustness memorization task in high dimension requires extremely overparametrized machine learning models.

Acknowledgments

I would first like to thank my advisers, Andrea Montanari and Sébastien Bubeck, for their guidance and inspiration throughout my Ph.D. Your research styles and tastes are quite different and I am lucky to have learned from you both.

I have enjoyed learning from and interacting with many peers, collaborators, mentors, and friends during my Ph.D. including Ahmed El Alaoui, Ryan Alweiss, Thomas Budzinski, Michael Celentano, Kabir Chandrasekher, Yining Chen, Timothy Chu, Christian Coester, Sean Cotner, Persi Diaconis, Ronen Eldan, Zhou Fan, Ofer Grossman, Meghal Gupta, Xiaoyu He, Brice Huang, Tom Hutchcroft, Zach Izzo, Bo'az Klartag, Yin Tat Lee, Ray Li, Yuanzhi Li, David Benjamin Lim, Allen Liu, Yang P. Liu, Haipeng Luo, Song Mei, Dan Mikulincer, Richard Peng, Yuval Rabani, Ashwin Sah, Mehtaab Sawhney, Alex Slivkins, Nike Sun, Yi Sun, Kevin Tian, Chen-Yu Wei, Joseph Woo, and Alex Zhai.

Many people were helpful to my mathematical development prior to graduate school. I thank Scott Sheffield, David Jerison, Ankur Moitra, Dongkwan Kim, Zuming Feng, Po-Shen Loh, Doug Klumpe, Henjin Chi, and Robert Fischer for their mentorship and guidance during college, high school, and middle school. Finally, a special thanks to the lovely Xiaoyue Gong and to my parents and family for their support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This introduction surveys the results in this thesis. Its aim is to convey at least roughly the statements and motivations for these results. To this end we have given a brief mention to certain topics when a more complete and not-overly-long explanation eluded us. The results themselves concern several quite different problems ranging from mixing times and spin glasses to machine learning. However common themes such as dimension-free behavior and sharp transitions are present throughout.

1.1 Cutoff for the Asymmetric Riffle Shuffle

In Chapter 2, we begin with the classical problem of riffle shuffling. In additional to being a ubiquitous procedure in real life, the riffle shuffle has led to beautiful mathematics. While one can ask many questions about such a process, the best studied is inarguably:

How many shuffles are needed to randomize the order of the deck?

Repeated riffle shuffling defines a Markov chain, because the distribution for the deck order X_{t+1} at time $t+1$ given the order X_t at time t is independent of the past. Of course, each X_t lives in the symmetric group $\mathfrak{S}_N$, where N is the number of cards in the deck. We assume the deck starts in a deterministic order X_0 (all choices of X_0 are equivalent by relabelling the cards). We would like to know how large t should be as a function of N for the distribution $\mu_{N,t}$ to become close to the uniform distribution $\mu_{N,\infty}$ on $\mathfrak{S}_N$. The Markov chain is said to *mix* once this occurs, though of course the amount of time required might depend on the precise notion of distance used.

As has become customary, we focus on mixing in total variation. Recall that the total variation

Figure 1.1: A riffle shuffle in progress.

distance $d_N^{\mathrm{TV}}(\mu,\nu)$ between two probability measures on the same space is defined by

$$d_N^{\mathrm{TV}}(\mu,\nu) = \sup_A |\mu(A) - \nu(A)|$$

where the supremum is taken over all measurable sets A. (In our case, $A \subseteq \mathfrak{S}_N$ would be a subset of the possible $N!$ permutations.) The total variation distance is a stringent notion. Indeed, a small total variation distance $d_N^{\mathrm{TV}}(\mu,\nu) \leq \varepsilon$ is equivalent to the existence of a coupling (x,y) such that $x \sim \mu$, $y \sim \nu$, and $\mathbb{P}[x = y] \geq 1 - \varepsilon$.

The total variation mixing time of the standard riffle shuffle was analyzed in [BD92], where it was shown that $\left(\frac{3}{2\log(2)} \pm o(1)\right)\log(N)$ shuffles are necessary and sufficient to mix an N card deck. [BD92] focused on the Gilbert-Shannon-Reeds (GSR) model of the riffle shuffle. In this model the N-card deck is first cut into parts of size A and $N - A$, for $A \sim \mathtt{Bin}(N, 1/2)$ drawn from a binomial distribution. In particular the deck is cut "roughly in half". Next, the cards are "riffled" together by generating a uniformly random interleaving of the two piles from the $\binom{N}{A}$ choices.

As above, given an arbitrary deterministic initial ordering for the cards, let $\mu_{N,K}$ denote the distribution for the state of the deck after K shuffles and $\mu_{N,\infty}$ the uniform distribution on all $N!$ permutations. Then [BD92] showed the following result.

Theorem 1 ([BD92])**.** *Fix $\varepsilon > 0$. If the sequence $(K_N)_{N\geq 1}$ satisfies $K_N \leq \left(\frac{3}{2\log(2)} - \varepsilon\right)\log(N)$,*

then the GSR shuffle satisfies

$$\lim_{N\to\infty} d_N^{\mathrm{TV}}(\mu_{N,K_N}, \mu_{N,\infty}) = 1.$$

On the other hand, if $K_N \geq \left(\frac{3}{2\log(2)} + \varepsilon\right)\log(N)$*, then*

$$\lim_{N\to\infty} d_N^{\mathrm{TV}}(\mu_{N,K_N}, \mu_{N,\infty}) = 0.$$

The above result establishes a sharp threshold at $\frac{3\log N}{2\log(2)}$ shuffles. At this time, the deck quickly transitions from unmixed to fully mixed. This behavior is known as "cutoff" and is surprisingly common in Markov chain theory. Setting $N = 52$ in Theorem 1 led to the moniker "seven shuffles suffice". In fact [BD92] showed even more precisely that cutoff occurs within a constant size window $\frac{3\log N}{2\log(2)} \pm O(1)$.

Our new contribution in Chapter 2, based on [Sel22], is to analyze an asymmetric generalization in which $A \sim \mathtt{Bin}(N, p)$ for general $p \in (0, 1)$. In this p-shuffle model, the riffling is identical to the GSR shuffle with $p = 1/2$, but the cuts are biased by p. This seemingly innocent modification destroys symmetry properties used in [BD92] and requires a completely new analysis. We establish cutoff at $(\overline{C}_p \pm o(1))\log N$ shuffles for an explicit constant $\overline{C}_p$ (see Figure 1.2). As expected, $\overline{C}_p$ is symmetric and minimized at $p = 1/2$. Thus, asymmetry can only slow mixing.

Some foundational ideas for our work, including most of lower bound on the mixing time, were previously introduced in [Lal00]. Moreover (as with Theorem 1), our result below extends to "multinomial" shuffles in which the deck is cut into more than 2 parts. We denote by $\mu_{N,K,p}$ the distribution of the deck after p-shuffling K times.

Theorem 2 ([Sel22])**.** *There exists a constant* $\overline{C}_p$ *such that the following holds. Fix* $\varepsilon > 0$*. If the sequence* $(K_N)_{N\geq 1}$ *satisfies* $K_N \leq \left(\overline{C}_p - \varepsilon\right)\log(N)$*, then the p-shuffle satisfies*

$$\lim_{N\to\infty} d_N^{\mathrm{TV}}(\mu_{N,K_N,p}, \mu_{N,\infty}) = 1.$$

On the other hand, if $K_N \geq \left(\overline{C}_p + \varepsilon\right)\log(N)$*, then*

$$\lim_{N\to\infty} d_N^{\mathrm{TV}}(\mu_{N,K_N,p}, \mu_{N,\infty}) = 0.$$

Prior to our work [Sel22], several interesting results were known for the asymmetric riffle shuffle. [ADS12] determined the mixing time in the more stringent separation and L^∞ senses with $O(1)$ cutoff window. The papers [Sta01, BD98b, BHR99] established connections to quasisymmetric functions and hyperplane arrangements, showing that the eigenvalues of the chain are all real despite its irreversibility. While these papers rely on exact identities, our approach does not and instead directly analyzes the time t distribution of the chain.

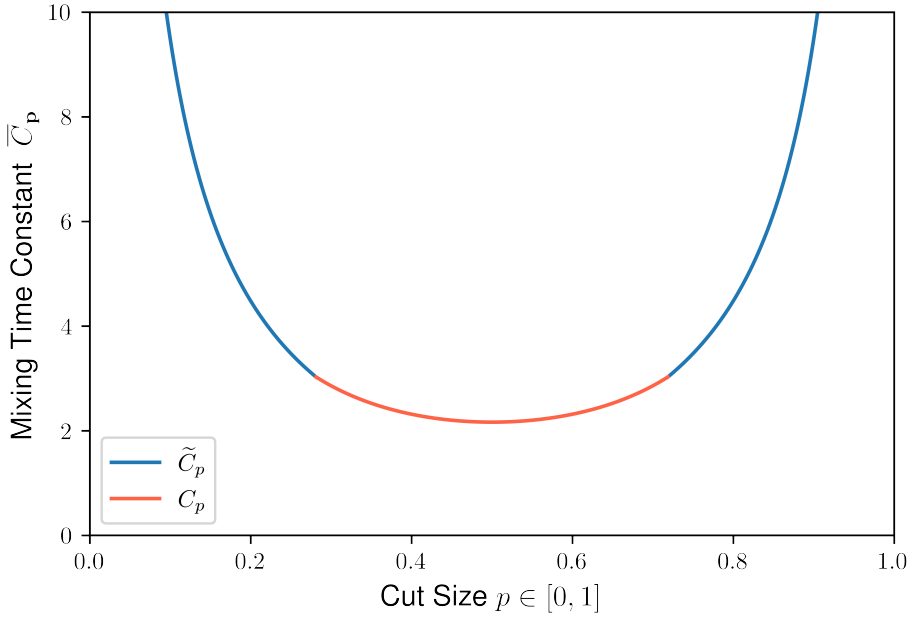


Figure 1.2: The values $\overline{C}_p$ are shown. The color-change points indicate non-smoothness of $\overline{C}_p$ at approximately $p \approx 0.28$ and $p \approx 0.72$.

Other aspects of riffle shuffles are surveyed in [Dia03], and several other choices for the interleaving process have been studied [Tho73, Mor09, Mor13, JM15]. Many seeming basic open problems remain, for example to generalize these results to non-binomial cuts.

Of course, the subject of Markov chain mixing extends far beyond riffle shuffling. First, many other interesting walks on the symmetric group have been thoroughly studied including random transpositions [DS81, Sch05], adjacent transpositions [Lac16] and random-to-random [BN19]. Other important distributions that can be efficiently sampled from via Markov chains include graph colorings [DG98, CDM$^+$19], the hardcore model, uniform spanning trees [Bro89, Ald90, Wil96, ALGV19], contingency tables [DG95], and log-concave measures [FKP94, SL19, LST20]. Several other randomized estimation algorithms employ Markov chain sampling as a fundamental primitive. For example, prominent algorithmis to estimate the volume of a convex body [LS93, LV06, JLLV21] or the number of perfect matchings in a graph [JS89] use such a strategy.

In the case of symmetric group, it is of course easy to sample a uniformly random permutation π directly by choosing $\pi(1), \pi(2), \ldots, \pi(N)$ sequentially and enforcing distinctness throughout. However for many examples mentioned above, a rapidly mixing Markov chain is indispensable for efficient sampling. For example as explained in [DV13], Markov chain sampling enables randomized algorithms to approximate the volume of a d-dimensional convex body $\mathcal{K}$ by evaluating the function

$1_{x\in\mathcal{K}}$ a polynomial (in d) number of times. This task requires exponentially many queries for a deterministic algorithm.

1.2 Algorithmic Stochastic Localization for the Sherrington-Kirkpatrick Model

Chapter 3 is the first of three chapters on spin glasses and is based on joint work [AMS22] with Ahmed El Alaoui and Andrea Montanari. We begin with the Sherrington-Kirkpatrick model, the probability measure on the Boolean cube $\{-1,+1\}^N$ defined for an inverse temperature parameter $\beta > 0$ by

$$\mu_{\beta,A}(\sigma) = e^{\frac{\beta}{2}\langle\sigma, A\sigma\rangle}\mathrm{d}\sigma.$$

We sample the $N \times N$ symmetric matrix A from the Gaussian Orthogonal Ensemble, which means A has independent entries (except that $A_{ij} = A_{ji}$) with distribution

$$\begin{aligned} A_{ii} &\sim \mathcal{N}(0, 2/N); \\ A_{ij} &\sim \mathcal{N}(0, 1/N), \quad i \neq j. \end{aligned}$$

The Sherrington-Kirkpatrick (SK) model was introduced in [SK75] to understand diluted magnetic materials such as CuMn and ZnO. It is a *mean-field* model because it ignores the 3-dimensional structure of physical space in favor of greater mathematical tractability. Indeed to view this model as a physical system, one considers N atoms such that the entry A_{ij} describes the interaction between atoms i and j; thus the SK model describes a "fully connected" system. In Chapter 3 we consider the high-temperature regime of the SK model with β small. We focus on the problem of sampling from this Gibbs measure using an efficient algorithm, e.g. one requiring time growing at most polynomially in the dimension N.

The Sherrington-Kirkpatrick model is known to exhibit a phase transition at $\beta = 1$. For $\beta < 1$, the behavior is known to be "replica symmetric" while for $\beta > 1$ replica symmetry breaking begins. This phase transition has a number of interpretations, but perhaps the simplest is that when $\beta < 1$, an i.i.d. pair $\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2 \sim \mu_{\beta,A}$ of Gibbs samples satisfy

$$\lim_{N\to\infty} \mathbb{P}\left[\left|\frac{\langle\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2\rangle}{N}\right| \leq \varepsilon\right] = 1$$

for any constant $\varepsilon > 0$. Here the probability is taken over the randomness of A as well as $(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$. However once replica symmetry breaks, the distribution of the *overlap* $\frac{\langle\boldsymbol{\sigma}_1,\boldsymbol{\sigma}_2\rangle}{N}$ becomes highly nontrivial and the Gibbs measure exhibits certain "hierarchical clustering" behavior.

In accordance with this, it is natural to believe that simple Markov chains should mix quickly in the replica-symmetric phase $\beta < 1$. The most prominent such chain is the *Glauber dynamics*: from deterministic initialization $\boldsymbol{x}^0 \in \{-1, +1\}^N$, one repeatedly chooses a uniformly random index $i \in [N]$ and resamples the coordinate $\boldsymbol{x}_i^t$ from the desired Gibbs measure conditioned on the other $N-1$ coordinates of $\boldsymbol{x}^t$. Unfortunately, classical techniques to upper bound the mixing time of Glauber dynamics such as the Dobrushin condition [AH87] only imply polynomial mixing at extremely high temerature $\beta^{-1} \geq \Omega(N^{1/2})$.

Rigorous progress on this problem has been achieved only recently. It was shown in [AJK$^+$21] that for $\beta < 1/4$, the Glauber dynamics mixes in $O(N \log N)$ time based on a modified log-Sobolev inequality. This result followed [EKZ21] which showed an $O(N^2)$ mixing time also for $\beta < 1/4$ by establishing a spectral gap. Conversely it is shown in [BAJ18] that mixing is exponentially slow in spin glass models satisfying a certain overlap gap condition (which is actually not expected to hold for the SK model).

We take a very different approach to efficient sampling based on Eldan's stochastic localization. This idea was introduced in [Eld13] and further developed in many works including [Eld16, LV17, LV18a, Eld20, Che21, KP21], see also the ICM survey [Eld22]. Following the approach of [EAM22], the stochastic localization process can be described as follows. Let μ be a compactly supported probability measure on $\mathbb{R}^N$, and let B_t be a standard $\mathbb{R}^N$-valued Brownian motion. Consider the diffusion defined by

$$\mathrm{d}X_t = x\mathrm{d}t + \mathrm{d}B_t, \quad X_0 = 0,\ x \sim \mu. \tag{1.2.1}$$

One may view X_t as a way to gradually reveal the point x. Indeed x is almost surely determined by the full path $(X_t)_{t\in[0,\infty)}$ since the law of large numbers implies that

$$\mathbb{P}\left[\lim_{t\to\infty} \frac{X_t}{t} = x\right] = 1. \tag{1.2.2}$$

At a finite time t, the conditional law of x given X_t is a random measure μ_t. Moreover the probability $\mu_t(S)$ is a martingale for any Borel set S, and thus $(\mu_t)_{t\geq 0}$ can be said to define a martingale on the space of probability measures on $\mathbb{R}^N$.

A key and non-obvious property is that the measures μ_t *localize*. For instance at time $t = \infty$, (1.2.2) implies that one has $\mu_\infty = \delta_x$ for $\mu_\infty = \lim_{t\to\infty} \mu_t$. This suggests that for large finite t, the measures μ_t should concentrate tightly. In fact, it is possible to make this quantitative and prove that the expected covariance matrix of μ_t is bounded above by $\frac{1}{t} \cdot I_N$.

Our approach to efficient sampling in Chapter 3 is to simulate (1.2.1) for a large constant amount of time, and then round the mean of μ_t to a corner in $\{-1, +1\}^N$. We show that this procedure succeeds in sampling from a probability distribution $\mu_{\beta,A}^{\mathrm{alg}}$ with $o(1)$ (normalized) Wasserstein distance $W_{2,N}$ from the true Gibbs distribution $\mu_{\beta,A}$. This means there exists a coupling (x, y) with

marginals $\mu_{\beta,A}$ and $\mu^{\rm alg}_{\beta,A}$ such that

$$\frac{1}{N}\mathbb{E}[\|x-y\|_2^2] \le o_{N\to\infty}(1).$$

A small Wasserstein distance is of course significantly weaker than a small variation distance. On the other hand, our algorithm succeeds in sampling for a larger range of temperatures $\beta < 1/2$ than the previous state of the art $\beta < 1/4$.

Theorem 3. *For any $\varepsilon > 0$ and $\beta < 1/2$ there exists a samping algorithm which inputs a random matrix $\boldsymbol{A}$ and outputs a random point $\boldsymbol{x}^{\rm alg} \in \{-1,+1\}^n$ with law $\mu^{\rm alg}_{\boldsymbol{A}}$ such that with probability $1 - o_N(1)$ over $\boldsymbol{A} \sim \mathsf{GOE}(N)$,*

$$W_{2,N}(\mu^{\rm alg}_{\boldsymbol{A}}, \mu_{\boldsymbol{A}}) \le \varepsilon\,. \tag{1.2.3}$$

The total complexity of this algorithm is $O(N^2)$.

The first key point is that the annealed law of X_t (i.e. averaged over x) is described by

$$\mathrm{d}X_t = \mathbb{E}^{x\sim\mu_t}[x]\mathrm{d}t + \mathrm{d}B_t, \quad X_0 = 0. \tag{1.2.4}$$

In other words, one needs only to repeatedly compute the *mean* of μ_t rather than understand the full high-dimensional distribution all at once. The approximate computation of $\mathbb{E}^{x\sim\mu_t}[x]$ and discretization analysis of (1.2.4) are both nontrivial but fall into the wheelhouse of high-dimensional statistics. They are achieved by combining a contiguity argument, an approximate message passing algorithm, and local convexity of the so-called TAP free energy.

Finally we complement Theorem 3 with an impossibility result. The sampling algorithm used in Theorem 3 is *stable* in the sense that (roughly speaking) it returns a similar output if the inverse temperature β and/or the matrix A are slightly perturbed. We show that such a stability property cannot hold for any sampling algorithm once $\beta > 1$ by an application of Chatterjee's theorem on disorder chaos. Conversely Chatterjee's theorem combined with the stability of our algorithm implies the following purely mathematical result for the true Gibbs measure $\mu_{\boldsymbol{A},\beta}$. Let $\boldsymbol{A}_s = \sqrt{1-s^2}\boldsymbol{A} + s\boldsymbol{A}'$ for $\boldsymbol{A}' \sim \mathsf{GOE}(N)$ independent of $\boldsymbol{A}$. Hence $\boldsymbol{A}_s$ is another GOE matrix correlated with $\boldsymbol{A}$, and is very close to $\boldsymbol{A}$ for small $\boldsymbol{A}_s$. Finally let p-lim denote a limit in probability.

Theorem 4. *Fix $\beta < 1/2$. With high probability Gibbs measure $\mu_{\boldsymbol{A},\beta}$ is Wasserstein-close to small perturbations $\mu_{\boldsymbol{A},\beta'}$ and $\mu_{\boldsymbol{A}_s,\beta}$ in the sense that*

1. $\lim_{s\to 0}\,\text{p-lim}_{N\to\infty} W_{2,N}(\mu_{\boldsymbol{A},\beta}, \mu_{\boldsymbol{A}_s,\beta}) = 0.$
2. $\lim_{\beta'\to\beta}\,\text{p-lim}_{N\to\infty} W_{2,N}(\mu_{\boldsymbol{A},\beta}, \mu_{\boldsymbol{A},\beta'}) = 0.$

In fact our use for Chatterjee's result is essentially to show that Theorem 4 fails when $\beta > 1$.

We believe that Theorem 4 holds for all $\beta < 1$, and that an improved analysis of our algorithm can be used to establish this.

1.3 Optimizing Mean-Field Spin Glasses: Background

In Chapter 4 we turn from sampling to optimization, again in the context of spin glasses. The results of this chapter are based on [AMS21] and [Sel21b]. The former is joint work with Ahmed El Alaoui and Andrea Montanari, and the latter is a follow-up work.

Here we consider the more general even mixed p-spin models. For each $p \in 2\mathbb{N}$, let $\mathbf{G}^{(p)} \in \left(\mathbb{R}^N\right)^{\otimes p}$ be an independent p-tensor with i.i.d. $\mathcal{N}(0,1)$ entries. Fix a sequence $(\gamma_p)_{p\in 2\mathbb{N}}$ with $\gamma_p \geq 0$ and $\sum_{p\in 2\mathbb{N}} 2^p \gamma_p^2 < \infty$. The mixed even p-spin Hamiltonian H_N is defined by

$$H_N(\boldsymbol{\sigma}) = \sum_{p\in 2\mathbb{N}} \frac{\gamma_p}{N^{(p-1)/2}} \langle \mathbf{G}^{(p)}, \boldsymbol{\sigma}^{\otimes p} \rangle.$$

We consider inputs $\boldsymbol{\sigma}$ in either the sphere $S_N = \{\boldsymbol{\sigma} \in \mathbb{R}^N : \sum_{i=1}^N \sigma_i^2 = N\}$ or the cube $\Sigma_N = \{-1,1\}^N$. These define, respectively, the *spherical* and *Ising* mixed p-spin glass models. The coefficients γ_p are customarily encoded in the *mixture function* $\xi(x) = \sum_{p\in 2\mathbb{N}} \gamma_p^2 x^p$. Note that $\widetilde{H}_N$ is equivalently described as the Gaussian process with covariance

$$\mathbb{E}\, \widetilde{H}_N(\boldsymbol{\sigma}^1)\widetilde{H}_N(\boldsymbol{\sigma}^2) = N\xi(\langle \boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2 \rangle / N).$$

For example the SK model discussed above is an Ising spin glass, in which $\gamma_2 = 1/2$ and $\gamma_k = 0$ for $k > 2$.

Our purpose in Chapter 4 is to shed light on a discrepancy between the asymptotic maximum values

$$\mathsf{OPT}^{\mathrm{Sp}} = \mathsf{OPT}^{\mathrm{Sp}}_{\xi} = \operatorname*{p\text{-}lim}_{N\to\infty} \frac{1}{N} \max_{\boldsymbol{\sigma}\in S_N} H_N(\boldsymbol{\sigma}), \qquad \mathsf{OPT}^{\mathrm{Is}} = \mathsf{OPT}^{\mathrm{Is}}_{\xi} = \operatorname*{p\text{-}lim}_{N\to\infty} \frac{1}{N} \max_{\boldsymbol{\sigma}\in \Sigma_N} H_N(\boldsymbol{\sigma})$$

and the maximum *efficiently computable* values of H_N over the same sets.

The values $\mathsf{OPT}^{\mathrm{Sp}}$ and $\mathsf{OPT}^{\mathrm{Is}}$ are given by the celebrated Parisi formula [Par79] which was proved for even models by [Tal06d, Tal06a] and in more generality by [Pan14]. While most often stated as a formula for the limiting free energy at inverse temperature β, the asymptotic maximum can be recovered as a $\beta \to \infty$ limit of the Parisi formula. Restricting for concreteness to the Ising case (we will state the analogous result for the spherical case in Section 5.2), the result can be expressed in the following form due to Auffinger and Chen [AC17b].

Define the function space

$$\mathscr{U} = \left\{ \zeta : [0,1) \to \mathbb{R}_{\geq 0} : \zeta \text{ is right-continuous and nondecreasing}, \int_0^1 \zeta(t)\mathrm{d}t < \infty \right\}. \tag{1.3.1}$$

For $\zeta \in \mathscr{U}$, define $\Phi_\zeta : [0,1] \times \mathbb{R} \to \mathbb{R}$ to be the solution of the following *Parisi PDE*.

$$\partial_t \Phi_\zeta(t,x) + \frac{1}{2}\xi''(t)\left(\partial_{xx}\Phi_\zeta(t,x) + \zeta(t)(\partial_x \Phi_\zeta(t,x))^2\right) = 0 \tag{1.3.2}$$

$$\Phi_\zeta(1,x) = |x|. \tag{1.3.3}$$

Existence and uniqueness properties for this PDE are well established and are reviewed in Subsection 5.6.1. The Parisi functional $\mathsf{P}^{\mathrm{Is}} = \mathsf{P}^{\mathrm{Is}}_{\xi,h} : \mathscr{U} \to \mathbb{R}$ is given by

$$\mathsf{P}^{\mathrm{Is}}(\zeta) = \Phi_\zeta(0,0) - \frac{1}{2}\int_0^1 t\xi''(t)\zeta(t)\ \mathrm{d}t. \tag{1.3.4}$$

Theorem 5 ([AC17b, CHL18])**.** *The following identity holds.*

$$\mathsf{OPT}^{\mathrm{Is}} = \inf_{\zeta \in \mathscr{U}} \mathsf{P}^{\mathrm{Is}}(\zeta). \tag{1.3.5}$$

Moreover the infimum is achieved at a unique $\zeta_* \in \mathscr{U}$

The minimizer $\zeta_* \in \mathscr{U}$ can be obtained as an appropriately renormalized zero-temperature limit of the corresponding minimizers in the positive temperature Parisi formula. These positive temperature minimizers roughly correspond to cumulative distribution functions for the overlap $\langle \boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2 \rangle / N$ of two independent samples from the Gibbs measure $e^{\beta H_N}/Z_N(\beta)$; this is why the functions ζ considered in Theorem 5 are nondecreasing.

Efficient algorithms to find an input $\boldsymbol{\sigma}$ achieving a large objective have recently emerged in a line of work initiated by [Sub21]. We reproduce a lightly modified version of his marvelous optimization algorithm for the spherical setting below.

The existence of a suitable $\boldsymbol{v}_k$ comes from the fact that for $\|\boldsymbol{\sigma}\|_2^2 = q$, the Hessian $\nabla^2 H_N(\boldsymbol{\sigma})$ restricted to the subspace $\boldsymbol{\sigma}_k^\perp$ has the law of a $GOE(N-1)$ matrix scaled by $\sqrt{\xi''(q)}$, which has maximum eigenvalue roughly $2\sqrt{\xi''(q)}$ with high probability. This would show that such a $\boldsymbol{v}_k$ exists with high probability if $\boldsymbol{\sigma}_k$ were independent of H_N. Of course this is not the case. However by a famous result of [AG97], a $GOE(N-1)$ matrix has maximum eigenvalue at **least** $2\sqrt{\xi''(q)} - \eta$ with probability at least $1 - e^{-c(\eta)N^2}$. Thanks to the N^2 in the exponent, one can show by a union bound over an ε-net of the radius $\sqrt{qN}$ sphere that such a $\boldsymbol{v}_k$ always exists with high probability. This argument circumvents the dependence of $\boldsymbol{\sigma}_k$ on H_N.

By summing the energy gain accumulated at each step and taking $\delta \to 0$, it follows that Subag's

Input: Input tensors $\boldsymbol{G}^{(k)}$, accuracy parameter $\varepsilon > 0$.
Output: A point $\boldsymbol{\sigma} \in \mathbb{R}^N$ with L^2 norm $\|\boldsymbol{\sigma}\|_2 = \sqrt{N}$ such that $H_N(\boldsymbol{\sigma})/N \geq \mathsf{ALG}_\xi^{\mathrm{Sp}} - \varepsilon$.
Initialize $\boldsymbol{\sigma}_1 = (\sqrt{N\delta}, 0, \dots, 0) \in \mathbb{R}^N$ **for** $k = 1, \cdots, K$ **do**
  Find a unit vector $\boldsymbol{v}_k \perp \boldsymbol{\sigma}_k$ such that

$$\langle \boldsymbol{v}_k, \nabla^2 H_N(\boldsymbol{\sigma}_k)\boldsymbol{v}_k\rangle \geq 2\xi''(k\delta)^{1/2} - \delta\varepsilon;$$
$$\langle \boldsymbol{v}_k, \nabla H_N(\boldsymbol{\sigma}_k)\rangle \geq 0.$$

  $\boldsymbol{\sigma}_{k+1} = \boldsymbol{\sigma}_k + \sqrt{N\delta}\boldsymbol{v}_k$.
end
return $\boldsymbol{\sigma}_K$

algorithm succeeds in fiding $\boldsymbol{\sigma}$ on the sphere of radius $\sqrt{N}$ such that $H_N(\boldsymbol{\sigma})/N \geq \mathsf{ALG}_\xi^{\mathrm{Sp}} - \varepsilon$, for

$$\mathsf{ALG}_\xi^{\mathrm{Sp}} = \int_0^1 \xi''(q)^{1/2}\mathrm{d}q.$$

This value turns out to coincide with the asymptotic ground state energy in some cases. In fact:

Proposition 1.3.1. *The following are equivalent:*

1. $\mathsf{ALG}_\xi^{\mathrm{Sp}} = \mathsf{OPT}_\xi^{\mathrm{Sp}}$.
2. $\xi''(q)^{-1/2}$ *is concave on* $(0, 1]$.

Qualitatively, the above conditions are also known to coincide with the model having no overlap gap, a phenomenon discussed further below. Shortly after, Montanari [Mon21] gave a more complicated approximate message passing algorithm for the Sherrington-Kirkpatrick model on the cube under an assumption of no overlap gap. In the next section we discuss this and subsequent works, as well as results suggesting that these algorithms are best possible.

1.4 A Brief Description of Approximate Message Passing

Here we review the general class of approximate message passing (AMP) algorithms. AMP algorithms are a flexible class of efficient algorithms based on a random matrix or, in our setting, mixed tensor. AMP was introduced in the setting of Gaussian random matrices in [Bol14, BM11b]; we will rely on extensions of these results to tensors.

We begin with an elementary fact. Given a $GOE(N)$ random matrix $\boldsymbol{A}$ and vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$ independent of $\boldsymbol{A}$, one has

$$\mathbb{E}\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}\boldsymbol{y}\rangle = \langle \boldsymbol{x}, \boldsymbol{y}\rangle \tag{1.4.1}$$

and in fact $\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}\boldsymbol{y}\rangle$ is within a $1 \pm o_N(1)$ factor of its expectation with $1 - o_N(1)$ probability. Moreover the individual coordinate entries of $\boldsymbol{A}\boldsymbol{x}$ are essentially given by i.i.d. Gaussians. In the more general setting of a mixed p-spin glass Hamiltonian H_N with mixture function ξ, one analogously has

$$\mathbb{E}\langle \nabla H_N(\boldsymbol{x}), \nabla H_N(\boldsymbol{y})\rangle = \xi'(\langle \boldsymbol{x}, \boldsymbol{y}\rangle).$$

It is natural to ask what happens if one iterates these operations. For instance, does

$$\langle \boldsymbol{A}^2\boldsymbol{x}, \boldsymbol{A}^2\boldsymbol{y}\rangle \stackrel{?}{\approx} \langle \boldsymbol{x}, \boldsymbol{y}\rangle \tag{1.4.2}$$

also hold with high probability? In fact (1.4.2) is **not true**. For instance, we cannot apply (1.4.1) to $(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}\boldsymbol{y})$ because both vectors are dependent on $\boldsymbol{A}$. The idea of AMP is to explicitly account for the dependence causing (1.4.2) to fail. Conditioned on the product $\boldsymbol{A}\boldsymbol{x}$, the conditional law of $\boldsymbol{A}$ has "most of its randomness" left, and can be analyzed directly. The result is a precise description for the behavior of rather general iterative algorithms which consist of a constant number of multiplications-by-$\boldsymbol{A}$, or evaluations of $\nabla H_N(\cdot)$. In fact this description applies even when a non-linear function is applied coordinatewise between these gradient evaluations.

To specify an AMP algorithm, we fix a probability distribution p_0 on $\mathbb{R}$ with finite second moment and a sequence $f_0, f_1, \dots$ of Lipschitz functions $f_\ell : \mathbb{R}^{\ell+1} \to \mathbb{R}$, with $f_{-1} = 0$. The functions f_ℓ will often be referred to as *non-linearities*. We begin by taking $\boldsymbol{z}^0 \in \mathbb{R}^N$ to have i.i.d. coordinates $(z_i^0)_{i\in[N]} \sim p_0$. Then we recursively define $\boldsymbol{z}^1, \boldsymbol{z}^2, \dots$ via

$$\boldsymbol{z}^{\ell+1} = \nabla \widetilde{H}_N(f_\ell(\boldsymbol{z}^0, \dots, \boldsymbol{z}^\ell)) - \sum_{j=1}^{\ell} d_{\ell,j} f_{j-1}(\boldsymbol{z}^0, \dots, \boldsymbol{z}^{j-1}), \tag{1.4.3}$$

$$d_{\ell,j} = \xi''\left(\left\langle f_\ell(\boldsymbol{z}^0, \dots, \boldsymbol{z}^\ell) f_{j-1}(\boldsymbol{z}^0, \dots, \boldsymbol{z}^{j-1})\right\rangle_N\right) \cdot \mathbb{E}\left[\frac{\partial f_\ell}{\partial Z^j}(Z^0, \dots, Z^\ell)\right]. \tag{1.4.4}$$

Here $Z^0 \sim p_0$ while $(Z^\ell)_{\ell\geq 1}$ is an independent centered Gaussian process with covariance $Q_{\ell,j} = \mathbb{E}[Z^\ell Z^j]$ defined recursively by

$$Q_{\ell+1,j+1} = \xi'\left(\mathbb{E}\left[f_\ell\left(Z^0, \cdots, Z^\ell\right) f_j\left(Z^0, \cdots, Z^j\right)\right]\right), \quad \ell, j \geq 0. \tag{1.4.5}$$

The key property of AMP, stated below in Proposition 4.2.3, is that for any ℓ the empirical distribution of the N sequences $(\boldsymbol{z}_i^0, \boldsymbol{z}_i^1, \boldsymbol{z}_i^2, \dots, \boldsymbol{z}_i^\ell)_{i\in[N]}$ converges in distribution to the Gaussian process $(Z^0, Z^1, \dots, Z^\ell)$ as $N \to \infty$. This is called *state evolution*.

Definition 1.4.1. *For non-negative integers n, m the function $\psi : \mathbb{R}^n \to \mathbb{R}^m$ is* pseudo-Lipschitz *if*

for some constant L and any $x, y \in \mathbb{R}^n$,

$$||\psi(x) - \psi(y)|| \leq L(1 + ||x|| + ||y||)||x - y||.$$

Proposition 1.4.2 ([AMS21, Proposition 3.1])**.** *For any pseudo-Lipschitz $\psi : \mathbb{R}^{\ell+1} \to \mathbb{R}$, the AMP iterates satisfy*

$$\operatorname*{p-lim}_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \psi\left(\boldsymbol{z}_i^0, \cdots, \boldsymbol{z}_i^\ell\right) = \mathbb{E}\left[\psi\left(Z^0, \cdots, Z^\ell\right)\right]. \tag{1.4.6}$$

An early motivation for AMP was to give an analog of belief propagation, which is best suited for trees (or at least graphs with no short cycles) to the setting of dense graphs. The basic connection is that a random symmetric matrix can be viewed as the adjacency matrix of a weighted graph. This point of view is explained in [DMM10].

The majority of the AMP literature has focused on the matrix case. The state evolution result above for tensors was suggested in [RM14] and established in [AMS21]; we give the proof in the Appendix of this thesis. In the past decade, state evolution results have been established for many situations such as orthogonally invariant random matrices, semirandom matrices, matrices with i.i.d. sub-Gaussian entries, and more [JM13, BLM15, BMN19, CL21, Fan22]. Notably, the matrix $\boldsymbol{A}$ can include a signal, e.g. be of the form

$$\boldsymbol{A} = \boldsymbol{G}^{(2)} + \boldsymbol{x}_0^{\otimes 2}$$

for $\boldsymbol{G}^{(2)} \sim GOE(N)$. This extension is crucially used in Chapter 3. Because state evolution holds for essentially arbitrary non-linearities f_ℓ, it allows a great deal of flexibility in solving problems involving random matrices or tensors.

We close this discussion with a brief comparison of two different flavors of AMP algorithm. The first involves a simple fixed-point iteration using the same memory-free non-linearity $f_\ell(Z^0, \ldots, Z^\ell) = f(Z^\ell)$ at all times. In such a case, one can often prove that the (abstract) state evolution iterates Z^ℓ converge to a limit (with Gaussian law) as $\ell \to \infty$. This implies a similar convergence result for the true iterates $\boldsymbol{z}^\ell$ on time-scales which grow very slowly with N. Such an AMP algorithm is an important ingredient in Chapter 3, and serves as the first phase of the algorithm of Chapter 4 below.

On the other hand, [Mon21] introduced an *incremental* AMP (IAMP) algorithm of a very different flavor. The rough idea here is to simulate a Brownian motion by making the iterate Z^ℓ behave like the increment $B_{(\ell+1)\delta} - B_{\ell\delta}$ of a Brownian motion B_t, where $\delta > 0$ is a small constant. One can then define auxilliary functions of this discretized Brownian motion which correspond to diffusions driven by B_t. Such an IAMP procedure comprises the second phase of the main algorithm in Chapter 4. IAMP behaves quite differently from the fixed-point iteration above. For example we show in Chapter 4 that IAMP algorithms can be **branched** to output **many** far apart solutions of essentially the same quality.

1.5 Optimizing Mean-Field Spin Glasses: New Results

The main result of Chapter 4 in the Ising case can be described as follows. For a function $f : \mathbb{R} \to \mathbb{R}$ and interval J, let $\|f\|_{\mathrm{TV}(J)}$ denote the total variation of f on J, expressed as the supremum over partitions:

$$\|f\|_{\mathrm{TV}(J)} = \sup_n \sup_{t_0<t_1<\cdots<t_n, t_i \in J} \sum_{i=1}^{n} |f(t_i) - f(t_{i-1})|.$$

Let $\mathscr{L} \supseteq \mathscr{U}$ denote the set of functions given by

$$\mathscr{L} = \left\{ \begin{array}{ll} \zeta : [0,1) \to \mathbb{R}_{\geq 0} : & \zeta \text{ right-continuous}, \|\xi'' \cdot \zeta\|_{\mathrm{TV}[0,t]} < \infty \text{ for all } t \in [0,1), \\ & \displaystyle\int_0^1 \xi''(t)\zeta(t)\ \mathrm{d}t < \infty \end{array} \right\}. \tag{1.5.1}$$

It turns out (see Subsection 5.6.1) that the definition of P^{Is} above extends from $\mathscr{U}$ to $\mathscr{L}$. Therefore we may define $\mathsf{ALG}^{\mathrm{Is}} = \mathsf{ALG}^{\mathrm{Is}}_{\xi,h}$ by

$$\mathsf{ALG}^{\mathrm{Is}} = \inf_{\zeta \in \mathscr{L}} \mathsf{P}^{\mathrm{Is}}(\zeta). \tag{1.5.2}$$

Note that $\mathsf{ALG}^{\mathrm{Is}} \leq \mathsf{OPT}^{\mathrm{Is}}$ trivially holds. We have $\mathsf{ALG}^{\mathrm{Is}} = \mathsf{OPT}^{\mathrm{Is}}$ if the infimum in (5.1.9) is attained by some $\zeta \in \mathscr{U}$, and otherwise $\mathsf{ALG}^{\mathrm{Is}} < \mathsf{OPT}^{\mathrm{Is}}$. The following result is proved in Chapter 4 and is from [AMS21, Sel21b]. Below the "efficient AMP algorithms" considered use a constant (N-independent) number of steps. This results in computation time linear in the description length of H_N when ξ is a polynomial, assuming oracle access to the minimizer $\zeta_*^{\mathscr{L}} \in \mathscr{L}$ and corresponding solution $\Phi_{\zeta_*^{\mathscr{L}}}$ to (5.1.6).

Theorem 6. *Assume there exists $\zeta_*^{\mathscr{L}} \in \mathscr{L}$ such that $\mathsf{P}^{\mathrm{Is}}(\zeta_*^{\mathscr{L}}) = \mathsf{ALG}^{\mathrm{Is}}$. Then for any $\varepsilon > 0$, there exists an efficient AMP algorithm $\mathcal{A} : \mathcal{H}_N \to [-1,1]^N$ such that*

$$\mathbb{P}[H_N(\mathcal{A}(H_N))/N \geq \mathsf{ALG}^{\mathrm{Is}} - \varepsilon] \geq 1 - e^{-cN}, \quad c = c(\varepsilon) > 0.$$

In fact, fix also a finite ultrametric space (X, d_X) with diameter at most $\sqrt{2}$. Then there exists a $|X|$-tuple of efficient AMP algorithms outputting points $\{\boldsymbol{\sigma}_x | x \in X\}$ in $[-1,1]^N$ such that

$$\frac{H_N(\boldsymbol{\sigma}_x)}{N} \in [\mathsf{P}(\gamma_*) - \varepsilon, \mathsf{P}(\gamma_*) + \varepsilon], \quad x \in X, \tag{1.5.3}$$

$$\frac{||\boldsymbol{\sigma}_x - \boldsymbol{\sigma}_y||}{\sqrt{N}} \in [d_X(x,y) - \varepsilon, d_X(x,y) + \varepsilon], \quad x, y \in X \tag{1.5.4}$$

with probability $1 - e^{-cN}$.

The non-equality $\mathsf{ALG}^{\mathrm{Is}} < \mathsf{OPT}^{\mathrm{Is}}$ has a natural interpretation in terms of the optimizer ζ_* of (5.1.7). Namely, it implies that $\zeta_* \in \mathscr{U}$ is not strictly increasing (see Chapter 4 for a more

precise condition). Thus, in the case that ζ_* is strictly increasing, the above result implies that $\mathsf{ALG}^{\mathrm{Is}} = \mathsf{OPT}^{\mathrm{Is}}$; this was the condition assumed in [Mon21].

The construction of ultrametric configurations rather than single solutions in Theorem 6 is related to the overlap gap property discussed below. Indeed as we will see, the existence of such ultrametric configurations can be taken a posteriori as a purely geometric **definition** for the algorithmic thresholds $\mathsf{ALG}^{\mathrm{Is}}$ and $\mathsf{ALG}^{\mathrm{Sp}}$.

As discussed already, it is natural to ask what the best algorithms for optimizing the random function H_N are. Of course it seems difficult to establish any limitations on the power of general polynomial-time algorithms for such a task as this would require essentially resolving at least RP vs NP. However one might still hope to characterize the power of natural classes of algorithms that include gradient descent and approximate message passing. To this end, we define the following distance on the space $\mathcal{H}_N$ of Hamiltonians H_N. We identify H_N with its disorder coefficients $(\mathbf{G}^{(p)})_{p\in 2\mathbb{N}}$, which we concatenate (in an arbitrary but fixed order) into an infinite vector $\mathbf{g}(H_N)$. We equip $\mathcal{H}_N$ with the (possibly infinite) distance

$$\|H_N - H'_N\|_2 = \|\mathbf{g}(H_N) - \mathbf{g}(H'_N)\|_2.$$

A consequence of our results in Chapter 5 (obtained in [HS22] in collaboration with Brice Huang) is that no suitably *Lipschitz* function $\mathcal{A} : \mathcal{H}_N \to [-1,1]^N$ can surpass the asymptotic value $\mathsf{ALG}^{\mathrm{Is}}$.

Theorem 7 ([HS22])**.** *Let $\tau, \varepsilon > 0$ be constants. For N sufficiently large, any τ-Lipschitz $\mathcal{A} : \mathcal{H}_N \to [-1,1]^N$ satisfies*

$$\mathbb{P}\left[H_N(\mathcal{A}(H_N))/N \geq \mathsf{ALG}^{\mathrm{Is}} + \varepsilon\right] \leq \exp(-cN), \quad c = c(\xi, \varepsilon, \tau) > 0.$$

The algorithms of Theorem 6, as well as general gradient based algorithms on dimension-free time scales, are $O(1)$-Lipschitz in this sense. While the approach of [Sub21] is not Lipschitz, its performance is captured by AMP as explained in [AMS21, Remark 2.2].[1] Hence in tandem with these constructive results, Theorem 7 identifies the exact asymptotic value achievable by Lipschitz algorithms $\mathcal{A} : \mathcal{H}_N \to [-1,1]^N$ (assuming the existence of a minimizer $\zeta_* \in \mathscr{L}$ as required in Theorem 6). We also give analogous algorithms and impossibility results for spherical spin glasses, in which there is no question of existence of a minimizer $\zeta_*^{\mathscr{L}}$ on the algorithmic side. Let us finally remark that the rate e^{-cN} in Theorem 7 is best possible up to the value of c, being achieved even for the trivial algorithm $\mathcal{A}(H_N) = (1, 1, \ldots, 1)$ which ignores the disorder tensors $\boldsymbol{G}^{(k)}$ entirely.

[1] We also outline a similar impossibility result for a family of variants of [Sub21] in Subsection 5.3.7.

1.6 The Branching Overlap Gap Property

Theorem 7 states that the objective value ALG achieved in Theorem 6 is best possible within the class of algorithms with dimension-free Lipschitz dependence on the entries of the tensors $\boldsymbol{G}^{(k)}$. In particular, this condition is verified by standard optimization algorithms such as gradient descent, Langevin dynamics, and approximate message passing on dimension-free time scales. Here we discuss the main construction used in the proof.

First, it is worth mentioning the high-level heuristic that algorithmic hardness in a random optimization problem ought to be linked with the geometry of the solution space. One version of this connection was proposed in [ACO08, COE15] based on a *shattering* phase transition: at large constraint density the solution space breaks into exponentially many small components, for suitable random instances of k-SAT, q-coloring, and maximum independent set. Intuitively, shattering seems to pose a problem for algorithms, especially those based on local search. Other predictions based on the *clustering*, *condensation* [KMRT$^+$07] and *freezing* [ZK07] transitions have also been suggested.

Our approach is based on an extension of the overlap gap property (OGP). In the past several years, a line of such works [GS14, RV17a, GS17, CGPR19, GJ21, GJW20a, Wei22, GK21a, BH21, GJW21] have developed the OGP framework, turning properties of the solution space geomtry into rigorous algorithmic hardness results against restricted families of algorithms.

An overlap gap property argument typically considers pairs of points $(\boldsymbol{\sigma}_0, \boldsymbol{\sigma}_1)$ with medium overlap and large energy as "forbidden structures". One aims to show that such configurations do not exist at all. If fact in most cases, one needs to consider an ensemble of many correlated instances $H_N^0, H_N^1, \dots$ of the optimization problem and show that points $(\boldsymbol{\sigma}_0, \boldsymbol{\sigma}_1)$ with medium overlap cannot be good solutions to any pair H_N^i, H_N^j of problems in this ensemble. If this stronger property can be established, one gradually deforms the Hamiltonian H_N and argues that a "stable" algorithm will trace a continuous path, hence constructing such a forbidden configuration and obtaining a contradiction. The main technical difficulty in such an argument is to prove that such forbidden structures are completely absent with high probability.

In Chapter 4 we introduce a much richer type of forbidden structure. Namely, we consider arbitrary ultrametric trees of solutions with diameter at most $\sqrt{2N}$. We call this the *Branching OGP*. The definition involves an ensemble of "ultrametrically correlated" Hamiltonians H_N^u for indices $u = u_1 u_2 \dots u_D \in [k]^D$ corresponding to the leaves of a k-ary tree with depth D, such that H_N^u and H_N^v are p_d correlated if $u_i = v_i$ for $i \in [d]$ but not $i = d{+}1$, where $0 = p_0 \le p_1 \le \cdots \le p_D = 1$ is a finite increasing sequence. We consider a separate input $\boldsymbol{\sigma}^u$ for each H_N^u.

We show using concentration of measure that a Lipschitz algorithm $\mathcal{A} : \mathcal{H}_N \to [-1,1]^N$ applied to such an ensemble constructs with high probability an approximate ultrametric space of outputs

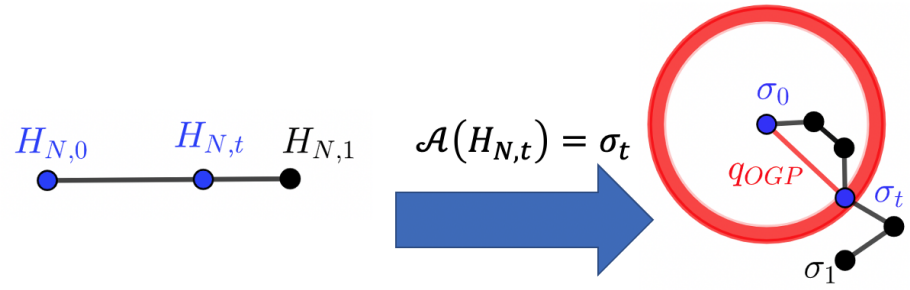


Figure 1.3: A cartoon overlap gap property argument. By a discrete-time version of the intermediate value theorem, one shows that a "stable" algorithm must produce an output $\boldsymbol{\sigma}_t$ at medium distance q_{OGP} from a previous output $\boldsymbol{\sigma}_0$. Hence if one can prove that no pair of good solutions can be at this medium distance from each other (thus establishing an OGP), one can rule out certain classes of algorithms from solving the random optimization problem well.

$\boldsymbol{\sigma}_u^{\mathrm{alg}} = \mathcal{A}(H_N^u)$. This is because by Gaussian concentration of measure, if H_N^u and H_N^v have p_d-correlated disorder, then the overlap

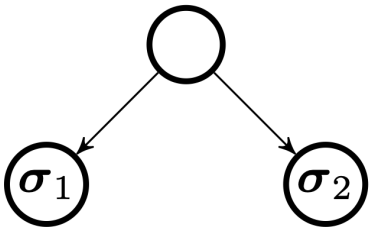
$$R(\mathcal{A}(H_N^u), \mathcal{A}(H_N^v))$$

concentrates with high probability. More precisely,

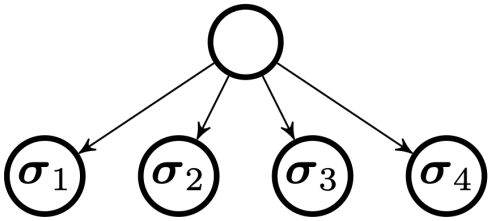
$$\mathbb{P}[R(\mathcal{A}(H_N^u), \mathcal{A}(H_N^v)) \in [q_d - \eta, q_d + \eta] \geq 1 - e^{-cN}$$

for a number $q_d = \chi_{\mathcal{A}}(p_d)$, arbitrarily small $\eta > 0$, and a constant $c(\xi, \mathcal{A}, \eta) > 0$ independent of p_d. In other words, a Lipschitz algorithm turns an ultrametric ensemble of Hamiltonians into an approximate ultrametric of outputs $(\boldsymbol{\sigma}_u^{\mathrm{alg}})_{u \in [k]^D}$. Using an extension of the Guerra-Talagrand interpolation, we upper-bound the total energy on any configuration $(\boldsymbol{\sigma}_u^{\mathrm{alg}})_{u \in [k]^D}$ with such an ultrametric overlap structure. This upper bound turns out to be smaller than $\mathsf{ALG} + \varepsilon$ once the ultrametric tree has sufficiently large depth $D = D(\varepsilon)$, with the choices $\delta = 1/D$ and $q_d = d\delta$.

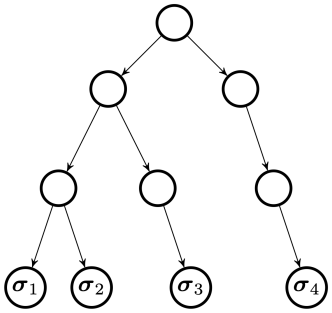
Finally we prove that our methods are in some necessary to identify the threshold ALG within the overlap gap property framework, at least for spherical spin glasses. More precisely we consider the result of an overlap gap property that can only use (arbitrary, possibly highly non-symmetric) ultrametric forbidden structures whose corresponding rooted trees can only contain full binary trees of **bounded** size at most D. We show that whenever $\mathsf{ALG}^{\mathrm{Sp}} < \mathsf{OPT}^{\mathrm{Sp}}$, the corresponding OGP-based threshold is bounded below by $\mathsf{ALG}^{\mathrm{Sp}} + \varepsilon_D$ for ε_D depending only on D and ξ. Thus, to establish $\mathsf{ALG}^{\mathrm{Sp}}$ as the exact threshold via an OGP, one essentially **must** use the full power of the branching OGP. Several intermediate-strength OGPs were used in previous work, as summarized in Figure 1.4.



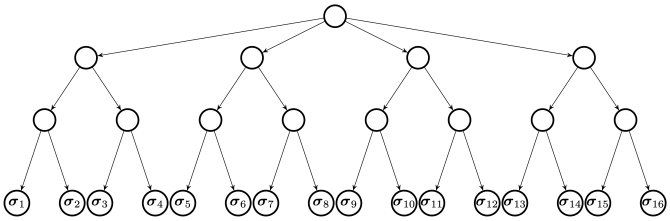
(a) **Classic OGP**: $\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2$ have medium overlap.



(b) **Star OGP**: many solutions, medium overlaps.



(c) **Ladder OGP**: medium "multi-overlaps" between $\boldsymbol{\sigma}_i$ and $\{\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_{i-1}\}$.



(d) **Branching OGP**: many solutions in an ultrametric tree.

Figure 1.4: Schematics of forbidden structures in overlap gap property arguments. The classic, star, and ladder OGPs have been used in several works to prove algorithmic hardness results in random optimization problems. The results of Chapter 5 are obtained by the branching overlap gap property.

1.7 Chasing Convex Bodies

We now turn our attention away from spin glasses and to a problem in real-time or *online* decision making. Let X be a d-dimensional normed space and $K_1, K_2, \ldots, K_T \subseteq X$ a finite sequence of convex bodies. In the *chasing convex bodies* problem, a player starting at $x_0 = 0 \in X$ learns the sets K_t one at a time, and after observing K_t moves to a point $x_t \in K_t$. The player's cost is the total path length

$$\text{cost}(x_1, \ldots, x_T) = \sum_{t=1}^{T} ||x_t - x_{t-1}||. \tag{1.7.1}$$

Denote the smallest cost (in hind-sight) among all such sequences by

$$\text{cost}(K_1, \ldots, K_T) = \min_{(y_t \in K_t)_{t \le T}} \sum_{t=1}^{T} ||y_t - y_{t-1}||.$$

The player's goal is to ensure that

$$\text{cost}(x_1, \ldots, x_T) \leq \alpha_d \cdot \text{cost}(K_1, \ldots, K_T) \tag{1.7.2}$$

holds for any sequence $K_1, \ldots, K_N$, where the *competitive ratio* α_d is as small as possible and is independent of N. We make no assumptions on the sets K_t and in fact allow them to be chosen adversarially, even possibly depending on the algorithm's previous choices.

We remark that unlike the algorithmic questions considered in the previous chapters, the issue of computational efficiency is now of secondary importance. Rather, the core difficulty is that the points $x_n = x_n(K_1, \ldots, K_n)$ must depend only on the sets revealed so far, i.e. the decisions must be made in real time. An online algorithm achieving (6.1.1) for some finite α_d is said to be α_d-*competitive*, and the smallest possible α_d among all online algorithms is the *competitive ratio* of the chasing convex bodies problem. The literature on competitive ratios for algorithmic problems is vast and includes scheduling [Gra66], self-organizing lists [ST85], efficient covering [AAA03], safely using machine-learned advice [BB00, KPS18, LV18b, WZ20], and the famous k-server problem [MMS90, Gro91, KP95, BBMN15].

The finiteness of the competitive ratio for convex body chasing was first posed in [FL93], which proved the case $d = 2$. In the past few years, the problem has seen renewed interest thanks to several applications such as efficiently powering data centers. The basic idea is that a point $x \in \mathbb{R}^d$ represents the state of (in this case) a data center. The data center receives a time-varying demand and meets this demand using resources of different types. Thus K_t represents the feasible region of server configurations which meet the demand at time t. If turning servers on and off is costly, then minimizing the objective in (1.7.1) is a natural goal. In fact one may consider the seemingly more general problem of chasing convex *functions* with cost

$$\text{cost}(x_1, \ldots, x_T) = \sum_{t=1}^{T} ||x_t - x_{t-1}|| + f_t(x_t)$$

for convex functions $f_n : \mathbb{R}^d \to \mathbb{R}_+$. This problem gives a lot of flexibility in modelling for instance, the time-t energy cost from leaving a server on.

In a prior work with Bubeck, Lee, and Li [BLLS19] we resolved this problem, showing that $\alpha_d \leq 2^{O(d)}$ for all $d \geq 1$. In Chapter 6, which is based on [Sel20], we give a d-competitive algorithm for chasing convex bodies in any normed space. This is an exponential improvement over the aforementioned result, and the competitive ratio of d is exactly optimal in the ℓ^∞ norm. (Indeed for random requests $K_t = \{x : x_t = \varepsilon_i\}$ for $t \in [d]$ with i.i.d. Rademacher variables $\varepsilon_i \in \{\pm 1\}$, no algorithm can be better than d-competitive in the ℓ^∞ norm even in expectation.) In the Euclidean norm, our algorithm is $O(\sqrt{d \log T})$ competitive after T steps, nearly matching a $\sqrt{d}$ lower bound

(proved by the same construction) for moderately sized T. Moreover the algorithm generalizes to chasing convex functions: the competitive ratio $\min(d, \sqrt{d \log T})$ simply increases by 1.

Our improved algorithm is based on the Steiner point, a classical object in convex geometry first defined in 1840 [Ste40]. Given a convex body K in a finite-dimensional normed space X, its Steiner point $s(K) \in K$ is an interior point whose definition we will not present here. Our idea is to follow the *functional Steiner point* of a suitable convex function. The function $W_t : \mathbb{R}^d \to \mathbb{R}_+$ used is the *work function*

$$W_t(x) \equiv \min_{\substack{(x_1,\ldots,x_t): \\ x_s \in K_s \ \forall s \in [t]}} \left(\|x - x_t\| + \sum_{s=1}^{t} \|x_s - x_{s-1}\| \right).$$

This work function essentially encodes the "effective total cost" of the sets $K_1, \ldots, K_t$ seen so far (and has an analogous definition for chasing convex functions).

In another previous work [BKL+20], we used the ordinary Steiner point to solve a special "nested" case of chasing convex bodies, together with Bubeck, Klartag, Lee, and Li. Concurrently with the main result presented in Chapter 6, a related algorithm for chasing convex bodies with $O(d)$ competitive ratio in the Euclidean norm was obtained in [AGGT21]. In fact we explain their algorithm at the end of Chapter 6 and show in a precise sense that it is almost the same as the functional Steiner point.

1.8 A Universal Law of Robustness via Isoperimetry

Chapter 7 is based on joint work [BS21] with Sébastien Bubeck. The motivation is the massive scale of modern machine learning, which often employs models with 100 times as many trainable parameters as examples. This is quite different from what most statistics and learning theory predicts, and moreover deep learning models are believed (and in some regimes, proved) to memorize essentially arbitrary labelled data.

We present a model for memorizing in high-dimension in which a surprisingly large number of parameters are required in order to memorize using a function with a small Lipschitz constant. The Lipschitz constant of a predictor is a natural proxy for its vulnerability to small adversarial input perturbations, which are a major concern in applications such as computer vision. An informal statement of our main result is as follows. Below, $\mathrm{poly}(n, d)$ denotes a quantity at most $(1 + n + d)^c$ for some constant c.

Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^d \to \mathbb{R}$ and let $(x_i, y_i)_{i=1}^n$ be i.i.d. input-output pairs in $\mathbb{R}^d \times [-1, 1]$. Assume that:

1. $\mathcal{F}$ admits a $\mathrm{poly}(n, d)$-Lipschitz parametrization by p real parameters, each of size at most $\mathrm{poly}(n, d)$.

2. The distribution μ of the covariates x_i has log-Sobolev constant $\Omega(d)$ (e.g. is uniform on the d-dimensional unit sphere), or is a mixture of $n^{0.99}$ such distributions.

3. The expected conditional variance of the output (i.e., the "noise level") is strictly positive, denoted $\sigma^2 \equiv \mathbb{E}^{\mu}[Var[y|x]] > 0$.

Then, with high probability over the sampling of the data, one has simultaneously for all $f \in \mathcal{F}$:

$$\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2 \leq \sigma^2 - \varepsilon \;\Rightarrow\; \mathrm{Lip}(f) \geq \widetilde{\Omega}\left(\frac{\varepsilon}{\sigma}\sqrt{\frac{nd}{p}}\right).$$

Here $\widetilde{\Omega}$ indicates a lower bound up to logarithmic factors.

Part I

Mixing and Sampling

Chapter 2

Cutoff for the Asymmetric Riffle Shuffle

2.1 Introduction

The riffle shuffle is among the most common methods to randomize a deck of cards. We study a parameterized model for riffle shuffles called p-shuffles, defined as follows for any $p \in (0,1)$. From a sorted deck of N cards, first remove the top $\texttt{Bin}(N,p)$ cards to create a top and a bottom pile. Next, interleave the two piles according to the following rule. If the piles currently have sizes A and B, the next card is dropped from the first pile with probability $\frac{A}{A+B}$. Conditioned on the pile sizes, this rule gives a uniformly random interleaving.

The case $p = \frac{1}{2}$, known as the Gilbert-Shannon-Reeds (GSR) shuffle, is perhaps the most natural model for riffle shuffling. It was analyzed by Bayer and Diaconis in [BD92] following work of Aldous ([Ald83, Example 4.17]); they proved that $\left(\frac{3}{2\log(2)} \pm o(1)\right)\log(N)$ shuffles are necessary and sufficient to randomize a deck. More precisely for any $\varepsilon > 0$, as $N \to \infty$ the total variation distance of the deck from a uniform permutation tends to 1 after $\left\lfloor \left(\frac{3}{2\log(2)} - \varepsilon\right)\log(N) \right\rfloor$ shuffles, and tends to 0 after $\left\lfloor \left(\frac{3}{2\log(2)} + \varepsilon\right)\log(N) \right\rfloor$ shuffles. In fact they showed that the total variation distance decays exponentially in C after $\frac{3\log(N)}{2\log(2)} + C$ shuffles.

By contrast, determining the mixing time for general p-shuffles has remained open. This discrepancy is because of a special property underpining the analysis in [BD92]: the deck order after a fixed number of GSR shuffles is uniformly random conditioned on its number of *rising sequences*. Therefore to understand the mixing time it suffices to understand how the number of rising sequences is distributed. This distribution turns out to admit a simple closed form, which enables

explicit analysis and a sharp understanding of the rate of convergence. When $p \neq \frac{1}{2}$ this conditional uniformity no longer holds and the problem becomes more complicated.

p-shuffles were introduced in [DFP92, Example 7] and further studied in [Lal96, Ful98, Lal00]. These works established upper and lower bounds of order $\log(N)$ on the mixing time, but with differing constant factors. Interestingly the eigenvalues of the p-shuffle chain are given explicitly by certain power sum symmetric functions. This follows from general results regarding random walks on hyperplane arrangements — see [BHR99, BD98b, Sta01] or the survey [Zha09].

Several aspects of riffle shuffles are surveyed in [Dia03]. Other interesting models arise from modifying the interleaving probabilities, such as the Thorpe shuffle [Tho73, Mor09, Mor13] and clumpy shuffle [JM15].

2.1.1 Main Result

The main result of this chapter is that all p-shuffles exhibit cutoff. More generally, let $\mathbf{p} = (p_0, \ldots, p_{k-1})$ be a discrete probability distribution with $p_i > 0$ for each i. We show cutoff for the more general $\mathbf{p}$-shuffles, which were also introduced in [DFP92]. To define such a shuffle, one first generates a multinomial $(N, \mathbf{p})$ vector $(n_0, \ldots, n_{k-1})$ so that each n_i has marginal distribution $n_i \sim \mathtt{Bin}(N, p_i)$ and $\sum_{i=0}^{k-1} n_i = N$ holds almost surely. One then splits the N cards into k piles by taking the top n_0 cards off the top to form the first pile, the next n_1 cards to form the second pile, and so on.

Interleaving the k piles into a single pile is done similarly to the $k = 2$ case. Namely, if the current remaining pile sizes are $A_0, \ldots, A_{k-1}$, then the next card is dropped from pile i with probability

$$\frac{A_i}{A_0 + A_1 + \cdots + A_{k-1}}. \tag{2.1.1}$$

This latter phase is again equivalent to interleaving the k piles uniformly at random conditioned on their sizes. Note that the asymmetry of $\mathbf{p}$ appears only in the first phase to determine the pile sizes and does not directly enter the second phase. When $\mathbf{p} = \left(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}\right)$, we recover the k-shuffle which is the k-partite analog of the GSR shuffle. k-shuffles exhibit cutoff after $\frac{3\log(N)}{2\log k} \pm O(1)$ steps by the same rising sequence analysis as in the $k = 2$ case ([BD92]).

To state our main result for general $\mathbf{p}$-shuffles, we must define several constants. With arbitrary tie-breaking, set $i_{\mathtt{max}} = \arg\max_{i \in \{0,1,\ldots,k-1\}}(p_i)$ and $p_{\mathtt{max}} = p_{i_{\mathtt{max}}}$. Similarly define $i_{\mathtt{min}}$ and $p_{\mathtt{min}}$. Define the functions

$$\phi_{\mathbf{p}}(t) = \sum_{i=0}^{k-1} p_i^t, \qquad \psi_{\mathbf{p}}(t) = -\log \phi_{\mathbf{p}}(t).$$

Define the positive constant $\theta_{\mathbf{p}}$ by the identity $\psi_{\mathbf{p}}(\theta_{\mathbf{p}}) = 2\psi_{\mathbf{p}}(2)$, i.e.

$$\phi_{\mathbf{p}}(\theta_{\mathbf{p}}) = \sum_{i=0}^{k-1} p_i^{\theta_{\mathbf{p}}} = \left(\sum_{i=0}^{k-1} p_i^2\right)^2 = \phi_{\mathbf{p}}(2)^2.$$

This uniquely determines $\theta_{\mathbf{p}}$ because $\phi_{\mathbf{p}}$ and $\psi_{\mathbf{p}}$ are strictly monotone. Finally define the constants $C_{\mathbf{p}}, \widetilde{C}_{\mathbf{p}}$, and $\overline{C}_p$ as follows.

$$\begin{aligned} C_{\mathbf{p}} &= \frac{3+\theta_{\mathbf{p}}}{4\psi_{\mathbf{p}}(2)} = \frac{3+\theta_{\mathbf{p}}}{2\psi_{\mathbf{p}}(\theta_{\mathbf{p}})}, \\ \widetilde{C}_{\mathbf{p}} &= \frac{1}{\log(1/p_{\mathtt{max}})}, \\ \overline{C}_{\mathbf{p}} &= \max(\widetilde{C}_{\mathbf{p}}, C_{\mathbf{p}}). \end{aligned}$$

We can now state our main result.

Theorem 8. *The* $\mathbf{p}$*-shuffles undergo total variation cutoff after* $\overline{C}_{\mathbf{p}} \log(N)$ *steps. That is, for any* $\varepsilon > 0$*,*

$$\lim_{N\to\infty} d_N^{\mathrm{TV}}(\lfloor (1-\varepsilon)\overline{C}_{\mathbf{p}} \log(N)\rfloor) = 1, \tag{2.1.2}$$

$$\lim_{N\to\infty} d_N^{\mathrm{TV}}(\lfloor (1+\varepsilon)\overline{C}_{\mathbf{p}} \log(N)\rfloor) = 0. \tag{2.1.3}$$

Here $d_N^{\mathrm{TV}}(K)$ *denotes the total variation distance from uniform after* $\mathbf{p}$*-shuffling* K *times.*

It is easy to see that $\overline{C}_{\mathbf{p}}$ is symmetric and continuous in the entries of $\mathbf{p}$. In the next proposition we show that for any k, the fastest possible mixing for any $\mathbf{p} = (p_0, \ldots, p_{k-1})$ occurs in the symmetric case $\mathbf{p} = \left(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}\right)$.

Proposition 2.1.1. *For any* k*,* $\overline{C}_{\mathbf{p}}$ *has minimum value* $\frac{3}{2\log k}$ *achieved uniquely at* $\mathbf{p} = \left(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}\right)$*. Moreover for any* $\mathbf{p}$*,*

$$C_{\mathbf{p}} \in \left[\frac{3}{2\psi_{\mathbf{p}}(2)}, \frac{7}{4\psi_{\mathbf{p}}(2)}\right) \quad \text{and} \quad \widetilde{C}_{\mathbf{p}} \in \left[\frac{1}{\psi_{\mathbf{p}}(2)}, \frac{2}{\psi_{\mathbf{p}}(2)}\right).$$

It also follows from Proposition 2.1.1 that for any $\mathbf{p}$, cutoff occurs in total variation occurs strictly sooner than in the L^∞ and separation distances. Quite precise results for these alternative notions of mixing are shown in [ADS12] by different methods, for the same asymmetric riffle shuffles that we study. In particular, cutoff occurs in both of these distances after $\frac{2\log(N)}{\psi_{\mathbf{p}}(2)} \pm O(1)$ shuffles. Recall that separation and L^∞ distance always upper-bound total variation distance, so only the strictness of the resulting inequality

$$\frac{2}{\psi_{\mathbf{p}}(2)} > \overline{C}_{\mathbf{p}}$$

between mixing time growth rates is non-trivial.

Proof of Proposition 2.1.1. When $\mathbf{p} = \left(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}\right)$ it is easy to see that $\theta_{\mathbf{p}} = 3$ and $\phi_{\mathbf{p}}(2) = \frac{1}{k}$. Therefore

$$C_{\mathbf{p}} = \frac{3}{2\log k} > \frac{1}{\log k} = \widetilde{C}_{\mathbf{p}}.$$

The value $\phi_{\mathbf{p}}(2)$ is symmetric and strictly convex in $\mathbf{p}$, hence achieves unique minimum at $\mathbf{p} = \left(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}\right)$. Moreover $\theta_{\mathbf{p}} \geq 3$ always holds as Cauchy–Schwarz implies

$$\phi_{\mathbf{p}}(2)^2 = \left(\sum_{i=0}^{k-1} p_i^2\right)^2 \leq \left(\sum_{i=0}^{k-1} p_i^3\right) \cdot \left(\sum_{i=0}^{k-1} p_i\right) = \sum_{i=0}^{k-1} p_i^3 = \phi_{\mathbf{p}}(3).$$

Therefore $C_{\mathbf{p}}$ achieves unique minimum at $\mathbf{p} = \left(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}\right)$, hence the first result. Moreover $\theta_{\mathbf{p}} < 4$ also holds because

$$\phi_{\mathbf{p}}(2)^2 = \left(\sum_{i=0}^{k-1} p_i^2\right)^2 > \sum_{i=0}^{k-1} p_i^4 = \phi_{\mathbf{p}}(4).$$

This shows that $C_{\mathbf{p}} \in \left[\frac{3}{2\psi_{\mathbf{p}}(2)}, \frac{7}{4\psi_{\mathbf{p}}(2)}\right)$. It remains to estimate $\widetilde{C}_{\mathbf{p}}$, and the claimed bounds amount to showing

$$\sum_{i=0}^{k-1} p_i^2 \leq p_{\mathtt{max}} < \sqrt{\sum_{i=0}^{k-1} p_i^2}.$$

The left inequality holds because

$$\sum_{i=0}^{k-1} p_i^2 \leq \sum_{i=0}^{k-1} p_i p_{\mathtt{max}} = p_{\mathtt{max}}$$

and the right inequality is clear. □

Our primary contribution is proving the upper bound (2.1.3), i.e. that the mixing time is at most $\overline{C}_{\mathbf{p}} \log(N)$. In Section 2.3 we reduce (2.1.3) to the estimation of a certain exponential moment, which occupies Sections 2.4 and 2.5. In the other direction, Lalley showed mixing time lower bounds of both $\widetilde{C}_{\mathbf{p}} \log(N)$ and $C_{\mathbf{p}} \log(N)$ in [Lal00]. However the latter result required $\mathbf{p} \approx \left(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}\right)$ to be close to uniform. ([Lal00] only considered the case $k = 2$, but the arguments work identically for larger k.) In Section 2.6 we generalize the $C_{\mathbf{p}} \log(N)$ lower bound to all $\mathbf{p} = (p_0, \ldots, p_{k-1})$ by refining Lalley's approach. For the sake of continuity, several of our notational choices, such as the constants $C_{\mathbf{p}}$ and $\widetilde{C}_{\mathbf{p}}$, are adopted from [Lal00]. However we reversed the sign of $\psi_{\mathbf{p}}$ from [Lal00] so that $\psi_{\mathbf{p}}(t) > 0$ for all $t > 1$.

Approximate Mixing Times $\overline{C}_{\mathbf{p}} \log N$ for p-Shuffles						
Deck Size	$p = 0.5$	$p = 0.6$	$p = 0.7$	$p = 0.8$	$p = 0.9$	$p = 0.95$
52	8.6	9.2	11.3	18	37	77
104	10.1	10.8	13.3	21	44	90
208	11.6	12.4	15.3	24	51	104
520	13.5	14.5	17.9	28	59	122
N	$2.16 \log N$	$2.32 \log N$	$2.86 \log N$	$4.5 \log N$	$9.5 \log N$	$19.5 \log N$

Table 2.1: The values $\overline{C}_{\mathbf{p}} \log N$ are shown for varying deck sizes N and $\mathbf{p} = (p, 1-p)$. These values should be taken as a rough guide because our results are asymptotic in N.

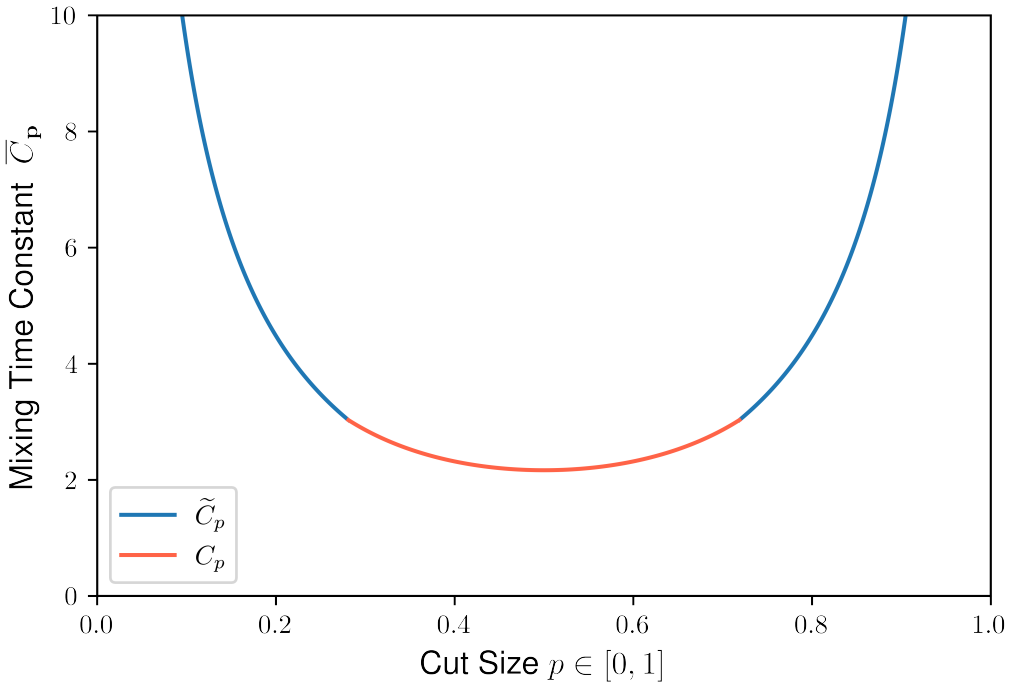


Figure 2.1: The values $\overline{C}_{\mathbf{p}}$ for $\mathbf{p} = (p, 1-p)$ are shown. The blue and red depict the transitions between $C_{\mathbf{p}}$ and $\widetilde{C}_{\mathbf{p}}$, which occur at $p \approx 0.28$ and $p \approx 0.72$. As $p \to 0$, the divergence is $\overline{C}_{\mathbf{p}} = \frac{1}{\log(1/(1-p))} = \frac{1}{p} + O(1)$.

2.2 Preliminaries

Let $P_{\mathbf{p}}$ denote the probability measure on the symmetric group $\mathfrak{S}_N$ given by applying a $\mathbf{p}$-shuffle to the identity. Given two discrete probability vectors $\mathbf{p} = (p_0, \ldots, p_{k-1})$ and $\mathbf{q} = (q_0, \ldots, q_{\ell-1})$ define their convolution

$$\mathbf{p} * \mathbf{q} \equiv (p_0 q_0, p_0 q_1, \ldots, p_0 q_{\ell-1}, p_1 q_0, \ldots, p_{k-1} q_{\ell-1}).$$

This convolution turns out to correspond to shuffle composition.

Proposition 2.2.1 ([DFP92, Example 7])**.** *Performing a* $\mathbf{q}$*-shuffle followed by a* $\mathbf{p}$*-shuffle is equivalent to performing a* $(\mathbf{p} * \mathbf{q})$*-shuffle. That is,*

$$P_{\mathbf{p}} * P_{\mathbf{q}} = P_{\mathbf{p}*\mathbf{q}}.$$

Proposition 2.2.1 yields an explicit description for the distribution $P_{\mathbf{p}^{*K}}$ of a deck after K shuffles. For instance in the "symmetric" setting of [BD92], it implies that composing a k_1-shuffle and a k_2-shuffle results in a $k_1 k_2$-shuffle. It will actually be more convenient for us to work with the inverse permutations. We now explain how to do this, following [Lal00]. First define a distribution on sequences

$$S = (s_1, \ldots, s_N)$$

of length K strings as follows. Generate N strings of length K, all with i.i.d. $\mathbf{p}$-random digits in

$$[k]_0 = \{0, \ldots, k-1\}.$$

S is obtained by sorting these strings into increasing lexicographic order

$$s_1 \leq_{\mathtt{lex}} s_2 \leq_{\mathtt{lex}} \cdots \leq_{\mathtt{lex}} s_N.$$

Recall that the lexicographic order on strings of the same length is just given by comparing their base k values. In general, the lexicographically smaller of two different $[k]_0$-strings is the one with the smaller digit at the first place where their digits differ, or is the shorter string if one string is a prefix of the other.

Next define the associated *shuffle graph* $G = G(S)$ on vertex set

$$[N] = \{1, 2, \ldots, N\}$$

in which $i, i+1 \in V(G)$ are neighbors if and only if $s_i = s_{i+1}$, and no other edges are in G. Hence G is a union of disjoint paths, which we call G*-components*. (We say S and $G = G(S)$ are $\mathbf{p}$-random when they are constructed in this way.) Finally choose a uniformly random permutation $\pi \in \mathfrak{S}_N$ and

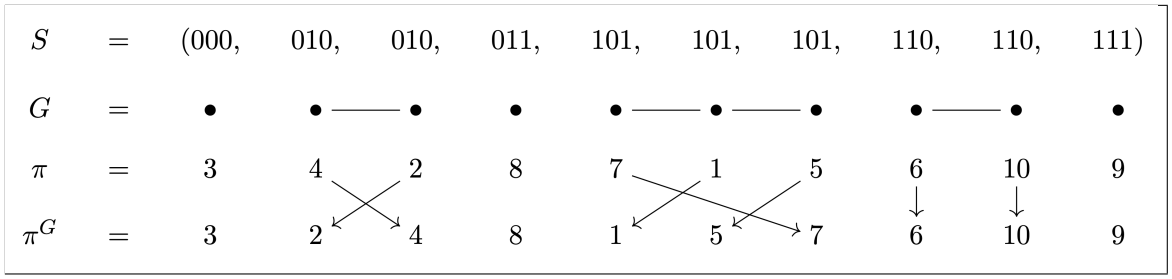


Figure 2.2: In this example with $N = 10$ strings in $[k]_0^K = [2]_0^3$, the lexiographically sorted sequence of strings S leads to the shuffle graph $G = G(S)$. The permutation $\pi \in \mathfrak{S}_N$ is then transformed into π^G by sorting within each G-component. By Proposition 2.2.2, the inverse $(\pi^G)^{-1}$ of the resulting permutation has distribution $P_{\mathbf{p}^{*K}}$.

define its G-modification π^G by, within each G-component, sorting the values $\pi(i)$ into increasing order. The next proposition states that π^G is exactly the inverse permutation of a $\mathbf{p}^{*K}$-shuffled deck.

Proposition 2.2.2 ([Lal00, Lemma 3])**.** *Let $\pi \in \mathfrak{S}_N$ be uniformly random and $G = G(S)$ be $\mathbf{p}$-random as defined above, for some fixed positive integers N and K. Then the distribution of $(\pi^G)^{-1}$ is exactly $P_{\mathbf{p}^{*K}}$. In particular, the total variation distance of π^G from uniform equals $d_N^{\mathrm{TV}}(K)$.*

In other words, the inverse permutation of a shuffled deck can be generated by starting with a uniformly random permutation π, and then modifying π to create π^G which is increasing on an independently random set of subintervals in $[N]$. After more and more shuffles, these subintervals shrink in distribution, leading eventually to mixing. In fact, L^∞ and separation mixing both correspond to G having no edges with high probability, see [Lal00, Corollary 3] and [ADS12]. However because G is random, total variation mixing can and does occur sooner. We refer the reader to [Lal00, Section 2] for more explanation and examples regarding Proposition 2.2.2. In brief, the N sequences $s_i \in [k]_0^K$ correspond to the sequences of pile-types that each of the N cards in the deck appears in during the shuffles. The sorting within G-components corresponds to the fact that if two cards are in the same pile during all K of the riffle shuffles, then their relative order must be preserved.

Throughout the remainder of this chapter, we work **entirely** with this transformed problem. Namely we will show that for $K \geq (1+\varepsilon)\overline{C}_p \log N$ the permutation π^G has total variation distance $o(1)$ from uniform, while for $K \leq (1-\varepsilon)\overline{C}_p \log N$ this distance is $1 - o(1)$.

2.2.1 Intuition Based on an Independent Point Process

There are two main obstructions to mixing which lead to the separate lower bounds of $\widetilde{C}_{\mathbf{p}}$ and $C_{\mathbf{p}}$. The simpler obstruction is that if $K \leq (\widetilde{C}_{\mathbf{p}} - \varepsilon)\log(N)$, then some strings will typically occur many

times, so π^G will contain an abnormally long increasing substring of length $N^{\Omega(1)}$. Indeed, from the definition $\widetilde{C}_{\mathbf{p}} = \frac{1}{\log(1/p_{\max})}$ it follows that after $K \le (\widetilde{C}_{\mathbf{p}} - \varepsilon)\log(N)$ shuffles, the expected number of strings with $s_j = i^K_{\max}$ is

$$\begin{aligned}\mathbb{E}\left|\{j \in [N] : s_j = i^K_{\max}\}\right| &= p^K_{\max} N \\ &\ge N^{-(\widetilde{C}_{\mathbf{p}} - \varepsilon)\log(1/p_{\max})+1} \\ &\ge N^{\Omega_\varepsilon(1)}.\end{aligned}$$

Since the number of such strings is binomially distributed, it is well-concentrated around its mean. Therefore with probability $1 - o(1)$ the $\mathbf{p}$-random shuffle graph G contains a length $N^{\Omega_\varepsilon(1)}$ path, and so π^G contains an increasing contiguous substring of the same length. However in a uniformly random permutation π, the probability to have an increasing substring of length $\ell \ge \log N$ is at most $N/(\ell!) = o(1)$. Therefore the total variation distance from uniform is $1 - o(1)$ when $K \le (\widetilde{C}_{\mathbf{p}} - \varepsilon)\log(N)$.

The more complicated obstruction to mixing comes from a fractal set of predictable locations (referred to as "cold spots" in [Lal00]) which tend to contain many G-edges. This obstruction, as well as our approach to the upper bound, can be motivated by an independent point process heuristic. (See also the last section of [Lal00].) Suppose we observe $\sigma \in \mathfrak{S}_N$ which is generated by either $\sigma = \pi$ or $\sigma = \pi^G$ for uniformly random $\pi \in \mathfrak{S}_N$ and $\mathbf{p}$-random G. Since the transformation $\pi \to \pi^G$ simply arranges small subintervals into increasing order, let us suppose that we observe only the ascent set $A(\sigma) = \{i : \sigma(i) < \sigma(i+1)\}$. As a heuristic, we may treat $A(\sigma)$ as an independent point process on edges in both the uniform $\sigma = \pi$ and shuffled $\sigma = \pi^G$ distributions. Specifically, for each $i \in [N-1]$ let

$$\eta_i \equiv \mathbb{P}[(i, i+1) \in E(G)].$$

be the probability for $(i, i+1)$ to be an edge in G. Then

$$\mathbb{P}[(i, i+1) \in A(\pi)] = \frac{1}{2}$$

while, roughly speaking,

$$\mathbb{P}[(i, i+1) \in A(\pi^G)] \approx \frac{1 + \eta_i}{2}.$$

(Technically $\mathbb{P}[(i, i+1) \in A(\pi^G)]$ should also depend on η_{i-1} and η_{i+1} but we will ignore this point.) This heuristic suggests that the likelihood ratio

$$\frac{\mathbb{P}^{\pi \in \mathfrak{S}_N}[\pi^G = \sigma]}{\mathbb{P}^{\pi \in \mathfrak{S}_N}[\pi = \sigma]}$$

evaluated at a uniformly random $\sigma \in \mathfrak{S}_N$ behaves like the random product

$$\prod_{i \in [N-1]} (1 \pm \eta_i)$$

where the $\pm$ signs are i.i.d. uniform. This product is close to 0 in probability (so mixing has not occured) if $\sum_i \eta_i^2 \gg 1$, and is close to 1 in probability (so mixing has occured) if $\sum_i \eta_i^2 \ll 1$.

Next observe that even without heuristic assumptions,

$$\sum_i \eta_i^2 = \mathbb{E}[E(G, G')]$$

is the expected size of the edge-intersection

$$E(G, G') \equiv E(G) \cap E(G')$$

of two independent **p**-random shuffle graphs G and G'. Therefore it is natural to guess that mixing occurs once $|E(G, G')|$ is typically small. Indeed, the quantity $|E(G, G')|$ will be crucial throughout. Let us finally summarize how it and related quantities appear in the proofs.

To lower bound the mixing time, one identifies deterministic "cold spot" sets $H \subseteq [N]$ which typically contain at least $|H|^{\frac{1}{2}+\delta}$ G-edges and shows that this implies non-mixing (see Proposition 2.6.1). The existence of such sets H implies in general that $\mathbb{E}[|E(G, G')|] \gg 1$ (Remark 2.6.1). Moreover in the independent point process model, the existence of such sets H is essentially equivalent to $\sum_i \eta_i^2 \gg 1$. Indeed, if $\sum_i \eta_i^2 \gg N^\delta$ then by the dyadic pigeonhole principle it follows that for some positive integer n there are at least $\Omega(2^{2n} N^{\delta/3})$ values $i \in [N-1]$ with $\eta_i \in [2^{-n}, 2^{-n+1}]$. These values of i can be taken for the set H.

On the other hand, it can happen that $\mathbb{E}[|E(G, G')|] \ll 1$ holds strictly before the onset of total variation mixing. This requires that $p_{\mathtt{max}} > \max(p_0, p_{k-1})$ and in particular $k \geq 3$ — see Remark 2.5.1. Instead as explained in Section 2.3, we reduce the mixing time upper bound (2.1.3) to showing that suitably truncated **exponential** moments of $|E(G, G')|$ are small. Estimating these exponential moments is rather involved. Our strategy is outlined just before the beginning of Subsection 2.3.2, and the proof occupies Sections 2.4 and 2.5.

These exponential moments arise naturally from considering (after some truncation) a chi-squared upper bound for total variation distance (see Lemma 2.3.3). In fact this seems to be a generally applicable method to upper-bound the total variation distance from a mixture of distributions with "random hidden structure" to a "null distribution" by controlling the interaction between two independent copies of the "structure" (in our setting, the graph G). For instance, a related observation was made in [MP12, Proposition 3.2] and later exploited in [LS16, LS17] to analyze information

percolation for the Ising model (see also the exposition [LS15]).

2.2.2 Notation

For any $M \geq 1$ the set $[k]_0^M$ consists of all length M strings with digits in $[k]_0$. (All strings will have digits in $[k]_0$.) Let

$$\mathcal{S} \subseteq ([k]_0^K)^N$$

denote the set of all lexicographically non-decreasing sequences $S = (s_1, \ldots, s_N)$ of N strings with length K each. Let $\mathcal{G}$ denote the set of all shuffle graphs, i.e. subgraphs of the path graph on N vertices. For $G \in \mathcal{G}$, let $\mathcal{C}(G)$ be the set of G-components, i.e. connected components of G.

Define $\mu_{\mathbf{p},M}$, often abbreviated as just $\mu_{\mathbf{p}}$, to be the probability measure on $[k]_0^M$ with each digit independently $\mathbf{p}$-random. By abuse of notation, we also use $\mu_{\mathbf{p},M}$ or simply $\mu_{\mathbf{p}}$ to denote the associated $\mathbf{p}$-random distributions on $\mathcal{S}$ or $\mathcal{G}$. We sometimes use square brackets to denote strings written out by their digits. For instance $[(k-1)(k-1)]$ indicates the string with two digits of $(k-1)$ while $[(k-1)(k-1)0^{K-2}]$ denotes the string with two initial $(k-1)$-digits followed by $K-2$ final 0-digits. We also occasionally use brackets to denote digits of a string, so for instance the digit expansion of a string x may be written

$$x = x[1]x[2]\ldots x[M] \in [k]_0^M.$$

We write $\mathbb{E}^\sigma, \mathbb{E}^\pi, \mathbb{P}^\sigma$, and $\mathbb{P}^\pi$ to denote expectations or probabilities taken over uniformly random permutations σ or π in $\mathfrak{S}_N$. We similarly write $\mathbb{E}^S$ to indicate expectation over $S \sim \mu_{\mathbf{p},K}$. We will continue to use $E(G, G') = E(G) \cap E(G')$ to denote the edge-intersection of $G, G' \in \mathcal{G}$. S' and $G' = G(S')$ will always denote independent copies of S and G.

The following definitions are used to prove Lemma 2.3.9 at the end of Section 2.3, and otherwise do not appear until Section 2.4. For each string

$$x = x[1]x[2]\ldots x[M] \in [k]_0^M$$

with $M \geq 1$ a positive integer, define

$$t_x \equiv \mathbb{P}^{y \sim \mu_{\mathbf{p},M}}[y <_{\texttt{lex}} x], \tag{2.2.1}$$

$$\lambda_x \equiv \mathbb{P}^{y \sim \mu_{\mathbf{p},M}}[y = x] = \prod_{i=1}^{M} p_{x[i]}, \tag{2.2.2}$$

$$J_x \equiv [t_x, t_x + \lambda_x). \tag{2.2.3}$$

Hence the intervals $(J_x)_{x\in[k]_0^M}$ partition $[0,1)$ for any fixed M. Note that

$$t_x + \lambda_x = \mathbb{P}^{y\sim\mu_{\mathbf{p},M}}[y \leq_{\mathtt{lex}} x].$$

It will often be useful to observe that to sample a $\mathbf{p}$-random string $x \in [k]_0^M$, one may equivalently sample a uniform random variable $a \in [0,1]$ and take the unique x with $a \in J_x$. Similarly to sample $(s_1,\dots,s_N) \in \mathcal{S}$, one may instead sample uniform i.i.d.

$$a'_1,\dots,a'_N \in [0,1],$$

sort them into increasing order

$$0 \leq a_1 \leq \cdots \leq a_N \leq 1,$$

and finally choose $s_i \in [k]_0^K$ such that $a_i \in J_{s_i}$ for each $i \in [N]$.

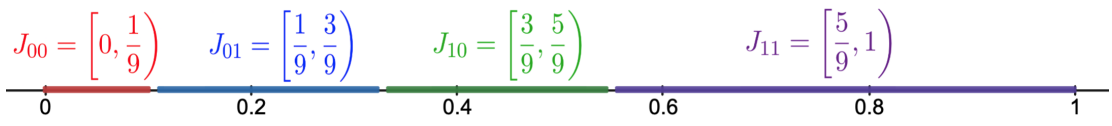


Figure 2.3: The partition $[0,1) = \bigcup_{x\in[k]_0^M} J_x$ with $k = 2, M = 2$, and $(p_0, p_1) = \left(\frac{1}{3}, \frac{2}{3}\right)$.

2.3 Upper Bound Approach

Here we explain some of the ingredients used to prove the mixing time upper bound (2.1.3). In Subsection 2.3.1 we present the more conceptual parts, ultimately reducing (2.1.3) to a certain exponential moment estimate. In Subsection 2.3.2 we prove a few other lemmas used in Subsection 2.3.1. This section might be viewed as an extended setup for the more difficult parts of the proof. For instance the constant $C_{\mathbf{p}}$ does not explicitly enter until the next section. However we emphasize that the results of this section are both specific to the particular problem considered and essential to understand the remainder of the argument.

2.3.1 High-Level Approach

We begin by carefully examining the Radon–Nikodym derivative between the distributions of π^G and π where π is a uniformly random permutation. For each $G \in \mathcal{G}$, let $\mathcal{C}(G) = \{G_1,\dots,G_j\}$ be the G-components, and suppose that each G_i contains v_i vertices. Then it is easy to see that the

map $\mathfrak{S}_N \to \mathfrak{S}_N$ given by $\pi \to \pi^G$ is $\left(\prod_{i=1}^{j} v_i!\right)$ to 1. Moreover its image consists of those σ with $\sigma^G = \sigma$. Therefore

$$\mathbb{P}^\pi[\pi^G = \sigma] = 1_{\sigma=\sigma^G} \cdot \frac{\prod_{i=1}^{j} v_i!}{N!}, \quad \sigma \in \mathfrak{S}_N\,.$$

As a consequence, for fixed $G \in \mathcal{G}$ the Radon–Nikodym derivative $f_{G,\sigma}$ of π^G with respect to π is given by

$$\begin{aligned} f_{G,\sigma} &\equiv \frac{\mathbb{P}^\pi[\pi^G = \sigma]}{\mathbb{P}^\pi[\pi = \sigma]} \\ &= N! \cdot \mathbb{P}^\pi[\pi^G = \sigma] \\ &= 1_{\sigma^G=\sigma} \cdot \prod_{i=1}^{j} v_i! \\ &= \frac{1_{\sigma^G=\sigma}}{\mathbb{P}^\pi[\pi^G = \pi]}. \end{aligned}$$

Observe that for fixed $G \in \mathcal{G}$,

$$\overset{\sigma}{\mathbb{E}}[f_{G,\sigma}] = 1. \tag{2.3.1}$$

On the other hand for fixed σ and $\mu_{\mathbf{p},K}$-random $G = G(S)$, we may apply the law of total expectation to the second definition of $f_{G,\sigma}$ above. This implies that for fixed σ,

$$\mathbb{P}^{\pi,S}[\pi^{G(S)} = \sigma] = \frac{\mathbb{E}^S[f_{G(S),\sigma}]}{N!}.$$

Therefore the total variation distance to uniform after K shuffles is given by

$$d_N^{\mathrm{TV}}(K) = \frac{1}{2} \cdot \overset{\sigma}{\mathbb{E}} \left|\mathbb{E}^S[f_{G(S),\sigma} - 1]\right|.$$

Next, we use a chi-squared upper bound for total variation distance after removing exceptional sequences from $\mathcal{S}$. To carry this out, given a partition $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_0$ (where $\mathcal{S}_1$ consists of "typical" sequences), write

$$\begin{aligned} \mathbb{E}^\sigma \left|\mathbb{E}^S[f_{G(S),\sigma} - 1]\right| &\le \mathbb{E}^\sigma \left|\mathbb{E}^S[(f_{G(S),\sigma} - 1)1_{S\in\mathcal{S}_1}]\right| + \mathbb{E}^\sigma \left|\mathbb{E}^S[(f_{G(S),\sigma} - 1)1_{S\in\mathcal{S}_0}]\right| \\ &\le \mathbb{E}^\sigma \left|\mathbb{E}^S[(f_{G(S),\sigma} - 1)1_{S\in\mathcal{S}_1}]\right| + \mathbb{E}^{\sigma,S}[(f_{G(S),\sigma} + 1)1_{S\in\mathcal{S}_0}] \\ &\overset{(2.3.1)}{=} \mathbb{E}^\sigma \left|\mathbb{E}^S[(f_{G(S),\sigma} - 1)1_{S\in\mathcal{S}_1}]\right| + 2\mu_{\mathbf{p}}(\mathcal{S}_0). \end{aligned} \tag{2.3.2}$$

Take S' to be an independent copy of S and define for any shuffle graphs $G, G' \in \mathcal{G}$

$$f_{G,G'} \equiv \mathbb{E}^\sigma[f_{G,\sigma} f_{G',\sigma}].$$

We now use the Cauchy–Schwarz inequality to upper-bound the main term of (2.3.2) via

$$\begin{aligned}\left(\mathbb{E}^{\sigma}\left|\mathbb{E}^{S}[(f_{G(S),\sigma}-1)1_{S\in\mathcal{S}_1}]\right|\right)^2 &\le \mathbb{E}^{\sigma}\left[\left(\mathbb{E}^{S}[(f_{G(S),\sigma}-1)1_{S\in\mathcal{S}_1}]\right)^2\right]\\ &= \mathbb{E}^{\sigma}\mathbb{E}^{S,S'}[(f_{G(S),\sigma}-1)(f_{G(S'),\sigma}-1)1_{S,S'\in\mathcal{S}_1}]\\ &= \mathbb{E}^{\sigma}\mathbb{E}^{S,S'}[(f_{G(S),\sigma}f_{G(S'),\sigma}-1)1_{S,S'\in\mathcal{S}_1}]\\ &= \mathbb{E}^{S,S'}\left[(f_{G,G'}-1)1_{S,S'\in\mathcal{S}_1}\right]. \end{aligned} \tag{2.3.3}$$

The second equality holds by switching the order of expectation and using (2.3.1). Starting from (2.3.3) and throughout the remainder of the chapter, we set $G = G(S), G' = G(S')$. Based on (2.3.3), to establish mixing it remains to show that $f_{G,G'}$ rarely exceeds 1 in an L^1 sense (modulo choosing $\mathcal{S}_1$ and $\mathcal{S}_0$).

We will upper-bound $f_{G,G'}$ using the number $|E(G,G')|$ of edges shared by G and G'. As motivation for why such a relationship should exist, observe that if no vertex $i \in [N]$ is incident to both a G-edge and a G'-edge, then $f_{G,\sigma}$ and $f_{G',\sigma}$ are **exactly** independent for $\sigma \in \mathfrak{S}_N$ uniformly random. Hence in this case we have the exact equality

$$f_{G,G'} = \mathbb{E}^{\sigma}[f_{G,\sigma}f_{G',\sigma}] = \mathbb{E}^{\sigma}[f_{G,\sigma}]\mathbb{E}^{\sigma}[f_{G',\sigma}] \overset{(2.3.1)}{=} 1.$$

In fact Lemma 2.3.1 below implies that $f_{G,G'} \le 1$ holds whenever $|E(G,G')| = 0$. In essence, incident but non-overlapping edges only reduce $f_{G,G'}$. It is now unsurprising that $f_{G,G'}$ can be bounded above by some function of $|E(G,G')|$. We show in Lemma 2.3.3 that this dependence is at most exponential when a condition called L-sparsity holds for both G and G'. The requirement of L-sparsity will be part of the eventual definition of $\mathcal{S}_1$.

In general, for any shuffle graphs G and G', define the new shuffle graph U to be their edge-union with U-components $\mathcal{C}(U)$. The next lemma shows how to estimate $f_{G,G'}$ based on the intersection structure of G and G'. The proof is deferred to the next subsection.

Lemma 2.3.1. *Suppose the U-components have vertex-sizes $(u_1,\dots,u_c)$. Then*

$$f_{G,G'} \le \prod_{\substack{1\le i\le c,\\ E(U_i)\cap E(G,G')\neq\emptyset}} (u_i!). \tag{2.3.4}$$

We now define the first condition that "typical" sequences in $\mathcal{S}_1$ must satisfy. The objective is to ensure that the u_i in Lemma 2.3.1 are uniformly bounded by some constant $L = L(\mathbf{p},\varepsilon)$. Let us point out that it is not enough to argue that $\max_i(u_i) \le L$ holds with high probability over random pairs (S,S'). Indeed, the truncation step (2.3.2) was used to remove $\mathcal{S}_0$ before applying Cauchy–Schwarz to introduce S'. There is no analogous way to remove an arbitrary low-probability

subset of **pairs** $(S, S') \in \mathcal{S}$. It is therefore important that the definition of L-sparsity below implies $\max_i(u_i) \leq L$ via separate restrictions on G and G'.

Definition 2.3.2. *For $L \geq 10$ a positive integer, a shuffle graph G is **L-sparse** if within any discrete interval $\{i, i+1, \ldots, i+L-1\} \subseteq [N]$ of L consecutive vertices, at most $L/3$ (of the possible $L-1$) edges are in $E(G)$.*

Lemma 2.3.3. *Suppose G and G' are L-sparse shuffle graphs. Then $f_{G,G'} \leq (L!)^{|E(G,G')|}$.*

Proof. We claim that $\max_i(u_i) \leq L$, i.e. each U-component contains at most L vertices. Indeed by L-sparsity, U contains at most $\frac{2L}{3} < L-1$ edges within each subinterval of L vertices, hence no such interval can be a connected subgraph of U. Therefore Lemma 2.3.1 implies that

$$f_{G,G'} \leq \prod_{\substack{1 \leq i \leq c, \\ E(U_i) \cap E(G,G') \neq \emptyset}} (L!).$$

By definition, $|E(G, G')|$ is at least the number of components U_i satisfying $E(U_i) \cap E(G, G') \neq \emptyset$. This completes the proof. □

Given Lemma 2.3.3, our main remaining task is to control the (truncated) exponential moments of $|E(G, G')|$. For technical reasons outlined at the end of this subsection, we will cover $E(G, G')$ by a union $E(G, G') = E_{\texttt{for}}(G, G') \cup E_{\texttt{back}}(G, G')$ of two sets which omit lexicographically late and early strings respectively. To ensure that $E(G, G')$ can be covered in this way for $G, G' \in \mathcal{S}_1$, we add a second restriction to the definition of $\mathcal{S}_1$ called regularity. This amounts to requiring that both prefixes [00] and $[(k-1)(k-1)]$ appear with roughly the expected frequency among the strings $(s_1, \ldots, s_N)$ of S.

Definition 2.3.4. *The sequence $S = (s_1, \ldots, s_N) \in \mathcal{S}$ of strings is **regular** if at most $\left(p_0^2 + \frac{p_0 p_{k-1}}{2}\right) N$ strings s_i begin with [00] (two consecutive 0 digits) and at most $\left(p_{k-1}^2 + \frac{p_0 p_{k-1}}{2}\right) N$ strings begin with $[(k-1)(k-1)]$ (two consecutive $(k-1)$ digits).*

Lemma 2.3.5. *For any $\mathbf{p}$ and $\varepsilon > 0$ there exist $L = L(\mathbf{p}, \varepsilon) \in \mathbb{Z}^+$ and $\delta = \delta(\mathbf{p}, \varepsilon) > 0$ such that the following holds. Consider a $\mathbf{p}$-random sequence $S = (s_1, \ldots, s_N)$ of strings of length $K \geq (\widetilde{C}_{\mathbf{p}} + \varepsilon) \log(N)$. Then with probability $1 - O(N^{-\delta})$, S is regular and $G(S)$ is L-sparse.*

The proof is deferred to the next subsection. $\mathcal{S}_1$ can now be defined: it consists of the regular sequences S for which $G(S)$ is L-sparse for $L = L(\mathbf{p}, \varepsilon)$ as in Lemma 2.3.5. Then Lemma 2.3.5 exactly states that

$$\mu_{\mathbf{p}}(\mathcal{S}_0) = O(N^{-\delta})$$

for some small $\delta = \delta(\mathbf{p}, \varepsilon)$.

We remark that the convergence rate $O(N^{-\delta})$ eventually appears as the upper bound for the total variation distance to uniformity (see (2.3.5) and the next displayed equations in that proof). The rate $O(N^{-\delta})$ seems to be tight in e.g. Proposition 2.4.1 via Lemma 2.4.15. As a result we use this rate in the statement of Lemma 2.3.5 although it could be improved. In fact the probability for S to be regular is at least $1 - e^{-\Omega_{\mathbf{p}}(N)}$. The probability for $G(S)$ to be L-sparse can be made at most e^{-CN} for any desired $C > 0$, if $L = L(C, \mathbf{p}, \varepsilon)$ is taken sufficiently large.

Next we show how to cover $E(G, G')$ when S and S' are regular.

Definition 2.3.6. *Let $E_{\mathtt{for}}(G)$ consist of all edges $(i, i+1) \in E(G)$ for which the strings $s_i = s_{i+1}$ do **not** begin with prefix $[(k-1)(k-1)]$. Let $E_{\mathtt{for}}(G, G') = E_{\mathtt{for}}(G) \cap E_{\mathtt{for}}(G')$. Define $E_{\mathtt{back}}(G, G')$ in the same way but with $[(k-1)(k-1)]$ replaced by $[00]$.*

Lemma 2.3.7. *If $S, S' \in \mathcal{S}$ are regular, then*

$$|E(G, G')| \leq |E_{\mathtt{for}}(G, G')| + |E_{\mathtt{back}}(G, G')|.$$

Proof. Regularity implies that $E_{\mathtt{for}}(G, G')$ contains all shared edges $(i, i+1) \in E(G, G')$ with

$$i \leq (1 - p_{k-1}^2 - (p_0 p_{k-1}/2))N,$$

and $E_{\mathtt{back}}(G, G')$ contains all shared edges $(i, i+1) \in E(G, G')$ with

$$i \geq (p_0^2 + (p_0 p_{k-1}/2))N.$$

Since

$$p_0^2 + p_0 p_{k-1} + p_{k-1}^2 < (p_0 + p_{k-1})^2 \leq 1$$

we obtain

$$(p_0^2 + (p_0 p_{k-1}/2))N \leq (1 - p_{k-1}^2 - (p_0 p_{k-1}/2))N.$$

Therefore

$$E_{\mathtt{for}}(G, G') \cup E_{\mathtt{back}}(G, G') = E(G, G')$$

which implies the result. □

Using symmetry to suppress the identical case of $E_{\mathtt{back}}$, to establish the mixing time upper bound in Theorem 8 it remains to verify the following lemma.

Lemma 2.3.8. *For any $\mathbf{p}$ and positive reals ε and t, there is $\delta = \delta(\mathbf{p}, \varepsilon, t)$ such that if $K \geq (\overline{C}_{\mathbf{p}} + \varepsilon)\log(N)$ then*

$$\mathbb{E}[e^{t \cdot |E_{\mathtt{for}}(G, G')|}] \leq 1 + O(N^{-\delta}).$$

Indeed, the mixing time upper bound (2.1.3) in Theorem 8 easily follows from the results above as we show now.

Proof of (2.1.3). Let $\delta > 0$ be sufficiently small depending on $(\mathbf{p}, \varepsilon, L, t)$, some of which are yet to be chosen. By (2.3.2) and (2.3.3),

$$\begin{aligned} d_N^{\mathrm{TV}}(K) &= \frac{1}{2} \cdot \mathbb{E}^{\sigma} \left| \mathbb{E}^{S}[f_{G(S),\sigma}] - 1 \right| \\ &\leq \frac{1}{2} \cdot \sqrt{\mathbb{E}^{S,S'}\left[(f_{G,G'} - 1) 1_{S,S' \in \mathcal{S}_1}\right]} + \mu_{\mathbf{p}}(\mathcal{S}_0). \end{aligned} \tag{2.3.5}$$

(It follows from (2.3.3) that the expression inside the square-root is non-negative.) Since $\mu_{\mathbf{p}}(\mathcal{S}_0) = O(N^{-\delta})$ by Lemma 2.3.5, it remains to estimate $\mathbb{E}^{S,S' \in \mathcal{S}}\left[(f_{G,G'} - 1) 1_{S,S' \in \mathcal{S}_1}\right]$. Using Lemma 2.3.3 in the first step, then Lemma 2.3.7 and finally Lemma 2.3.8 with $t = 2\log(L!)$, we obtain

$$\begin{aligned} \mathbb{E}^{S,S' \in \mathcal{S}}\left[(f_{G,G'} - 1) 1_{S,S' \in \mathcal{S}_1}\right] &\leq \mathbb{E}^{S,S'}\left[\left((L!)^{|E(G,G')|} - 1\right) 1_{S,S' \in \mathcal{S}_1}\right] \\ &\leq \mathbb{E}\left[\left((L!)^{|E_{\mathtt{for}}(G,G')| + |E_{\mathtt{back}}(G,G')|} - 1\right) 1_{S,S' \in \mathcal{S}_1}\right] \quad \text{Lemma 2.3.7} \\ &\leq \mathbb{E}\left[(L!)^{|E_{\mathtt{for}}(G,G')| + |E_{\mathtt{back}}(G,G')|} - 1\right] \\ &\leq \frac{\mathbb{E}\left[(L!)^{2|E_{\mathtt{for}}(G,G')|}\right] + \mathbb{E}\left[(L!)^{2|E_{\mathtt{back}}(G,G')|}\right]}{2} - 1 \\ &\leq O(N^{-\delta}). \quad \text{Lemma 2.3.8} \end{aligned}$$

Combining the above, we conclude that $d_N^{\mathrm{TV}}(K) \leq O(N^{-\delta})$ when $K \geq (\overline{C}_{\mathbf{p}} + \varepsilon)\log(N)$. □

The above constitutes a complete proof for the upper bound, except that Lemmas 2.3.1, 2.3.5 and 2.3.8 are yet to be proved. The first two are not difficult and are handled in the next subsection. Lemma 2.3.8 is more challenging and its proof occupies Sections 2.4 and 2.5. We now outline our approach to Lemma 2.3.8, which starts from the following basic fact. Suppose $X \in \mathbb{N}$ is a non-negative integer-valued random variable satisfying the uniform hazard rate bound

$$\sup_{j \geq 0} \mathbb{P}[X \geq j + 1 | X \geq j] \leq O(N^{-\delta}) \tag{2.3.6}$$

for some $\delta > 0$. Then X is stochastically dominated by a geometric random variable with mean $O(N^{-\delta})$, and therefore $\mathbb{E}[e^{tX}] = 1 + O(e^t N^{-\delta}) = 1 + o(1)$ for any constant t. To prove Lemma 2.3.8, we will implement this idea with $X = |E_{\mathtt{for}}(G, G')|$. We explore G and G' by revealing their strings together in order, so that

$$(s_1, \ldots, s_i) \quad \text{and} \quad (s'_1, \ldots, s'_i)$$

have been revealed at time $i \in [N]$. We show that at *any* time, the expected number of unrevealed

edges in $E_{\texttt{for}}(G, G')$ is at most $O(N^{-\delta})$. That is, almost surely,

$$\mathbb{E}\left[\left|E_{\texttt{for}}(G,G')|_{\{i,i+1,\ldots,N\}}\right| \;\middle|\; (s_1,\ldots,s_i,s'_1,\ldots,s'_i)\right] \le O(N^{-\delta}). \tag{2.3.7}$$

(Here we write $E_{\texttt{for}}(G,G')|_{\{i,i+1,\ldots,N\}}$ to indicate the set of edges $(j, j+1) \in E_{\texttt{for}}(G,G')$ with $j \ge i$.) The estimate (2.3.7) readily implies Lemma 2.3.8 analogously to the above discussion on (2.3.6). See Lemma 2.5.4 for a detailed proof.

As a first step towards establishing (2.3.7), in Section 2.4 we prove for $K \ge (\overline{C}_{\mathbf{p}} + \varepsilon)\log(N)$ the weaker first moment bound

$$\mathbb{E}\left[|E(G,G')|\right] \le O(N^{-\delta}). \tag{2.3.8}$$

In Section 2.5 we use (2.3.8) to show (2.3.7). The idea is to group the set of possible future strings

$$\{s \in [k]_0^K : s \ge_{\texttt{lex}} s_i\}$$

into a small number of blocks. Here each block B_x consists of all strings beginning with some prefix $x \in [k]_0^M$ (where $M = M(x)$ depends on x). Such a block B_x is essentially equivalent to a copy of $[k]_0^{K-M}$. The idea is to first estimate the left-hand side of (2.3.7) by a sum over blocks (using Cauchy–Schwarz several times), and to then estimate the contribution of each block using (2.3.8). The total number of blocks will always be $O(\log N) \le N^{o(1)}$. Therefore summing over blocks is no problem (up to adjusting the value of δ slightly) as long as the hypothesis of (2.3.8) applies "within" each block.

To illustrate the key reason for introducing $E_{\texttt{for}}$, let us explain why (2.3.7) can be false if $E_{\texttt{for}}(G,G')$ is replaced by $E(G,G')$. Suppose that $s_i = s'_i = [(k-1)^K]$ holds for some $i \in [N]$. Then conditioning on (s_i, s'_i) forces $s_j = s'_j = [(k-1)^K]$ for all $j > i$. Hence $E(G,G')$ must contain all the edges $(i, i+1), (i+1, i+2), \ldots, (N-1, N)$ and so

$$\mathbb{E}\left[\left|E(G,G')|_{\{i,i+1,\ldots,N\}}\right| \;\middle|\; (s_1,\ldots,s_i,s'_1,\ldots,s'_i)\right] = N - i + 1.$$

However working with $E_{\texttt{for}}(G,G')$ prevents such situations by halting exploration once either s_i or s'_i becomes too lexicographically late. This circumvents the above obstruction because the left-hand side of (2.3.7) is trivially 0 unless a lot of "space" in $[k]_0^K$ remains available for future strings $(s_{i+1},\ldots,s_N,s'_{i+1},\ldots,s'_N)$.

In fact, this guaranteed available space is also directly helpful in implementing the block decomposition strategy outlined above. Namely for any prefix $x \in [k]_0^M$, it ensures that the distribution for the number of strings $(s_{i+1},\ldots,s_N)$ starting with x cannot increase too much from conditioning on $(s_1,\ldots,s_i)$ (see Lemma 2.5.6 for a precise statement). This is important because when applying (2.3.8) to the block of strings starting with some prefix x, we replace N by the number of strings

starting with x (and also replace K by $K - M$). In short, we must ensure that the hypothesis of (2.3.8) holds within each block.

2.3.2 Proof of Lemmas 2.3.1 and 2.3.5

Here we prove Lemmas 2.3.1 and 2.3.5, thus reducing the proof of the mixing time upper bound (2.1.3) to establishing Lemma 2.3.8.

Proof of Lemma 2.3.1. Let $(v_1, \ldots, v_a)$ be the vertex-sizes of the G-components and $(w_1, \ldots, w_b)$ be the vertex-sizes of the G'-components.

We first claim that

$$f_{G,G'} = \frac{\left(\prod_{i=1}^{a} v_i!\right) \cdot \left(\prod_{j=1}^{b} w_j!\right)}{\prod_{\ell=1}^{c} u_\ell!}. \tag{2.3.9}$$

Indeed this follows by writing

$$\begin{aligned} f_{G,G'} &= \mathbb{E}^\sigma[f_{G,\sigma} f_{G',\sigma}] \\ &= \mathbb{E}^\sigma \left[1_{\sigma^G = \sigma} \cdot 1_{\sigma^{G'} = \sigma} \cdot \left(\prod_{i=1}^{a} v_i! \right) \cdot \left(\prod_{j=1}^{b} w_j! \right) \right] \\ &= \mathbb{E}^\sigma \left[1_{\sigma^U = \sigma} \right] \cdot \left(\prod_{i=1}^{a} v_i! \right) \cdot \left(\prod_{j=1}^{b} w_j! \right) \\ &= \frac{\left(\prod_{i=1}^{a} v_i!\right) \cdot \left(\prod_{j=1}^{b} w_j!\right)}{\prod_{\ell=1}^{c} u_\ell!}. \end{aligned}$$

Decomposing the product in (2.3.9) based on the components $U_i \in \mathcal{C}(U)$ implies

$$f_{G,G'} = \prod_{\ell} f_{G,G',U_\ell} \tag{2.3.10}$$

where we define

$$f_{G,G',U_\ell} \equiv \frac{\left(\prod_{i:G_i \subseteq U_\ell} v_i!\right) \cdot \left(\prod_{j:G'_j \subseteq U_\ell} w_j!\right)}{u_\ell!}.$$

Observe that in general, for any positive integers $m_1, \ldots, m_n, M$ with

$$\sum_{i=1}^{n} (m_i - 1) \leq M - 1,$$

one has $\prod_{i=1}^{n} m_i! \leq M!$. Indeed both sides can be written as a product of at most $M - 1$ integers at least 2, and the $M - 1$ numbers appearing in the product for $M!$ are clearly larger. In particular, this holds for $M = u_\ell$ whenever $m_1, \ldots, m_n$ are the vertex-sizes of edge-disjoint subinterval graphs

of $V(U_\ell)$. It directly implies

$$\prod_{i:G_i\subseteq U_\ell} v_i! \le u_\ell!,$$
$$\prod_{j:G'_j\subseteq U_\ell} w_j! \le u_\ell!$$

from which it follows that $f_{G,G',U_\ell} \le u_\ell!$ holds. Moreover if U_ℓ does not contain any edge in $E(G,G')$ then the G-components and G'-components are collectively edge-disjoint. Hence for such U_ℓ,

$$\left(\prod_{i:G_i\subseteq U_\ell} v_i!\right)\cdot\left(\prod_{j:G'_j\subseteq U_\ell} w_j!\right) \le u_\ell!$$

which implies $f_{G,G',U_\ell} \le 1$. Substituting these estimates into (2.3.10) implies (2.3.4). □

The next lemma is used to show Lemma 2.3.5.

Lemma 2.3.9. *For $K \ge (\widetilde{C}_{\mathbf{p}}+\varepsilon)\log(N)$, there is $\delta(\mathbf{p},\varepsilon) > 0$ so that the following holds. Conditioned on any initial strings $s_1, s_2, \ldots, s_i$, none of which begin with $[(k-1)(k-1)]$, the conditional probability that $s_i = s_{i+1}$ is at most $O(N^{-\delta})$.*

Proof. Recall the definitions (2.2.1), (2.2.2), (2.2.3) and the subsequent discussion. We use the sampling model of N i.i.d.-then-sorted uniform random variables, letting

$$0 \le a_1 \le a_2 \le \cdots \le a_N \le 1$$

be uniformly random before being sorted and then choosing s_j such that $a_j \in J_{s_j}$ for each $1 \le j \le N$.

Recall that we condition on s_i. Let us now condition further on the value $a_i \in J_{s_i}$. Then the remaining numbers a_j for $j > i$ are, up to sorting, conditionally i.i.d. in $[a_i, 1]$. The crucial observation is that the interval $[a_i, 1]$ has length lower bounded by $1 - a_i \ge p_{k-1}^2 \ge p_{\mathtt{min}}^2$. Indeed, $a_i < 1 - p_{k-1}^2$ is equivalent to the assumption that s_i does not begin with $[(k-1)(k-1)]$. (For instance, note that $J_{[(k-1)(k-1)]} = [1-p_{k-1}^2, 1)$.) Meanwhile the length of J_{s_i} is λ_{s_i}.

Combining these observations, it follows that the conditional distribution for the number of $j > i$ with $s_j = s_i$ is stochastically dominated by a $\mathtt{Bin}(N, p_{\mathtt{min}}^{-2}\lambda_{s_i})$ random variable, regardless of the value a_i. Averaging over the unknown a_i, the same stochastic domination holds conditioned on just $(s_1, \ldots, s_i)$. Since $K \ge (\widetilde{C}_{\mathbf{p}} + \varepsilon)\log(N)$ was assumed,

$$\lambda_{s_i} \le (p_{\mathtt{max}})^K \le N^{-1-\delta}.$$

The result now follows. □

Proof of Lemma 2.3.5. The regularity readily follows from Chernoff estimates so we focus only on the L-sparsity. First, Lemma 2.3.9 implies that $\mathbb{P}[s_{i+1} = s_i | (s_1, \ldots, s_i)] \leq O(N^{-\delta})$ whenever $s_i <_{\texttt{lex}} [(k-1)(k-1)]$. A simple Markovian coupling now implies that the set of edges formed by strings $s_i <_{\texttt{lex}} [(k-1)(k-1)]$ is stochastically dominated by instead choosing each edge independently with probability $O(N^{-\delta})$. By symmetry the same holds for edges formed by strings starting with $[(k-1)(k-1)]$. Call these ordinary edges and final edges, respectively.

A simple Chernoff bound implies that for $L \geq 1000\delta^{-1}$, each interval $\{i, i+1, \ldots, i+L-1\}$ of L consecutive vertices contains at most $L/6$ ordinary edges and at most $L/6$ final edges with probability $1 - O_L\left(\frac{1}{N^2}\right)$. Since $L/6 + L/6 = L/3$, union bounding over at most N such length-L intervals shows that L-sparsity holds with probability at least $1 - O(N^{-1}) \geq 1 - O(N^{-\delta})$. □

2.4 Upper Bounding the Expected Shared Edges

Define the constant

$$\underline{C}_{\mathbf{p}} \equiv \max\left(C_{\mathbf{p}}, \frac{1}{\log(1/p_0)}, \frac{1}{\log(1/p_{k-1})}\right) \leq \overline{C}_{\mathbf{p}}.$$

The purpose of this section is to prove the following crucial result.

Proposition 2.4.1. *For any $\varepsilon > 0$, if $K \geq \left(\underline{C}_{\mathbf{p}} + \varepsilon\right)\log(N)$ holds then*

$$\mathbb{E}\left[|E(G, G')|\right] \leq O(N^{-\Omega_{\mathbf{p}}(\varepsilon)}).$$

We eventually need to control the (truncated) *exponential* moments of $E(G, G')$. However Proposition 2.4.1 is the most involved part of upper-bounding the mixing time, and the value $C_{\mathbf{p}} = \frac{3+\theta_{\mathbf{p}}}{4\psi_{\mathbf{p}}(2)}$ emerges naturally in its proof. We note that for our main goal of establishing cutoff, proving Proposition 2.4.1 only for $K \geq \left(\overline{C}_{\mathbf{p}} + \varepsilon\right)\log(N)$ would suffice just as well. However there is no additional difficulty in proving Proposition 2.4.1 as stated. Moreover the case $\underline{C}_{\mathbf{p}} \neq \overline{C}_{\mathbf{p}}$ amounts to an interesting discrepancy between the first moment and exponential moment behavior of $|E(G, G')|$. See Remark 2.5.1 for more discussion of this point.

Let us mention that after further preparation in Subsection 2.4.1, we provide in Subsection 2.4.2 a proof outline for Proposition 2.4.1.

2.4.1 Preparation for the Upper Bound Proof

We now introduce several more technical definitions. As a convention, $\mathbf{p}$ and ε will be treated as fixed, while $\delta = \delta(\mathbf{p}, \varepsilon)$ will be taken sufficiently small. As before G and G' will always be independent $\mathbf{p}$-random shuffle graphs. Moreover s will denote strings of length K while x will denote strings of

arbitrary length $M \leq K$.

Lexicographic Subintervals and Blocks

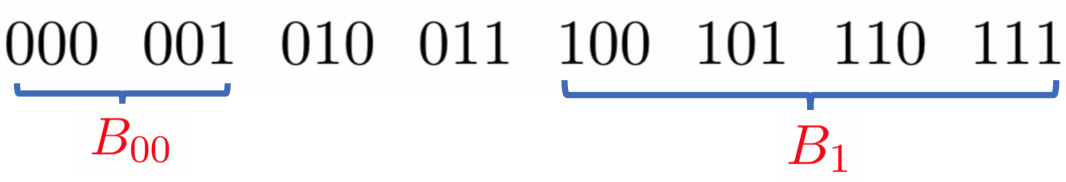


Figure 2.4: The blocks B_{00} and B_1 are shown for $k = 2$ and $K = 3$.

For a string x of length M, define its *block* $B_x \subseteq [k]_0^K$ to be the set of strings of length K beginning with x. Hence B_x consists of k^{K-M} strings. Given a lexicographically sorted sequence $(s_1, \ldots, s_N) \in \mathcal{S}$ of strings, define the discrete interval $\mathcal{I}(B_x) \subseteq [N]$ by

$$\mathcal{I}(B_x) \equiv \{i \in [N] : s_i \in B_x\} = \{\iota(x), \iota(x) + 1, \ldots, \tau(x)\}.$$

In general, we define

$$\iota(x) = |\{i \in [N] : s_i <_{\texttt{lex}} x\}| + 1, \qquad \tau(x) = |\{i \in [N] : s_i <_{\texttt{lex}} x \text{ or } s_i \in B_x\}|.$$

This ensures $|\mathcal{I}(B_x)| = \tau(x) - \iota(x) + 1$ even if $\mathcal{I}(B_x)$ is empty. Observe that for fixed x (recall the definitions (2.2.1) and (2.2.2)),

$$|\mathcal{I}(B_x)| \sim \texttt{Bin}(N, \lambda_x), \tag{2.4.1}$$

$$\iota(x) \sim \texttt{Bin}(N, t_x) + 1, \tag{2.4.2}$$

$$\tau(x) \sim \texttt{Bin}(N, t_x + \lambda_x). \tag{2.4.3}$$

Finally define G_{B_x} to be the induced subgraph of G with vertex set $\mathcal{I}(B_x)$, which retains the edges $(i, i+1) \in E(G)$ such that $s_i = s_{i+1} \in B_x$. Denote its edge set by $E(G_{B_x})$.

Entropy

We will require the entropy function. Given any k-tuple of non-negative real numbers $(a_0, \ldots, a_{k-1})$ with sum $a_{\mathtt{tot}}$, let

$$H(a_0, \ldots, a_{k-1}) = \frac{\sum_{i=0}^{k-1} a_i \log\left(\frac{a_{\mathtt{tot}}}{a_i}\right)}{a_{\mathtt{tot}}}$$

be the entropy of the discrete probability distribution with weights $(a_i/a_{\mathtt{tot}})_{i=0}^{k-1}$. If $a_0 = \cdots = a_{k-1} = 0$ then set $H(a_0, \ldots, a_{k-1}) = 0$. The following result allows approximation of multinomial coefficients using entropy. (The values $a_i \log(N)$ correspond to the normalization in Definition 2.4.3 just below.)

Proposition 2.4.2 ([CS04, Lemma 2.2])**.** *Fix $A \geq 0$ and any non-negative real numbers $a_0, \ldots, a_{k-1} \in [0, A]$ satisfying $a_i \log(N) \in \mathbb{Z}$,*

$$N^{a_{\mathtt{tot}} H(a_0,\ldots,a_{k-1}) - o_N(1)} \leq \binom{a_{\mathtt{tot}} \log(N)}{a_0 \log(N), \ldots, a_{k-1} \log(N)} \leq N^{a_{\mathtt{tot}} H(a_0,\ldots,a_{k-1})}.$$

Here the term $o_N(1)$ tends to 0 for any fixed A as $N \to \infty$, uniformly in the values $a_0, \ldots, a_{k-1} \in [0, A]$.

The following special definitions will also be convenient. For $t > 0$, let $\mathbf{p}^t$ be the probability distribution on $[k]_0$ given by $(\mathbf{p}^t)_i = \frac{p_i^t}{\phi_{\mathbf{p}}(t)}$. Define

$$I(\mathbf{p}, \mathbf{p}^t) \equiv D_{\mathtt{KL}}(\mathbf{p}^t \,||\, \mathbf{p}) + H(\mathbf{p}^t) = \sum_{i=0}^{k-1} (\mathbf{p}^t)_i \log(1/p_i) = \sum_{i=0}^{k-1} \frac{p_i^t \log(1/p_i)}{\phi_{\mathbf{p}}(t)} > 0. \tag{2.4.4}$$

It is not difficult to verify the identity

$$H(\mathbf{p}^t) = t \cdot I(\mathbf{p}, \mathbf{p}^t) - \psi_{\mathbf{p}}(t), \quad t \in \mathbb{R}^+. \tag{2.4.5}$$

Digit Profile

Here we define several notions based on the *digit profile* of a string, which tracks how many of each digit a string contains, as well as initial digits of 0 or $k-1$.

Definition 2.4.3. *For a string $x \in [k]_0^M$, the **digit profile** of x is the $(k+2)$-tuple*

$$(b_0(x), b_{k-1}(x), c_0(x), \ldots, c_{k-1}(x)) \in (\mathbb{Z}/\log N)^{k+2}$$

of non-negative real numbers summing to $b_0 + b_{k-1} + \sum_{i=0}^{k-1} c_i = \frac{M}{\log(N)}$ defined as follows. $b_0 \log(N)$ is the number of initial 0-digits in x and $b_{k-1} \log(N)$ is the number of initial $(k-1)$-digits (so

$\min(b_0, b_{k-1}) = 0$). *After the first* $(b_0 + b_{k-1})\log(N)$ *digits,* x *contains* $c_i \log(N)$ *digits of* i *for each* $i \in [k]_0$.

The normalization $\frac{1}{\log N}$ above is taken so that the total sum $\frac{M}{\log N}$ is of constant order. We next define constants depending on the digit profile of x. Let

$$c_{\mathtt{tot}}(x) = \sum_{i=0}^{k-1} c_i(x)$$

be the number of digits in x after the initial 0 or initial $(k-1)$ digits. Also define

$$\begin{aligned}
c_L(x) &\equiv 1 - b_0 \log\left(\frac{1}{p_0}\right) - b_{k-1}\log\left(\frac{1}{p_{k-1}}\right) - \sum_{i=0}^{k-1} c_i \log\left(\frac{1}{p_i}\right) = 1 + \log_N(\lambda_x),\\
c_F(x) &\equiv \frac{1 - b_0 \log\left(\frac{1}{p_0}\right) - b_{k-1}\log\left(\frac{1}{p_{k-1}}\right)}{2},\\
c_D(x) &\equiv c_L(x) - c_F(x) = \frac{1 - b_0 \log\left(\frac{1}{p_0}\right) - b_{k-1}\log\left(\frac{1}{p_{k-1}}\right)}{2} - \sum_{i=0}^{k-1} c_i \log\left(\frac{1}{p_i}\right),\\
c_E(x) &\equiv \left(\frac{M-K}{\log N}\right)\psi_{\mathbf{p}}(2) = \left(b_0 + b_{k-1} + c_{\mathtt{tot}} - \frac{K}{\log N}\right)\psi_{\mathbf{p}}(2) < 0,\\
c_X(x) &\equiv c_{\mathtt{tot}} H(c_0, \ldots, c_{k-1}) + 5c_L - 2c_F + 2c_E.
\end{aligned}$$

Finally say x is δ-*stable* if

$$c_L(x) - c_F(x) \in [\delta, 2\delta]. \tag{2.4.6}$$

The typical size of $|\mathcal{I}(B_x)|$ is N^{c_L} while N^{c_F} is the order of fluctuations for $\iota(x)$ and $\tau(x)$. c_E is related to the typical number of G-edges coming from strings in B_x. c_X is related to the typical number of G-edges coming from strings of the same digit profile as x. Note that when $b_0 = b_{k-1} = 0$ we have $c_F = \frac{1}{2}$. As explained in the next subsection, this corresponds to $\iota(x)$ and $\tau(x)$ having fluctuations of order $N^{1/2}$.

2.4.2 Proof Outline for Proposition 2.4.1

We now outline the proof of Proposition 2.4.1. Except for the end of this outline we will only consider strings x with $b_0(x) = b_{k-1}(x) = 0$ so that the interval $J_x \in [0,1]$ is a constant distance from the boundary points $\{0,1\}$. We will take $\delta \ll \varepsilon$ to be a small constant, and simply write δ when a constant multiple such as 4δ would be technically correct. Since we are targeting an upper bound $O\left(N^{-\Omega_{\mathbf{p}}(\varepsilon)}\right)$ in Proposition 2.4.1, factors of $N^{O(\delta)}$ can usually be thought of as small.

The first idea is to start from the empty block $B_\emptyset = [k]_0^K$ and recursively refine the partition of

$[k]_0^K$ by decomposing a block B_x into k smaller blocks via

$$B_x = \bigcup_{i \in [k]_0} B_{xi}.$$

For example when $k = 2$ such a refinement might proceed as

$$B_\emptyset \to B_0 \cup B_1 \to B_{00} \cup B_{01} \cup B_1 = [2]_0^K.$$

We recursively refine the partition $B_\emptyset$ in this way until each block B_x in the partition has size $\mu_{\mathbf{p}}(B_x) \approx N^{-\frac{1}{2}+\delta}$; this is formally carried out in Lemma 2.4.5. The set of strings x used in the resulting partition is denoted by $\mathcal{L}_{\texttt{stable}}$, so that we obtain

$$[k]_0^K = \bigcup_{x \in \mathcal{L}_{\texttt{stable}}} B_x, \quad \text{and} \quad [N] = \bigcup_{x \in \mathcal{L}_{\texttt{stable}}} \mathcal{I}(B_x). \tag{2.4.7}$$

as in Lemma 2.4.6. The first and last indices $\iota(x)$ and $\tau(x)$ of $\mathcal{I}(B_x)$ are (non-independent) binomial random variables with N trials, hence each fluctuate by at most $O(N^{1/2})$ with high probability.

The upshot of the above is that the random set $\mathcal{I}(B_x)$ agrees with a discrete deterministic interval of size $|NJ_x \cap \mathbb{Z}| \approx N^{\frac{1}{2}+\delta}$ up to boundary fluctuations $|\iota(x) - Nt_x|$ and $|\tau(x) - N(t_x + \lambda_x)|$ which are smaller than $N^{\frac{1}{2}+\delta}$ with high probability. Because the random interval $\mathcal{I}(B_x)$ has typical size of larger order than the fluctuations of its left and right endpoints, we may think of $\mathcal{I}(B_x)$ as being nearly deterministic. In line with this intuition, we show in Lemma 2.4.10 that given any $i \in [N]$ there exist adjacent $x_{i,1}, x_{i,2} \in \mathcal{L}_{\texttt{stable}}$ such that $i \in \mathcal{I}(B_{x_{i,1}}) \cup \mathcal{I}(B_{x_{i,2}})$ holds with extremely high probability. Combining this with AM-GM, we show in Lemma 2.4.11 that $\mathbb{E}[|E(G, G')|]$ is upper bounded by the expected number of shared edges from pairs (G_{B_x}, G'_{B_x}) of matching blocks as follows.

$$\begin{aligned} \mathbb{E}[|E(G, G')|] &\lesssim \sum_{x \in \mathcal{L}_{\texttt{stable}}} \mathbb{E}[E(G_{B_x}, G'_{B_x})] \\ &= \sum_{x \in \mathcal{L}_{\texttt{stable}}} \sum_{i=1}^{N-1} \mathbb{P}[(i, i+1) \in E(G_{B_x})]^2. \end{aligned} \tag{2.4.8}$$

Here the informal notation $\lesssim$ hides a constant factor and a negligible additive term.

Our next objective is to upper-bound for each x the probability $\mathbb{P}[(i, i+1) \in E(G_{B_x})]$ appearing in (2.4.8). We do this by conditioning on the multiset S_x of strings appearing in $\mathcal{I}(B_x)$ and averaging over the still-random external strings. Although this conditioning determines the size and internal edge-structure of $\mathcal{I}(B_x)$, the *position* of $\mathcal{I}(B_x)$ is conditionally random. Indeed the position of the interval $\mathcal{I}(B_x)$ depends on the number of external strings lexicographically smaller than x, which we

have not conditioned on. This shift is conditionally binomial with order $N^{1/2}$ fluctuations. Crucially, these fluctuations "homogenize" the edge locations within each block B_x. Indeed averaging over these external shifts, it follows that

$$\max_{i\in[N-1]} \mathbb{P}[(i,i+1) \in E(G_{B_x})|S_x] \lesssim \frac{|E(G_{B_x})|}{N^{1/2}}. \tag{2.4.9}$$

It is not difficult to control the typical size $|E(G_{B_x})|$. Moreover since the location of $\mathcal{I}(B_x)$ is almost deterministic, the above probability is negligibly small for all but $O(\mathbb{E}[|\mathcal{I}(B_x)|]) = O(N^{\frac{1}{2}+\delta})$ values of i. Combining these considerations leads to an upper bound on

$$\sum_{i=1}^{N-1} \mathbb{P}[(i,i+1) \in E(G_{B_x})]^2. \tag{2.4.10}$$

The precise bound requires some care to state and is given in Lemma 2.4.14.

The preceding argument allowed us to estimate (2.4.10) for each x. In light of (2.4.8), it remains to sum over $x \in \mathcal{L}_{\texttt{stable}}$. The last key point is that all $x \in \mathcal{L}_{\texttt{stable}}$ with a given digit profile contribute essentially identically. Moreover there are only $\log(N)^{O(1)} \leq N^{o(1)}$ possible digit profiles. It therefore suffices to count the number of $x \in \mathcal{L}_{\texttt{stable}}$ with each digit profile and then determine the maximum total contribution of any fixed digit profile. This count is easily approximated using Proposition 2.4.2. The resulting maximum turns out to be achieved when x has digit frequencies approximately given by $\mathbf{p}^{\theta_{\mathbf{p}}}$, which leads to the appearance of the constant $C_{\mathbf{p}}$.

So far, this outline considered only blocks B_x with $b_0(x) = b_{k-1}(x) = 0$. When $b_0(x)$ or $b_{k-1}(x)$ is large the fluctuations of $\iota(x)$ and $\tau(x)$ shrink, simply because the variance $Np(1-p)$ of a $\texttt{Bin}(N,p)$ random variable shrinks when p is close to 0 or 1. To handle such prefixes x requires a slightly revised definition of $\mathcal{L}_{\texttt{stable}}$. In general the fluctuations of $\iota(x)$ and $\tau(x)$ should be slightly smaller than the typical size of $\mathcal{I}(B_x)$; this is precisely the definition of δ-stability in (2.4.6). It turns out that the resulting maximization problem over digit profiles nearly reduces to considering those with $b_0 = b_{k-1} = 0$. Indeed by an elementary linearity argument (see (2.4.18)), the only other digit profiles that must be considered are the degenerate cases with $c_0 = c_1 = \cdots = c_{k-1} = 0$ in which x consists of all 0 digits or all $(k-1)$ digits. These cases are much simpler and lead to the requirement that

$$\underline{C}_{\mathbf{p}} \geq \max\left(\frac{1}{\log(1/p_0)}, \frac{1}{\log(1/p_{k-1})}\right).$$

During a first reading of the next subsection it may be easier to focus on the main case $b_0 = b_{k-1} = 0$ so that the proofs match the outline above more closely.

Finally, we remark that the estimates outlined after (2.4.8) lead to the inequality

$$\mathbb{E}[|E(G,G')|] \lesssim N^{O(\delta)} \sum_{x \in \mathcal{L}_{\texttt{stable}}} \frac{\mathbb{E}[|E(G_{B_x})|]^2}{\mathbb{E}[|\mathcal{I}(B_x)|]}.$$

Hence for the purpose of counting edges in $E(G,G')$, each block B_x behaves approximately like an i.i.d. point process of edges in $\mathcal{I}(B_x)$ with x-dependent edge probability $\frac{\mathbb{E}[|E(G_{B_x})|]}{\mathbb{E}[|\mathcal{I}(B_x)|]}$. In fact (2.4.9) states that this holds more precisely at the level of individual edge probabilities. These hold precisely because the boundary fluctuations of $\mathcal{I}(B_x)$ are only slightly smaller than $\mathbb{E}[|\mathcal{I}(B_x)|]$, so that the homogenizing effect of the random shifts is near-total. Somewhat fancifully, one might then view the partition (2.4.7) as analogous to an ergodic or pure state decomposition.

2.4.3 The Partition into Stable Blocks

We now turn to a tree-based partition of $[k]_0^K$ into blocks B_x. Define the k-ary rooted tree $\mathcal{T} = \mathcal{T}_{k,K}$ of depth K which consists of all $[k]_0$-strings of length M at each level $0 \le M \le K$. Here the children of $s \in [k]_0^M$ are the concatenations $s0, s1, \ldots, s(k-1) \in [k]_0^{M+1}$. Hence the leaves of $\mathcal{T}$ are given by $[k]_0^K$ while the root of $\mathcal{T}$ is the empty string $\emptyset$. Recall from Subsection 2.4.1 the definition

$$c_D(x) = c_L(x) - c_F(x) = \frac{1 - b_0 \log\left(\frac{1}{p_0}\right) - b_{k-1}\log\left(\frac{1}{p_{k-1}}\right)}{2} - \sum_{i=0}^{k-1} c_i \log\left(\frac{1}{p_i}\right).$$

Moreover, as in Proposition 2.4.1, we assume throughout that $K \ge (\underline{C}_{\mathbf{p}} + \varepsilon)\log(N)$ holds for some small, fixed $\varepsilon > 0$.

Lemma 2.4.4. *Let x be the parent of y in $\mathcal{T}$. Then*

$$0 \le c_D(x) - c_D(y) \le O\left(\frac{1}{\log(N)}\right). \tag{2.4.11}$$

Moreover c_D takes value $c_D(\emptyset) = \frac{1}{2}$ at the root, while $c_D(s) \le -\Omega_{\mathbf{p}}(\varepsilon)$ for any leaf $s \in [k]_0^K$.

Proof. The values $b_0, b_{k-1}, c_0, \ldots, c_{k-1}$ each change by $O(1/\log(N))$ between neighboring vertices in $\mathcal{T}$, which shows that

$$|c_D(x) - c_D(y)| \le O\left(\frac{1}{\log(N)}\right).$$

Moreover since c_D is decreasing in each coordinate of the digit profile it follows that $c_D(x) - c_D(y) \ge 0$. This concludes the proof of (2.4.11).

When $x = \emptyset$ is the root, $b_0 = b_{k-1} = c_0 = \cdots = c_{k-1} = 0$, and so $c_D(\emptyset) = \frac{1}{2}$. Finally for any leaf

$s \in [k]_0^K$ of $\mathcal{T}$ we have

$$b_0(s) + b_{k-1}(s) + \sum_{i=0}^{k-1} c_i(s) = K \geq \underline{C}_{\mathbf{p}} + \varepsilon.$$

Since $t \to \log(\frac{1}{t})$ is decreasing and positive for $t \in (0,1)$,

$$\begin{aligned} c_L(s) - c_F(s) &= \frac{1}{2} - b_0 \cdot \frac{\log\left(\frac{1}{p_0}\right)}{2} - b_{k-1} \cdot \frac{\log\left(\frac{1}{p_{k-1}}\right)}{2} - \sum_{i=0}^{k-1} c_i \log\left(\frac{1}{p_i}\right) \\ &\leq \frac{1}{2} - \frac{(\underline{C}_{\mathbf{p}} + \varepsilon)\min(\log(1/p_0), \log(1/p_{k-1}), 2\log(1/p_{\mathtt{max}}))}{2} \end{aligned}$$

By definition $\underline{C}_{\mathbf{p}} \log(1/p_0) \geq 1$ and $\underline{C}_{\mathbf{p}} \log(1/p_{k-1}) \geq 1$. Moreover Proposition 2.1.1 implies

$$2\underline{C}_{\mathbf{p}} \log(1/p_{\mathtt{max}}) \geq \frac{2C_{\mathbf{p}}}{\widetilde{C}_{\mathbf{p}}} \geq 1.$$

Combining yields

$$\underline{C}_{\mathbf{p}} \cdot \min(\log(1/p_0), \log(1/p_{k-1}), 2\log(1/p_{\mathtt{max}})) \geq 1$$

which implies the result. □

Define the subtree $\mathcal{T}_{\mathtt{stable}} \subseteq \mathcal{T}$ to consist of all $x \in \mathcal{T}$ with $c_D(x) \geq 2\delta$, as well as all children of such x. Let $\mathcal{L}_{\mathtt{stable}}$ denote the set of leaves of $\mathcal{T}_{\mathtt{stable}}$. We say a finite rooted tree is a *full k-ary tree* if all of its vertices have either 0 or k children.

Lemma 2.4.5. *$\mathcal{T}_{\mathtt{stable}}$ is a full k-ary tree. Moreover $\mathcal{L}_{\mathtt{stable}}$ consists entirely of δ-stable strings. Finally all $x \in \mathcal{L}_{\mathtt{stable}}$ are strings of length in $[\Omega_{\mathbf{p},\delta}(\log N), K - \Omega_{\mathbf{p},\delta}(\log N)]$ and satisfy*

$$c_F(x) \geq \delta \quad \text{and} \quad c_L(x) \geq 2\delta.$$

Proof. First we explain why $\mathcal{T}_{\mathtt{stable}}$ is a full k-ary tree. The point is that since $c_D(x)$ is decreasing down $\mathcal{T}$ by Lemma 2.4.4, the set of strings x with $c_D(x) \geq 2\delta$ forms a subtree, and adding all children of such x therefore yields a full k-ary subtree.

Next, Lemma 2.4.4 shows $c_D(\emptyset) = \frac{1}{2}$ while $c_D(s) \leq -\Omega_{\mathbf{p}}(\varepsilon)$ for any s of length K, and also shows c_D has Lipschitz constant $O\left(\frac{1}{\log(N)}\right)$ on $\mathcal{T}$. It follows that $\mathcal{T}_{\mathtt{stable}}$ contains all of the first $\Omega(\log(N))$ levels of $\mathcal{T}$ but none of the last $\Omega(\log(N))$. As a result all $x \in \mathcal{L}_{\mathtt{stable}}$ have length in

$$[\Omega_{\mathbf{p},\delta}(\log(N)), K - \Omega_{\mathbf{p},\delta}(\log(N))].$$

The fact that all leaves are δ-stable holds because children were added in the definition of $\mathcal{T}_{\mathtt{stable}}$.

Indeed this definition combined with (2.4.11) implies that

$$c_D(x) \in [2\delta - O(1/\log N), 2\delta]$$

for all $x \in \mathcal{L}_{\texttt{stable}}$. Moreover recalling the definition (2.4.6), all $x \in \mathcal{L}_{\texttt{stable}}$ satisfy

$$c_F(x) + \delta \leq c_L(x).$$

Finally the inequality $c_L(x) \leq 2c_F(x)$ holds for any string x. Altogether, these inequalities imply $c_F(x) \geq \delta$ and therefore

$$c_L(x) \geq c_F(x) + \delta \geq 2\delta.$$

□

Lemma 2.4.6. *The following partitions (i.e. disjoint unions) hold:*

$$[k]_0^K = \bigcup_{x \in \mathcal{L}_{\texttt{stable}}} B_x \quad \text{and} \quad [N] = \bigcup_{x \in \mathcal{L}_{\texttt{stable}}} \mathcal{I}(B_x). \tag{2.4.12}$$

Proof. The first partition clearly implies the second. The first partition holds because $\mathcal{L}_{\texttt{stable}}$ consists of the leaves of $\mathcal{T}_{\texttt{stable}}$ and $\mathcal{T}_{\texttt{stable}} \subseteq \mathcal{T}$ is a full k-ary subtree. Indeed, it simply asserts that the subtrees of $\mathcal{T}$ rooted at each $x \in \mathcal{L}_{\texttt{stable}}$ partition the leaves of $\mathcal{T}$. □

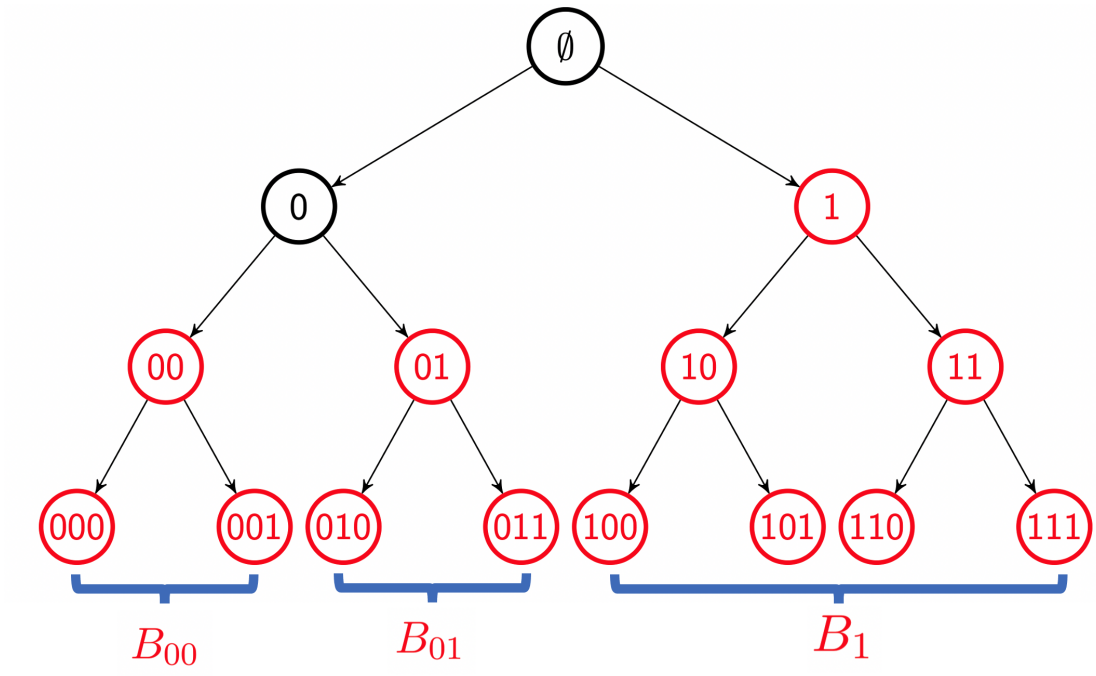


Figure 2.5: The first partition in Lemma 2.4.6 is shown in the case that $\mathcal{L}_{\texttt{stable}} = \{B_{00}, B_{01}, B_1\}$ with $(k, K) = (2, 3)$. It states that $[k]_0^K = [2]_0^3 = B_{00} \cup B_{01} \cup B_1$.

2.4.4 No Edge Intersections in Expectation

In this subsection we prove Proposition 2.4.1. As explained in the outline, the idea is to estimate $\mathbb{E}[|E(G, G')|]$ by a sum of individual contributions from each $x \in \mathcal{L}_{\texttt{stable}}$ and then control the total contribution from each digit profile.

Lemma 2.4.7. *Let $X \sim \texttt{Bin}(N, q)$ for some $q \in [0, 1]$. Then for $t \leq \sqrt{Nq(1-q)}$,*

$$\mathbb{P}\left[\left|X - \mathbb{E}[X]\right| \geq t\sqrt{Nq(1-q)}\right] \leq e^{-\Omega(t^2)}$$

holds uniformly over q.

Proof. This follows from Bernstein's inequality; see for instance [BLM13, Inequality (2.10)]. □

Lemma 2.4.8. *For any $x \in [k]_0^M$, either*

$$\min(t_x, 1 - t_x) = 0$$

or

$$\min(t_x, 1-t_x) \asymp_{\mathbf{p}} N^{-1+2c_F(x)}$$

holds. The same holds for $\min(t_x+\lambda_x, 1-t_x-\lambda_x)$. *Here* $\asymp_{\mathbf{p}}$ *denotes asymptotic equality for large* N *up to* $\mathbf{p}$*-dependent constant factors.*

Proof. We focus on $\min(t_x, 1-t_x)$ (as the two statements are symmetric) and assume x has a digit $x[i] \neq 0$ so that $t_x \neq 0$. If $x[1] = 0$ and $i > 1$ is minimal with $x[i] \neq 0$, then $b_0(x)\log(N) = i-1$ and so

$$t_x \asymp_{\mathbf{p}} p_0^{b_0(x)\log(N)} = N^{-1+2c_F}.$$

Similarly if $x[1] > 0$ and $i' > 1$ is minimal with $x[i'] \neq (k-1)$, then

$$1 - t_x - \lambda_x \asymp_{\mathbf{p}} p_{k-1}^{b_{k-1}(x)\log(N)} = N^{-1+2c_F}.$$

□

Lemma 2.4.9. *Let* $x \in \mathcal{L}_{\texttt{stable}}$ *have digit profile* $(b_0, b_{k-1}, c_0, \ldots, c_{k-1})$. *Then*

$$\mathbb{P}\left[\left||\mathcal{I}(B_x)| - N^{c_L}\right| \geq N^{\frac{c_L+\delta}{2}}\right] \leq e^{-\Omega(N^\delta)}, \tag{2.4.13}$$

$$\mathbb{P}\left[\left|\iota(x) - Nt_x]\right| \geq N^{c_F+\frac{\delta}{2}}\right] \leq e^{-\Omega(N^\delta)} \tag{2.4.14}$$

$$\mathbb{P}\left[\left|\tau(x) - N(t_x+\lambda_x)\right| \geq N^{c_F+\frac{\delta}{2}}\right] \leq e^{-\Omega(N^\delta)}. \tag{2.4.15}$$

Proof. First, inequality (2.4.13) follows immediately from (2.4.1), by applying Lemma 2.4.7 with $t = N^{\delta/2}$.

For inequality (2.4.14) we similarly recall the distribution of ι given by (2.4.2). From Lemma 2.4.8 it follows that unless $t_x = 0$ (in which case $\iota(x) = 1$ with probability 1),

$$\min(t_x, 1-t_x) \asymp N^{-1+2c_F}.$$

Then Lemma 2.4.7 with $t = N^{\delta/2}$ completes the proof of (2.4.14) as $\frac{\delta}{2} < \min(\frac{c_L}{2}, c_F)$ by Lemma 2.4.5. Inequality (2.4.15) is proved identically. □

The next lemma shows that for any $i \in [N]$, there are at most two blocks $B_{x_{i,1}}, B_{x_{i,2}}$ that i could plausibly appear in.

Lemma 2.4.10. *For each index* $i \in [N]$, *there exist* $x_{i,1}, x_{i,2} \in \mathcal{L}_{\texttt{stable}}$ *with*

$$\mathbb{P}[i \in \mathcal{I}(B_{x_{i,1}}) \cup \mathcal{I}(B_{x_{i,2}})] \geq 1 - e^{-\Omega(N^\delta)}.$$

Proof. Choose $x_{i,1} \in \mathcal{L}_{\texttt{stable}}$ so that $\frac{i}{N} \in J_x = [t_{x_{i,1}}, t_{x_{i,1}} + \lambda_{x_{i,1}})$, and without loss of generality assume

$$\frac{i}{N} \in \left[t_{x_{i,1}} + \frac{\lambda_{x_{i,1}}}{2}, t_{x_{i,1}} + \lambda_{x_{i,1}} \right).$$

Then we obtain

$$\begin{aligned} \iota(x_{i,1}) &\leq N t_{x_{i,1}} + \left| \iota(x_{i,1}) - N t_{x_{i,1}} \right| \\ &\leq i - \frac{N\lambda_{x_{i,1}}}{2} + \left| \iota(x_{i,1}) - N t_{x_{i,1}} \right|. \end{aligned}$$

As

$$N\lambda_{x_{i,1}} = N^{c_L(x_{i,1})} \geq N^{c_F(x_{i,1})+\delta},$$

using inequality (2.4.14) implies that

$$\mathbb{P}[\iota(x_{i,1}) \leq i] \geq 1 - e^{-\Omega(N^\delta)}$$

If $x_{i,1}$ is the lexicographically last element of $\mathcal{L}_{\texttt{stable}}$ then $\iota(x_{i,1}) \leq i$ already implies $i \in \mathcal{I}(B_{x_{i,1}})$. Otherwise using Lemma 2.4.6 we take $x_{i,2} \in \mathcal{L}_{\texttt{stable}}$ immediately lexicographically following $x_{i,1}$, so that $t_{x_{i,1}} + \lambda_{x_{i,1}} = t_{x_{i,2}}$. Reasoning identically to the above shows that

$$\mathbb{P}[\tau(x_{i,2}) \geq i] \geq 1 - e^{-\Omega(N^\delta)}.$$

If $\iota(x_{i,1}) \leq i \leq \tau(x_{i,2})$, then $i \in \mathcal{I}(B_{x_{i,1}}) \cup \mathcal{I}(B_{x_{i,2}})$ holds because $x_{i,1}$ and $x_{i,2}$ are consecutive in $\mathcal{L}_{\texttt{stable}}$. The result follows. □

Based on the previous lemma, we now upper-bound $\mathbb{E}[|E(G, G')|]$ by a sum over the individual blocks B_x. Recall that $E(G_{B_x}) \subseteq E(G)$ is the set of edges $(i, i+1) \in E(G)$ coming from strings $s_i = s_{i+1} \in B_x$.

Lemma 2.4.11. $\mathbb{E}[|E(G, G')|] \leq e^{-\Omega(N^\delta)} + 4 \sum_{x \in \mathcal{L}_{\texttt{stable}}} \sum_{i=1}^{N-1} \mathbb{P}[(i, i+1) \in E(G_{B_x})]^2$.

Proof. Lemma 2.4.10 and the AM-GM inequality imply

$$\begin{aligned}\mathbb{E}[|E(G,G')|] &\le \sum_{i=1}^{N-1} \mathbb{P}[(i,i+1)\in E(G,G')] \\ &\le e^{-\Omega(N^\delta)} + \sum_{i=1}^{N-1}\sum_{j_1,j_2\in\{1,2\}} \mathbb{P}[(i,i+1)\in E(G_{B_{x_{i,j_1}}}, G_{B_{x_{i,j_2}}})] \\ &\le e^{-\Omega(N^\delta)} + 2\sum_{i=1}^{N-1}\sum_{j\in\{1,2\}} \mathbb{P}[(i,i+1)\in E(G_{B_{x_{i,j}}})]^2 \\ &\le e^{-\Omega(N^\delta)} + 4\sum_{x\in\mathcal{L}_{\texttt{stable}}}\sum_{i=1}^{N-1} \mathbb{P}[(i,i+1)\in E(G_{B_x})]^2.\end{aligned}$$

□

The next lemma uses the quantity

$$c_E(x) = \left(\frac{M(x)-K}{\log N}\right)\psi_{\mathbf{p}}(2) = \left(b_0(x)+b_{k-1}(x)+c_{\texttt{tot}}(x)-\frac{K}{\log N}\right)\psi_{\mathbf{p}}(2) < 0$$

defined near the end of Subsubsection 2.4.1.

Lemma 2.4.12. *For any $x\in\mathcal{T}$,*

$$\mathbb{E}\left[|E(G_{B_x})| \;\middle|\; |\mathcal{I}(B_x)|\right] \le |\mathcal{I}(B_x)|^2 N^{c_E(x)}.$$

Proof. The right-hand side upper-bounds the expected number of pairs (i,j) with $s_i = s_j$ and $i,j\in\mathcal{I}(B_x)$, by summing over the $|\mathcal{I}(B_x)|^2$ pairs of pre-sorted strings in B_x. Indeed it is easy to see that for independent $\mu_{\mathbf{p},K}$-random strings s and s', and fixed $x\in[k]_0^M$,

$$\mathbb{P}[s = s' | s,s'\in B_x] = \phi_{\mathbf{p}}(2)^{-(K-M)} = N^{c_E(x)}.$$

□

The following lemma upper-bounds the probability for an edge $(i,i+1)$ to appear in $E(G_{B_x})$ as a function of x, uniformly over $i\in[N]$. The idea is that even conditioned on the value $|\mathcal{I}(B_x)|$ and the internal structure of $\mathcal{I}(B_x)$, the remaining randomness of the value $\iota(x)$ has a "homogenizing" effect.

Lemma 2.4.13. *For any $x\in\mathcal{L}_{\texttt{stable}}$ and index $i\in[N-1]$,*

$$\mathbb{P}[(i,i+1)\in E(G_{B_x})] \le 4N^{2c_L(x)-c_F(x)+c_E(x)+2\delta} + e^{-\Omega(N^\delta)}.$$

Proof. We condition on the multiset of strings $S_x \equiv [s_j | s_j \in B_x]$ appearing in B_x. We will show that if

$$|\mathcal{I}(B_x)| \le 2N^{c_L} \le N/2 \tag{2.4.16}$$

holds, then

$$\mathbb{P}[(i, i+1) \in E(G_{B_x}) | S_x] \le 4N^{2c_L(x) - c_F(x) + c_E(x) + 2\delta}.$$

This implies the desired result since by inequality (2.4.13),

$$\mathbb{P}[|\mathcal{I}(B_x)| \le 2N^{c_L}] \ge 1 - e^{-\Omega(N^\delta)}.$$

Observe that the multiset S_x determines the values $|E(G_{B_x})|$ and $|\mathcal{I}(B_x)| = |S_x|$, and in fact determines the entire set $E(G_{B_x})$ up to shifts. Given S_x, it is easy to see that $\iota(x)$ has conditional law

$$\iota(x) \sim \texttt{Bin}\left(N - |\mathcal{I}(B_x)|, \frac{t_x}{1-\lambda_x}\right) + 1.$$

From (2.4.16), we have $N - |\mathcal{I}(B_x)| \ge N/2$. Because any $x \in \mathcal{L}_{\texttt{stable}}$ has length $\Omega(\log(N))$ by Lemma 2.4.5, it follows that $\lambda_x \le \frac{1}{2}$ for all $x \in \mathcal{L}_{\texttt{stable}}$ when N is large enough. Therefore Lemma 2.4.8 gives $t_x = 0$ or $t_x \ge \Omega(N^{-1+2c_F})$. Similarly

$$1 - \frac{t_x}{1-\lambda_x} = \frac{1 - t_x - \lambda_x}{1 - \lambda_x} \ge \Omega(N^{-1+2c_F})$$

unless $1 - t_x - \lambda_x = 0$.

Let us now split into two cases, the first being that

$$\min(t_x, 1 - t_x - \lambda_x) > 0.$$

In this case we conclude that $\iota(x) - 1$ is binomial with number of trials $N - |\mathcal{I}(B_x)| \ge N/2$ and total variance $\Omega(N^{2c_F})$. Recalling that $c_F(x) \ge \delta$ for $x \in \mathcal{L}_{\texttt{stable}}$, the Lindeberg condition implies that conditionally on S_x, $\iota(x)$ satisfies a central limit theorem with standard deviation $\Omega\left(N^{c_F(x)}\right)$. Since $\iota(x) - 1$ is binomial, this implies a pointwise bound on its probability mass function. Explicitly, we may apply either [Pit97, Equation 25] or the combination of [Pit97, Equation 24] and [Can80, Theorem B] to obtain

$$\max_{j \in [N]} \mathbb{P}[\iota(x) = j | S_x] \le N^{-c_F(x) + 2\delta}. \tag{2.4.17}$$

Next in the second case, assume that

$$\min(t_x, 1 - t_x - \lambda_x) = 0.$$

This simply means that x consists of all digits 0 or all digits $(k-1)$. Then $c_{\texttt{tot}}(x) = 0$ and so

$c_L = 2c_F \leq c_F + 2\delta$ implies $c_F \leq 2\delta$. Hence (2.4.17) holds in either case. As a result for any $i \in [N-1]$,

$$\begin{aligned}\mathbb{P}\left[(i, i+1) \in E(G_{B_x}) \mid S_x\right] &\leq |E(G_{B_x})| \cdot \max_{j \in [N]} \mathbb{P}\left[\iota(x) = j | S_x\right] \\ &\leq |E(G_{B_x})| \cdot N^{-c_F(x)+2\delta}.\end{aligned}$$

Applying Lemma 2.4.12 shows that when (2.4.16) holds,

$$\mathbb{P}\left[(i, i+1) \in E(G_{B_x}) \mid S_x\right] \leq 4N^{2c_L(x)-c_F(x)+c_E(x)+2\delta}.$$

□

Using Lemma 2.4.13, we can estimate each term appearing in Lemma 2.4.11.

Lemma 2.4.14. *For any $x \in \mathcal{L}_{\texttt{stable}}$,*

$$\sum_{i=1}^{N-1} \mathbb{P}[(i, i+1) \in E(G_{B_x})]^2 \leq 64N^{5c_L(x)-2c_F(x)+2c_E(x)+4\delta} + e^{-\Omega(N^\delta)}.$$

Proof. For those $i \in [N]$ with

$$i \in \left[Nt_x - N^{c_F+\frac{\delta}{2}}, N(t_x + \lambda_x) + N^{c_F+\frac{\delta}{2}}\right],$$

Lemma 2.4.13 implies

$$\mathbb{P}[(i, i+1) \in E(G_{B_x})] \leq 4N^{2c_L(x)-c_F(x)+c_E(x)+2\delta} + e^{-\Omega(N^\delta)}.$$

As $c_F + \frac{\delta}{2} \leq c_L - \frac{\delta}{2}$, the above applies to at most $2N^{c_L}$ values of i. For all other $i \in [N-1]$, inequalities (2.4.14) and (2.4.15) imply $\mathbb{P}[(i, i+1) \in E(G_{B_x})] \leq e^{-\Omega(N^\delta)}$. Combining and using $(a+b)^2 \leq 2a^2 + 2b^2$ yields

$$\begin{aligned}\sum_{i=1}^{N-1} \mathbb{P}[(i, i+1) \in E(G_{B_x})]^2 &\leq 2N^{c_L}\left(4N^{2c_L(x)-c_F(x)+c_E(x)+2\delta} + e^{-\Omega(N^\delta)}\right)^2 + Ne^{-\Omega(N^\delta)} \\ &\leq 64N^{5c_L(x)-2c_F(x)+2c_E(x)+4\delta} + e^{-\Omega(N^\delta)}.\end{aligned}$$

□

Having controlled the individual summands in Lemma 2.4.11 in terms of the digit profile of x, it remains to sum over $x \in \mathcal{L}_{\texttt{stable}}$. This amounts to determining the number of $x \in \mathcal{L}_{\texttt{stable}}$ with each possible digit profile, and then finding the maximum possible contribution of each digit profile.

Recalling the definition

$$c_X = c_{\mathtt{tot}} H(c_0, \dots, c_{k-1}) + 5c_L - 2c_F + 2c_E,$$

it follows from Lemma 2.4.14 and Proposition 2.4.2 that the contribution of a given digit profile x to the bound of Lemma 2.4.11 is roughly $N^{c_X(x)}$. The next lemma shows that $c_X(x)$ is uniformly negative over $x \in \mathcal{L}_{\mathtt{stable}}$ when $K \geq (\underline{C}_{\mathbf{p}} + \varepsilon)\log(N)$. Here we give a concise proof which does not provide much intuition for the constants $\theta_{\mathbf{p}}$ and $C_{\mathbf{p}}$. See Subsection 2.4.5 for another argument which is longer and less formal but probably more enlightening.

Lemma 2.4.15. *For $\delta = \delta(\mathbf{p}, \varepsilon)$ small enough, if $K \geq (\underline{C}_{\mathbf{p}} + \varepsilon)\log(N)$ then*

$$\max_{(b_0, b_{k-1}, c_0, \dots, c_{k-1}) \ \delta\text{-stable}} c_X(b_0, b_{k-1}, c_0, \dots, c_{k-1}) \leq -\Omega_{\mathbf{p}}(\varepsilon) < 0.$$

Proof. Let us extend the definitions of $c_{\mathtt{tot}}, c_F, c_L, c_E$, and c_X to be functions of arbitrary $(k+2)$-tuples $(b_0, b_{k-1}, c_0, \dots, c_{k-1}) \in (\mathbb{R}^+)^{k+2}$ which are constrained to satisfy $\min(b_0, b_{k-1}) = 0$. Having done this, we observe that $c_X = c_X(b_0, b_{k-1}, c_0, \dots, c_{k-1})$ is affine in t along the paths

$$t \in \mathbb{R} \to ((1 - t\alpha_{\mathbf{p}})\, b_0, (1 - t\alpha_{\mathbf{p}})\, b_{k-1}, (1+t)c_0, \dots, (1+t)c_{k-1}) \tag{2.4.18}$$

where $\alpha_{\mathbf{p}} \geq 0$ is chosen so that $c_L - c_F$ remains constant as t varies.

Therefore to conclude we only need to show $c_X \leq -\Omega(\varepsilon)$ at the endpoint cases, which take the forms $(b_0, b_{k-1}, 0, \dots, 0)$ and $(0, 0, c_0, \dots, c_{k-1})$ and which continue to satisfy $c_L - c_F \in [\delta, 2\delta]$. As either $b_0 = 0$ or $b_{k-1} = 0$, we assume without loss of generality that $b_{k-1} = 0$. In the case $(b_0, 0, \dots, 0)$, we get

$$\begin{aligned} c_X(b_0, 0, \dots, 0) &= 5 - 5b_0 \log\left(\frac{1}{p_0}\right) - 1 + b_0 \log\left(\frac{1}{p_0}\right) + 2\left(b_0 - \frac{K}{\log(N)}\right)\psi_{\mathbf{p}}(2) + 2\delta \\ &= 4\left(1 - b_0 \log\left(\frac{1}{p_0}\right)\right) + 2\left(b_0 - \frac{K}{\log(N)}\right)\psi_{\mathbf{p}}(2) + 2\delta \end{aligned}$$

From $c_L - c_F \in [\delta, 2\delta]$ we obtain

$$c_L - c_F = \frac{1 - b_0 \log\left(\frac{1}{p_0}\right)}{2} \in [\delta, 2\delta]$$

and so

$$b_0 \log\left(\frac{1}{p_0}\right) \in [1 - 4\delta, 1 - 2\delta].$$

Using also that

$$\frac{K}{\log N} \geq \underline{C}_{\mathbf{p}} + \varepsilon \geq \frac{1}{\log(1/p_0)} + \varepsilon,$$

we find

$$\begin{aligned} c_X(b_0, 0, \ldots, 0) &\leq 8\delta + 2\left(\frac{1-2\delta}{\log\left(\frac{1}{p_0}\right)} - \frac{1+\varepsilon}{\log\left(\frac{1}{p_0}\right)}\right)\psi_{\mathbf{p}}(2) + 2\delta \\ &\leq -\Omega_{\mathbf{p}}(\varepsilon) + 10\delta \\ &\leq -\Omega_{\mathbf{p}}(\varepsilon). \end{aligned}$$

The last inequality above holds because $\delta = \delta(\mathbf{p}, \varepsilon)$ is sufficiently small. We now turn to the main task of estimating $c_X(0, 0, c_0, \ldots, c_{k-1})$. We use the following identities and inequalities.

- $c_L - c_F \in [\delta, 2\delta]$.
- $c_F = \frac{1}{2}$.
- $H(\mathbf{p}^{\theta_{\mathbf{P}}}) = \theta_{\mathbf{p}} I(\mathbf{p}, \mathbf{p}^{\theta_{\mathbf{P}}}) - \psi_{\mathbf{p}}(\theta_{\mathbf{p}})$.
- $\psi_{\mathbf{p}}(\theta_{\mathbf{p}}) = 2\psi_{\mathbf{p}}(2)$.

To deal with the entropy term in c_X, we use the non-negativity of Kullback-Leibler divergence. For any discrete probability distribution $\mathbf{q} = (q_0, \ldots, q_{k-1})$ (with $\sum_{i=0}^{k-1} q_i = 1$),

$$\begin{aligned} H(q_0, \ldots, q_{k-1}) &= \sum_{i=0}^{k-1} q_i \log\left(\frac{1}{(\mathbf{p}^{\theta_{\mathbf{P}}})_i}\right) - D_{\mathrm{KL}}(\mathbf{q}, \mathbf{p}^{\theta_{\mathbf{P}}}) \\ &\leq \sum_{i=0}^{k-1} q_i \log\left(\frac{1}{(\mathbf{p}^{\theta_{\mathbf{P}}})_i}\right) \\ &= -\psi_{\mathbf{p}}(\theta_{\mathbf{p}}) + \theta_{\mathbf{p}} \sum_{i=0}^{k-1} q_i \log\left(\frac{1}{p_i}\right). \end{aligned}$$

Using the above estimate with $q_i = \frac{c_i}{c_{\mathtt{tot}}}$, we find

$$\begin{aligned} c_X(0, 0, c_0, \ldots, c_{k-1}) &= c_{\mathtt{tot}} H(c_0, \ldots, c_{k-1}) + 5(c_L - c_F) + \frac{3}{2} + 2c_E + 2\delta \\ &\leq -c_{\mathtt{tot}}\psi_{\mathbf{p}}(\theta_{\mathbf{p}}) + \theta_{\mathbf{p}} \sum_{i=0}^{k-1} c_i \log\left(\frac{1}{p_i}\right) + \frac{3}{2} + 2c_E + 12\delta \qquad (2.4.19) \\ &\leq \theta_{\mathbf{p}} \sum_{i=0}^{k-1} c_i \log\left(\frac{1}{p_i}\right) + \frac{3}{2} - 2(\underline{C}_{\mathbf{p}} + \varepsilon)\psi_{\mathbf{p}}(2) + 12\delta \\ &\leq \theta_{\mathbf{p}} \sum_{i=0}^{k-1} c_i \log\left(\frac{1}{p_i}\right) + \frac{3}{2} - 2\psi_{\mathbf{p}}(2)\underline{C}_{\mathbf{p}} - \Omega_{\mathbf{p}}(\varepsilon). \qquad (2.4.20) \end{aligned}$$

The last line again follows because δ is sufficiently small. Finally we recall the following:

$$\sum_{i=0}^{k-1} c_i \log\left(\frac{1}{p_i}\right) = 1 - c_L = \frac{1}{2} + O(\delta),$$
$$C_{\mathbf{p}} = \frac{3+\theta_{\mathbf{p}}}{4\psi_{\mathbf{p}}(2)} \leq \underline{C}_{\mathbf{p}} = \max\left(C_{\mathbf{p}}, \frac{1}{\log(1/p_0)}, \frac{1}{\log(1/p_{k-1})}\right).$$

Substituting into the estimate (2.4.20), we obtain

$$c_X(0,0,c_0,\ldots,c_{k-1}) \leq \frac{3+\theta_{\mathbf{p}}+O(\delta)}{2} - 2\psi_{\mathbf{p}}(2)\underline{C}_{\mathbf{p}} - \Omega_{\mathbf{p}}(\varepsilon) \leq -\Omega_{\mathbf{p}}(\varepsilon).$$

This completes the proof. □

Proposition 2.4.1 readily follows by combining the ingredients just established.

Proof of Proposition 2.4.1. We start from the upper bound in Lemma 2.4.11 and group the strings $x \in \mathcal{L}_{\texttt{stable}}$ by their digit profile. For each digit profile $(b_0, b_{k-1}, c_0, c_1, \ldots, c_{k-1})$, by Proposition 2.4.2 the number of corresponding blocks $x \in \mathcal{L}_{\texttt{stable}}$ is at most

$$\binom{c_{\texttt{tot}}\log(N)}{c_0\log(N),\ldots,c_{k-1}\log(N)} \leq N^{c_{\texttt{tot}}H(c_0,\ldots,c_{k-1})}.$$

Lemmas 2.4.14 and 2.4.15 imply that for each fixed digit profile $(b_0, b_{k-1}, c_0, c_1, \ldots, c_{k-1})$,

$$\sum_{\substack{x\in\mathcal{L}_{\texttt{stable}},\\ \text{Digit Profile}(x)=(b_0,\ldots,c_{k-1})}} \sum_{i=1}^{N-1} \mathbb{P}[(i,i+1)\in E(G_{B_x})]^2 \leq 64N^{c_{\texttt{tot}}H(c_0,\ldots,c_{k-1})+5c_L-2c_F+2c_E+2\delta} + e^{-\Omega(N^\delta)}$$
$$= 64N^{c_X+4\delta} + e^{-\Omega(N^\delta)}$$
$$\leq 64N^{-\Omega_{\mathbf{p}}(\varepsilon)} + e^{-\Omega(N^\delta)}.$$

Since there are at most $O(\log^{k+2}(N)) \leq N^{o(1)}$ total digit profiles $(b_0, b_{k-1}, \ldots, c_{k-1})$, Lemma 2.4.11 therefore yields the desired estimate

$$\mathbb{E}[|E(G,G')|] \leq 256N^{-\Omega_{\mathbf{p}}(\varepsilon)} + e^{-\Omega(N^\delta)}.$$

□

2.4.5 Informal Derivation of the Value $C_{\mathbf{p}}$

We saw the constant

$$\psi_{\mathbf{p}}(2) = -\log \sum_{i=0}^{k-1} p_i^2$$

arise naturally in Lemma 2.4.12, expressed via c_E. In this informal subsection, we will explain why the constants $\theta_{\mathbf{p}}$ and $C_{\mathbf{p}}$ appeared in the final stages of the proof above by determining "straightforwardly" how large $\frac{K}{\log N}$ must be for Lemma 2.4.15 to hold. We again view $c_X(c_0, \dots, c_{k-1})$ as a continuous function and restrict to the main case that $b_0 = b_{k-1} = 0$. Moreover we will set all $O(\delta)$ terms to zero for simplicity. For $x \in \mathcal{L}_{\mathtt{stable}}$ with $b_0(x) = b_{k-1}(x) = 0$, we have $c_L(x) = c_F(x) = 1/2$ which yields the constraint equation

$$\sum_{i=0}^{k-1} c_i \log(1/p_i) = \frac{1}{2}. \tag{2.4.21}$$

Setting $C = \frac{K}{\log N}$, we find from $c_L(x) = c_F(x) = 1/2$ that

$$c_X = \big(H(c_0, \dots, c_{k-1}) + 2\psi_{\mathbf{p}}(2)\big) \cdot c_{\mathtt{tot}} + \frac{3}{2} - 2C\psi_{\mathbf{p}}(2).$$

To maximize $c_X = c_X(c_0, \dots, c_{k-1})$ given the constraint (2.4.21), we set the gradient ∇c_X to be parallel to the constraint direction $\big(\log(1/p_0), \log(1/p_1), \dots, \log(1/p_{k-1})\big)$. (Without arguing too formally, one expects there are no issues of maxima occurring at the boundary because the entropy function is concave and its inward-normal derivative diverges when any coordinate approaches 0.) By writing out the definition of entropy one readily computes that the maximizer $(c_0^*, \dots, c_{k-1}^*)$ satisfies

$$\begin{aligned}
\theta \log(1/p_i) &= \frac{\partial}{\partial c_i}(c_X^*) \\
&= 2\psi_{\mathbf{p}}(2) + \log(c_{\mathtt{tot}}^*/c_i^*) - 1 + \sum_{j \in [k]_0} \frac{c_j^*}{c_{\mathtt{tot}}^*} \\
&= 2\psi_{\mathbf{p}}(2) + \log(c_{\mathtt{tot}}^*/c_i^*)
\end{aligned}$$

for some proportionality constant $\theta \in \mathbb{R}$. Recalling that $\psi_{\mathbf{p}}(t) = -\log \phi_{\mathbf{p}}(t)$ for $\phi_{\mathbf{p}}(t) = \sum_{i=0}^{k-1} p_i^t$, we obtain by rearranging

$$\begin{aligned}
\frac{c_i^*}{c_{\mathtt{tot}}^*} &= e^{2\psi_{\mathbf{p}}(2)} p_i^\theta \\
&= \frac{p_i^\theta}{\phi_{\mathbf{p}}(2)^2}.
\end{aligned}$$

Since $\sum_{i=0}^{k-1} \frac{c_i^*}{c_{\mathtt{tot}}^*} = 1$ it follows that $\phi_{\mathbf{p}}(\theta) = \phi_{\mathbf{p}}(2)^2$, i.e. $\theta = \theta_{\mathbf{p}}$. Moreover we find $\left(\frac{c_0^*}{c_{\mathtt{tot}}^*}, \dots, \frac{c_{k-1}^*}{c_{\mathtt{tot}}^*}\right) = \mathbf{p}^{\theta_{\mathbf{p}}}$. Solving for $c_{\mathtt{tot}}^*$ using (2.4.21) above yields

$$\frac{1}{c_{\mathtt{tot}}^*} = \frac{2}{\phi(\theta_{\mathbf{p}})} \sum_{i=0}^{k-1} p_i^{\theta_{\mathbf{p}}} \log(1/p_i) = 2I(\mathbf{p}, \mathbf{p}^{\theta_{\mathbf{p}}}).$$

Finally plugging back into the definition of c_X and recalling properties of $I(\mathbf{p}, \mathbf{p}^t)$,

$$\begin{aligned} c_X(c_0^*, \dots, c_{k-1}^*) &= \frac{H(\mathbf{p}^{\theta_{\mathbf{p}}}) + 2\psi_{\mathbf{p}}(2)}{2I(\mathbf{p}, \mathbf{p}^{\theta_{\mathbf{p}}})} + \frac{3}{2} - 2C\psi_{\mathbf{p}}(2) \\ &= \frac{H(\mathbf{p}^{\theta_{\mathbf{p}}}) + \psi_{\mathbf{p}}(\theta_{\mathbf{p}})}{2I(\mathbf{p}, \mathbf{p}^{\theta_{\mathbf{p}}})} + \frac{3}{2} - 2C\psi_{\theta_{\mathbf{p}}}(2) \\ &= \frac{3 + \theta_{\mathbf{p}}}{2} - 2C\psi_{\mathbf{p}}(2). \end{aligned}$$

Rearranging shows that $c_X^* < 0$ is equivalent to

$$C > C_{\mathbf{p}} = \frac{3 + \theta_{\mathbf{p}}}{4\psi_{\mathbf{p}}(2)} = \frac{3 + \theta_{\mathbf{p}}}{2\psi_{\mathbf{p}}(\theta_{\mathbf{p}})}.$$

Therefore we have "straightforwardly" recovered the statement of Lemma 2.4.15. Let us also point out that

$$\begin{aligned} c_{\mathtt{tot}} &= \frac{1}{2I(\mathbf{p}, \mathbf{p}^{\theta_{\mathbf{p}}})} = \frac{\theta_{\mathbf{p}}}{2(H(\mathbf{p}^{\theta_{\mathbf{p}}}) + \psi_{\mathbf{p}}(\theta_{\mathbf{p}}))} \\ &< \frac{\theta_{\mathbf{p}}}{2\psi_{\mathbf{p}}(\theta_{\mathbf{p}})} < \frac{3 + \theta_{\mathbf{p}}}{2\psi_{\mathbf{p}}(\theta_{\mathbf{p}})} \\ &= C_{\mathbf{p}} \leq \underline{C}_{\mathbf{p}}. \end{aligned}$$

Here we used (2.4.5) in the first line. Hence the maximizer we found corresponds to "real" blocks B_x with length $M \approx c_{\mathtt{tot}} \log N < K$.

Since this argument ignored $O(\delta)$ error terms and some details on boundary issues, we verified Lemma 2.4.15 directly in the previous section instead of making the informal argument rigorous. The main step of this verification was to use non-negativity of the Kullback-Leibler divergence $D_{\mathtt{KL}}(\mathbf{q}, \mathbf{p}^{\theta_{\mathbf{p}}})$ with $q_i = \frac{c_i}{c_{\mathtt{tot}}}$ in inequality (2.4.19). Given the argument above, this step becomes quite natural. Indeed c_X is linear in $(c_0, \dots, c_{k-1})$ except for the entropy term, so (2.4.19) simply linearizes this entropy term around the equality case $\left(\frac{c_0^*}{c_{\mathtt{tot}}^*}, \dots, \frac{c_{k-1}^*}{c_{\mathtt{tot}}^*}\right) \approx \mathbf{p}^{\theta_p}$.

2.5 Proof of Lemma 2.3.8

In this section we prove Lemma 2.3.8, whose statement is recalled now.

Lemma 2.3.8. *For any* $\mathbf{p}$ *and positive reals* ε *and* t*, there is* $\delta = \delta(\mathbf{p}, \varepsilon, t)$ *such that if* $K \geq (\overline{C}_{\mathbf{p}} + \varepsilon)\log(N)$ *then*

$$\mathbb{E}[e^{t\cdot|E_{\text{for}}(G,G')|}] \leq 1 + O(N^{-\delta}).$$

It was shown in Section 2.3 how to upper-bound the total variation distance from uniform after K $\mathbf{p}$-shuffles based on the exponential moment estimate above. Therefore establishing Lemma 2.3.8 will complete the proof of the mixing time upper bound (2.1.3).

2.5.1 Preparatory Lemmas

Define $F(a, b)$ to be the value $\mathbb{E}[|E(G, G')|]$ for i.i.d. $\mathbf{p}$-random shuffle graphs G and G' on decks of a cards with b shuffles. Proposition 2.4.1 provides the main upper bound on $F(a, b)$, stated as a bound on $F(N, K)$. The next lemma gives a much easier estimate we will use for small values of a and b.

Lemma 2.5.1. *For any non-negative integers* a *and* b*,*

$$F(a, b) \leq \min\left(a, a^2 \cdot \phi_{\mathbf{p}}(2)^b\right).$$

Proof. The bound $F(a, b) \leq a$ is obvious. The other bound

$$\mathbb{E}[|E(G, G')|] \leq \mathbb{E}[|E(G)|] \leq a^2\phi_{\mathbf{p}}(2)^b$$

follows by summing over all $\binom{a}{2}$ pairs of strings s_i, s_j as in Lemma 2.4.12. □

The next two lemmas allow us to upper-bound relatively complicated expected edge intersections based on simple expected edge intersections. They will be used below to estimate the left-hand side of (2.5.2) as a sum over the blocks in the decomposition (2.5.3).

Lemma 2.5.2. *Let* A *and* B *be independent random subsets of a finite set* $\mathcal{A}$*. Let* A' *and* B' *respectively be independent copies of* A *and* B*. Then*

$$\mathbb{E}[|A \cap B|] \leq \frac{\mathbb{E}[|A \cap A'|] + \mathbb{E}[|B \cap B'|]}{2}.$$

Proof. For each $a \in \mathcal{A}$ let $A_a = \mathbb{P}[a \in A]$ and $B_a = \mathbb{P}[a \in B]$. Then the statement reduces to showing $\sum_a A_a B_a \leq \frac{\sum_a (A_a^2 + B_a^2)}{2}$ which holds by AM-GM. □

Lemma 2.5.3. *Let* A *be a random subset of a finite set* $\mathcal{A}$ *and let* $\mathcal{F}$ *be a* σ*-algebra. Let* A' *be an independent copy of* A *and let* $A_{\mathcal{F}}$ *and* $A'_{\mathcal{F}}$ *be conditionally independent copies of* A *conditioned on* $\mathcal{F}$*. Then*

$$\mathbb{E}[|A \cap A'|] \leq \mathbb{E}[|A_{\mathcal{F}} \cap A'_{\mathcal{F}}|]. \tag{2.5.1}$$

Proof. For each element $a \in \mathcal{A}$, let $Q_a = \mathbb{P}[a \in A|\mathcal{F}]$. Let $P_a = \mathbb{P}[a \in A] = \mathbb{E}[Q_a]$. Then (2.5.1) amounts to showing

$$\sum_{a\in\mathcal{A}} P_a^2 \leq \sum_{a\in\mathcal{A}} \mathbb{E}[Q_a^2].$$

Since $\mathbb{E}[Q_a^2] - P_a^2 \geq 0$ is simply the variance of the random variable Q_a for each a, the result follows. □

2.5.2 The Edge-Exploration Process

We now define the exploration process mentioned at the end of Section 2.3, which explores a pair $(s_1, \ldots, s_N), (s'_1, \ldots, s'_N) \in \mathcal{S}$ of sorted string sequences in order starting from s_1, s'_1. At step i, the currently revealed strings are

$$(s_1, \ldots, s_i) \quad \text{and} \quad (s'_1, \ldots, s'_i)$$

which results in revealed subgraphs

$$G_i \subseteq G, \quad G'_i \subseteq G'$$

that grow with i. Explicitly, G_i and G'_i are simply the induced subgraphs of G and G' on the vertex set $\{1, 2, \ldots, i\}$. When either s_i or s'_i begins with the prefix $[(k-1)(k-1)]$ we stop the process. Essentially by definition, this process finds all edges in $E_{\texttt{for}}(G, G')$. As alluded to at the end of Section 2.3, the following lemma shows how to bound the exponential moments of $E_{\texttt{for}}(G, G')$ using this exploration process.

Lemma 2.5.4. *Suppose $\gamma > 0$ is such that the conditional expectation estimate*

$$\mathbb{E}[E_{\texttt{for}}(G, G') - E(G_i, G'_i)|\mathcal{F}_i] \leq \gamma \tag{2.5.2}$$

holds almost surely with $\mathcal{F}_i \equiv \sigma(s_1, \ldots, s_i, s'_1, \ldots, s'_i)$ for each $i \in [N]$. Then

$$\mathbb{E}[e^{t\cdot E_{\texttt{for}}(G,G')}] \leq 1 + 2e^t\gamma$$

for any $t > 0$ satisfying $e^t\gamma \leq \frac{1}{10}$.

Proof. Define for simplicity the random variable $X = E_{\texttt{for}}(G, G')$. For each $j \geq 0$ let $t_j = \inf\{i : E(G_i, G'_i) \geq j\}$. Then t_j is an stopping time, and if $t_j < \infty$ then $|E(G_{t_j}, G'_{t_j})| = j$ holds because $E(G_{i+1}, G'_{i+1}) - E(G_i, G'_i) \leq 1$ holds almost surely for each i. Morever when $t_j < \infty$ we have

$$\mathbb{P}[X > j|\mathcal{F}_{t_j}] \leq \gamma$$

due to the assumption (2.5.2). Of course the inequality $X \geq j$ implies that t_j is finite. Hence by

optional stopping, we may average the above display to conclude that

$$\mathbb{P}[X > j | X \geq j] \leq \gamma.$$

This means X has hazard rate at least that of a geometric random variable Y with

$$\mathbb{P}[Y = j] = (1-\gamma)\gamma^j, \quad j \geq 0.$$

Therefore X is stochastically dominated by Y. Using the assumption $e^t\gamma \leq \frac{1}{10}$, we find

$$\begin{aligned}
\mathbb{E}[e^{tX}] &\leq \mathbb{E}[e^{tY}] \\
&\leq (1-\gamma)\sum_{j\geq 0}(e^t\gamma)^j \\
&\leq \frac{1}{1-e^t\gamma} \\
&\leq 1 + 2e^t\gamma.
\end{aligned}$$

□

To analyze the exploration process we group the potential future strings which are lexicographically larger than s_i. Supposing that $s_i <_{\texttt{lex}} [(k-1)(k-1)]$ does not begin with $[(k-1)(k-1)]$, set

$$\texttt{Blocks}(s_i) = \texttt{Blocks}(s_i, [(k-1)(k-1)])$$

in the notation of Lemma 2.5.5 just below. By construction, $\texttt{Blocks}(s_i)$ consists of $O(\log N)$ blocks and

$$\{s \in [k]_0^K : s_i <_{\texttt{lex}} s <_{\texttt{lex}} [(k-1)(k-1)]\} = \bigcup_{x \in \texttt{Blocks}(s_i)} B_x. \tag{2.5.3}$$

The fact that $|\texttt{Blocks}(s_i)| \leq O(\log N) \leq N^{o(1)}$ will be used in the proof of Lemma 2.3.8 in the next subsection. It allows us to estimate a sum over $x \in \texttt{Blocks}(s_i)$ by its maximum term; see just before the start of Case 1 therein.

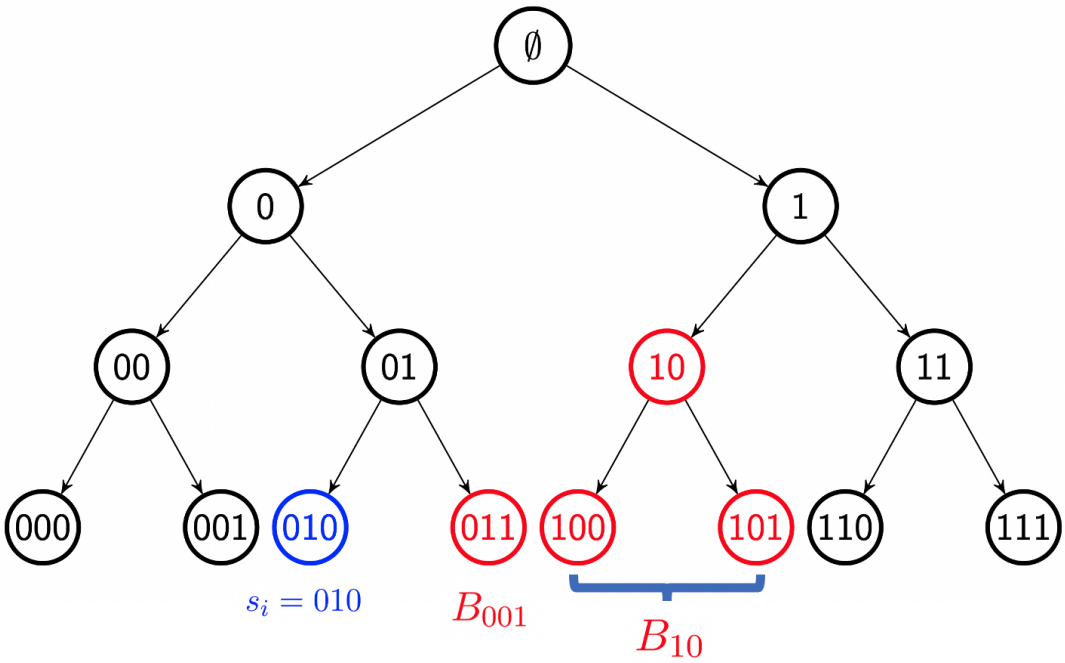


Figure 2.6: The decomposition of (2.5.3), guaranteed by Lemma 2.5.5, is shown when $s_i = 010$ with $(k, K) = (2, 3)$. It states that $\{s \in [2]_0^3 : 010 <_{\texttt{lex}} s <_{\texttt{lex}} 11\} = B_{001} \cup B_{10}$.

Lemma 2.5.5. *Let $s_a <_{\texttt{lex}} s_b$ be strings each of length at most K. Define the lexicographic interval*

$$I_{s_a,s_b} \equiv \{s \in [k]_0^K : s_a <_{\texttt{lex}} s <_{\texttt{lex}} s_b\}.$$

Then I_{s_a,s_b} can be written as a disjoint union of blocks

$$I_{s_a,s_b} = \bigcup_{x \in \texttt{Blocks}(s_a,s_b)} B_x$$

for some set $\texttt{Blocks}(s_a, s_b)$ containing at most $2Kk \leq O(\log N)$ strings, each of length at most K.

Proof. For $0 \leq M \leq K$, define

$$\overline{\texttt{Blocks}}^M(s_a, s_b) = \{x \in [k]_0^M : B_x \cap I_{s_a,s_b} \neq \emptyset\}$$

to be the set of all length-M strings x such that B_x has non-trivial intersection with I_{s_a,s_b}. Similarly define

$$\underline{\texttt{Blocks}}^M(s_a, s_b) = \{x \in [k]_0^M : B_x \subseteq I_{s_a,s_b}\}$$

to be the set of all length-M strings x such that B_x is contained inside I_{s_a,s_b}. Clearly $\underline{\texttt{Blocks}}^M(s_a,s_b) \subseteq \overline{\texttt{Blocks}}^M(s_a,s_b)$. Moreover the fact that I_{s_a,s_b} is a lexicographic interval means these sets differ in at most 2 elements, i.e.

$$\left|\overline{\texttt{Blocks}}^M(s_a,s_b) \backslash \underline{\texttt{Blocks}}^M(s_a,s_b)\right| \leq 2. \tag{2.5.4}$$

Define

$$\overline{\texttt{Blocks}}(s_a,s_b) = \bigcup_{0\leq M\leq K} \overline{\texttt{Blocks}}^M(s_a,s_b),$$
$$\underline{\texttt{Blocks}}(s_a,s_b) = \bigcup_{0\leq M\leq K} \underline{\texttt{Blocks}}^M(s_a,s_b).$$

Next, for any $s \in I_{s_a,s_b}$, note that all ancestors (prefixes) of s are contained in $\overline{\texttt{Blocks}}(s_a,s_b)$, while $\emptyset \notin \underline{\texttt{Blocks}}(s_a,s_b)$. Let y_s be the longest ancestor string of s with

$$y_s \notin \underline{\texttt{Blocks}}(s_a,s_b).$$

By definition $y_s \neq s$, so y_s has a child x_s which is also an ancestor of s (possibly $x_s = s$). By definition of y_s,

$$x_s \in \underline{\texttt{Blocks}}(s_a,s_b)$$

and so

$$B_{x_s} \subseteq I_{s_a,s_b}.$$

We claim the blocks B_{x_s} constructed in this way from $s \in I_{s_a,s_b}$ are pairwise equal or disjoint. Indeed if

$$B_{x_s} \subsetneq B_{x_{s'}}$$

then $x_{s'}$ is a prefix of y_s. However

$$y_s \notin \underline{\texttt{Blocks}}(s_a,s_b)$$

and

$$x_{s'} \in \underline{\texttt{Blocks}}(s_a,s_b)$$

which contradicts the fact that $\underline{\texttt{Blocks}}(s_a,s_b)$ is descendent-closed.

Above, we started from an arbitrary $s \in I_{s_a,s_b}$ and found a block B_{x_s} containing s. It follows that the distinct blocks B_{x_s} appearing in the above construction form a partition of I_{s_a,s_b}. Finally, note that by inequality (2.5.4), and the fact that y_s has length at most $K-1$, y_s ranges over a set of size at most $2K$. Hence x_s ranges over a set of size at most $2Kk$. This concludes the proof. $\square$

The next lemma shows that conditioning on the exploration process for G up to s_i with

$$s_i <_{\texttt{lex}} [(k-1)(k-1)]$$

does not dramatically increase the typical size of $\mathcal{I}(B_x)$ for any $x \in \texttt{Blocks}(s_i)$. The fact that $s_i <_{\texttt{lex}} [(k-1)(k-1)]$ is crucial here, as was discussed at the end of Subsection 2.3.1. Indeed conditioning on $s_i = [(k-1)^K]$ would imply that

$$s_i = s_{i+1} = \cdots = s_N = [(k-1)^K]$$

so that $E(G)$ contains all remaining potential edges $(i, i+1), (i+1, i+2), \ldots, (N-1, N)$. However when $s_i <_{\texttt{lex}} [(k-1)(k-1)]$, a constant fraction of the $\mu_{\mathbf{p},K}$-measure of $[k]_0^K$ remains not yet occupied, which prevents such an example from occuring.

Lemma 2.5.6. *Conditioned on $(s_1, \ldots, s_i)$ which satisfy $s_i <_{\texttt{lex}} [(k-1)(k-1)]$, for any $x \in \texttt{Blocks}(s_i)$ the conditional distribution of $|\mathcal{I}(B_x)|$ is stochastically dominated by a $\texttt{Bin}(N, p_{\texttt{min}}^{-2}\lambda_x)$ random variable.*

Proof. Condition further on the largest value $j \in [N]$ with $s_i = s_j$. Then we can generate all strings $(s_{j+1}, \ldots, s_N)$ by sampling i.i.d. random numbers $a'_{j+1}, \ldots, a'_N$ uniformly from $[t_{s_i} + \lambda_{s_i}, 1]$, sorting them into increasing order $a_{j+1} \le a_{j+2} \le \cdots \le a_N$, and choosing $s_\ell \in [k]_0^K$ such that $a_\ell \in J_{s_\ell}$ for $\ell \ge j+1$. There are $N - j \le N$ such random numbers a_ℓ, and $1 - (t_{s_i} + \lambda_{s_i}) \ge p_{\texttt{min}}^2$ because of the assumption that $s_i <_{\texttt{lex}} [(k-1)(k-1)]$. Therefore conditionally on j, each a'_i has probability at most $p_{\texttt{min}}^{-2}\lambda_x$ to fall into the interval J_x, which completes the proof. □

2.5.3 Proof of Lemma 2.3.8

We now complete the proof of Lemma 2.3.8. In light of Lemma 2.5.4 it remains to show that the conditional expectation for the number of unrevealed edges in $E_{\texttt{for}}(G, G')$, given by

$$\mathbb{E}[E_{\texttt{for}}(G, G') - E(G_i, G'_i)|\mathcal{F}_i],$$

is almost surely bounded by $O(N^{-\delta})$. The idea is to use Lemmas 2.5.2 and 2.5.3 to upper-bound this quantity by a sum over the future blocks appearing in (2.5.3), see Equation (2.5.6) in the proof below. Analyzing the summand corresponding to a block B_x for $x \in [k]_0^M$ amounts to a smaller version of the problem considered in Proposition 2.4.1 since B_x can be viewed as a copy of $[k]_0^{K-M}$. As a result, the summand for B_x has value $F(|\mathcal{I}_{B_x}|, K-M)$. This term can be estimated by Lemma 2.5.1 when $\mathbb{E}[|\mathcal{I}_{B_x}|] \le N^\delta$ is small (Cases 1 and 2 of the proof below) and by Proposition 2.4.1 when $\mathbb{E}[|\mathcal{I}_{B_x}|] \ge N^\delta$ is reasonably large (Case 3 of the proof).

Proof of Lemma 2.3.8. Take $\delta = \delta(\mathbf{p}, \varepsilon)$ sufficiently small, $\eta = \eta(\mathbf{p}, \varepsilon, \delta)$ smaller and $\zeta = \zeta(\mathbf{p}, \varepsilon, \delta, \eta)$ yet smaller. Define the following σ-algebras.

$$\mathcal{F}_i = \sigma(s_1, \dots, s_i, s'_1, \dots, s'_i),$$
$$\widetilde{\mathcal{F}}_i = \sigma\left(s_1, \dots, s_i, s'_1, \dots, s'_i, (\mathcal{I}(B_x))_{x \in \texttt{Blocks}(s_i)}\right).$$

(Note that the σ-algebras $\widetilde{\mathcal{F}}_i$ do not define a filtration as i varies.) Let

$$G_{u,1} = E_{\texttt{for}}(G) \backslash E(G_i)$$

consist of all so-far-unrevealed edges which do not involve strings beginning with $[(k-1)(k-1)]$. Let $G_{u,2}$ be a conditionally independent copy of $G_{u,1}$ given $\widetilde{\mathcal{F}}_i$ - equivalently this means $G_{u,2}$ is obtained by resampling $G_{u,1}$ conditioned to have the same sets $\mathcal{I}(B_x)$ for each $x \in \texttt{Blocks}(s_i)$. Define $G'_{u,1}, G'_{u,2}$ the same way for G'. Hence $G_{u,1}, G_{u,2}, G'_{u,1}, G'_{u,2}$ are shuffle graphs with all edge-endpoints in $\{i, i+1, \dots, N\}$.

We will show that at any time i in the exploration process, the expected number of unrevealed edges in $E_{\texttt{for}}(G, G')$ is bounded by

$$\mathbb{E}[|E(G_{u,1}, G'_{u,1})| \big| \mathcal{F}_i] \leq O(N^{-\zeta}).$$

By Lemma 2.5.4, this will complete the proof of Lemma 2.3.8 up to replacing ζ with δ. First, using Lemmas 2.5.2 and 2.5.3 conditionally on $\mathcal{F}_i$, we estimate the expected number of unrevealed edges by

$$\mathbb{E}[|E(G_{u,1}, G'_{u,1})| \big| \mathcal{F}_i] \leq \mathbb{E}\left[\frac{|E(G_{u,1}, G_{u,2})| + |E(G'_{u,1}, G'_{u,2})|}{2} \big| \mathcal{F}_i\right].$$

Therefore by symmetry it suffices to show that

$$\mathbb{E}\left[|E(G_{u,1}, G_{u,2})| \big| \mathcal{F}_i\right] \leq O(N^{-\zeta})$$

holds almost surely. By definition, conditioning on $\widetilde{\mathcal{F}}_i$ determines the interval $\mathcal{I}(B_x)$ for each such x. Moreover the remaining $K - M$ digits of each of the $|\mathcal{I}(B_x)|$ random strings in B_x are still i.i.d. $\mathbf{p}$-random. As a consequence,

$$\mathbb{E}[E(G_{u,1}, G_{u,2}) \big| \widetilde{\mathcal{F}}_i] = |\{j > i : s_j = s_i\}| + \sum_{x \in \texttt{Blocks}(s_i)} F(|\mathcal{I}(B_x)|, K - M). \tag{2.5.5}$$

Indeed recall from the start of Subsection 2.5.1 that $F(a, b)$ is the expected size of $E(G, G')$ when there are a cards and b shuffles. Thus (2.5.5) essentially holds by definition.

Next, the law of total expectation yields

$$\begin{aligned}\mathbb{E}[E(G_{u,1}, G_{u,2})\big|\mathcal{F}_i] &= \mathbb{E}\left[\mathbb{E}[E(G_{u,1}, G_{u,2})\big|\widetilde{\mathcal{F}}_i]\Big|\mathcal{F}_i\right] \\ &= \mathbb{E}\left[|\{j > i : s_j = s_i\}|\,\big|\mathcal{F}_i\right] + \sum_{x \in \texttt{Blocks}(s_i)} \mathbb{E}\left[F(|\mathcal{I}(B_x)|, K - M)\big|\mathcal{F}_i\right]. \quad (2.5.6)\end{aligned}$$

The first term on the right-hand side of (2.5.6) is controlled by Lemma 2.3.9, which implies

$$\mathbb{E}[|\{j > i : s_j = s_i\}|\big|\mathcal{F}_i] \le O(N^{-\zeta}).$$

To estimate the other (main) term on the right-hand side of (2.5.6), we will show for each $x \in \texttt{Blocks}(s_i)$ that

$$\mathbb{E}\left[F(|\mathcal{I}(B_x)|, K - M)\big|\mathcal{F}_i\right] \le O(N^{-\zeta}).$$

As $|\texttt{Blocks}(s_i)| = O(\log N) \le N^{o(1)}$ this suffices to finish the proof. We now split into three cases depending on the size of λ_x. In all cases below we let M denote the length of x. Case 3 (the main one) is where Proposition 2.4.1 is essential.

Let us emphasize that $|I(B_x)|$ is still random conditionally on $\mathcal{F}_i$. While we do not have good almost sure bounds on $|I(B_x)|$ itself, its conditional distribution is uniformly stochastically bounded by Lemma 2.5.6. Since we are estimating a conditional expectation given $\mathcal{F}_i$ and not $\widetilde{\mathcal{F}}_i$, this suffices for an almost sure bound.

Case 1: $\lambda_x \le N^{-1-\delta}$.addpunct. In this case, Lemmas 2.5.1 and 2.5.6 imply

$$\begin{aligned}\mathbb{E}\left[F(|\mathcal{I}(B_x)|, K - M)|\mathcal{F}_i\right] &\le \mathbb{E}[|\mathcal{I}(B_x)|] \\ &\le O(N^{-\zeta}).\end{aligned}$$

Case 2: $N^{-1-\delta} \le \lambda_x \le N^{-1+\delta}$.addpunct. In this case, Lemmas 2.5.6 and 2.4.7 imply that $|\mathcal{I}(B_x)| \le N^{2\delta}$ holds with probability $1 - e^{-\Omega(N^{\delta})}$. The fact $\lambda_x \le (p_{\texttt{max}})^M$ implies

$$\begin{aligned}M &\le \frac{\log(\lambda_x^{-1})}{\log(p_{\texttt{max}}^{-1})} \\ &\le \frac{(1+\delta)\log N}{\log(p_{\texttt{max}}^{-1})}\end{aligned}$$

In particular as $\delta \ll \varepsilon$ is sufficiently small this implies $K - M \ge \Omega_{\mathbf{p}}(\varepsilon)\log N$. Lemma 2.5.1 now

yields

$$\begin{aligned}\mathbb{E}\left[F(|\mathcal{I}(B_x)|, K-M)|\mathcal{F}_i\right] &\leq \mathbb{E}[|\mathcal{I}(B_x)|^2]\phi_{\mathbf{p}}(2)^{\Omega_{\mathbf{p}}(\varepsilon)\log(N)} \\ &\leq O\left(N^{2\delta-\Omega_{\mathbf{p}}(\varepsilon)}\right) \\ &\leq O(N^{-\zeta}).\end{aligned}$$

Case 3: $\lambda_x \geq N^{-1+\delta}$*.addpunct.* Similarly to the previous case, observe that

$$\begin{aligned}M &\leq \frac{\log(\lambda_x^{-1})}{\log(p_{\mathtt{max}}^{-1})} &(2.5.7)\\ &\leq \overline{C}_{\mathbf{p}}\log(\lambda_x^{-1}). &(2.5.8)\end{aligned}$$

We break into subcases depending on $|\mathcal{I}(B_x)|$. The first subcase is that $|\mathcal{I}(B_x)| \leq N^{\eta}$. Here the lower bound $K - M \geq \Omega_{\mathbf{p}}(\delta \log N)$ follows from inequality (2.5.8), and applying Lemma 2.5.1 yields

$$F(|\mathcal{I}(B_x)|, K-M) \leq N^{2\eta}\phi_{\mathbf{p}}(2)^{K-M} \leq N^{-\Omega_{\mathbf{p}}(\delta)}.$$

In the main subcase $|\mathcal{I}(B_x)| \in [N^{\eta}, 2p_{\mathtt{min}}^{-2}N\lambda_x]$ we obtain:

$$\begin{aligned}K - M &\geq (\overline{C}_{\mathbf{p}} + \varepsilon)\log(N\lambda_x) &(2.5.9)\\ &\geq \left(\overline{C}_{\mathbf{p}} + \frac{\varepsilon}{2}\right)\log(2p_{\mathtt{min}}^{-2}N\lambda_x) \\ &\geq \left(\overline{C}_{\mathbf{p}} + \frac{\varepsilon}{2}\right)\log|\mathcal{I}(B_x)|.\end{aligned}$$

Since $|\mathcal{I}(B_x)| \geq N^{\eta}$ tends to infinity with N, Proposition 2.4.1 implies

$$F(|\mathcal{I}(B_x)|, K-M) \leq O\left(|\mathcal{I}(B_x)|^{-\delta}\right) \leq O(N^{-\zeta}).$$

Finally the subcase $|\mathcal{I}(B_x)| \geq 2p_{\mathtt{min}}^{-2}N\lambda_x$ occurs with tiny probability $e^{-\Omega(N^{\delta})}$ by Lemmas 2.5.6 and 2.4.7. In this subcase we use the trivial bound $F(|\mathcal{I}(B_x)|, K-M) \leq N$. Combining subcases, we have established that whenever Case 3 holds,

$$\mathbb{E}\left[F(|\mathcal{I}(B_x)|, K-M)|\mathcal{F}_i\right] \leq O(N^{-\zeta}).$$

Combining cases (and substituting δ for ζ at the end) concludes the proof of Lemma 2.3.8. □

Remark 2.5.1. Recall that throughout Section 2.4, and in particular in Proposition 2.4.1, the weaker inequality $K \geq (\underline{C}_{\mathbf{p}} + \varepsilon) \log N$ sufficed where

$$\underline{C}_{\mathbf{p}} \equiv \max\left(C_{\mathbf{p}}, \frac{1}{\log(1/p_0)}, \frac{1}{\log(1/p_{k-1})}\right) \leq \overline{C}_{\mathbf{p}}.$$

This means that when $k > 2$, for some parameter choices such as $\mathbf{p} = (0.01, 0.98, 0.01)$, the expectation $\mathbb{E}[|E(G, G')|]$ becomes small before mixing occurs, so the exponential moments of $|E(G, G')|$ are still large. This discrepancy can be explained as follows. When K satisfies

$$\underline{C}_{\mathbf{p}} + \varepsilon < \frac{K}{\log N} < \overline{C}_{\mathbf{p}} - \varepsilon,$$

the graph G typically contains $N^{\Omega(1)}$-size connected components coming from strings with nearly all digits $i_{\mathtt{max}}$. In such situations $\mathbb{E}[|E(G, G')|] \leq o(1)$ is small by Proposition 2.4.1. However an easy pigeonhole argument on N copies of G shows that with $\Omega(1/N^2)$ probability, $E(G, G')$ contains an $N^{\Omega(1)}$-sized component formed by a large G-component and large G'-component overlapping. As a result $|E(G, G')|$ has large exponential moments. (Moreover this argument still applies if we initially require $S, S' \in \mathcal{S}_1$ to be "typical".)

In upper-bounding the mixing time, the bound $K \geq (\overline{C}_{\mathbf{p}} + \varepsilon) \log N$, as opposed to $K \geq (\underline{C}_{\mathbf{p}} + \varepsilon) \log N$, is necessary in two places. The first is in Lemma 2.3.9. The other occurs above in (2.5.9) where we needed to ensure that Proposition 2.4.1 yields an upper bound for $F(|\mathcal{I}(B_x)|, K - M)$. In the worst case, all M of x's digits might be $i_{\mathtt{max}}$. Then typically (at least when the right-hand side below is positive),

$$\log |\mathcal{I}(B_x)| \approx \log(N) - M \log(1/p_{\mathtt{max}}).$$

To apply Proposition 2.4.1, we thus need

$$K - M \geq \underline{C}_{\mathbf{p}}(\log N - M \log(1/p_{\mathtt{max}}))$$

to hold for any M making both sides positive. In particular, if we continuously increase M the right side must reach 0 before the left side, which implies $K \geq \frac{\log N}{\log(1/p_{\mathtt{max}})}$. On the other hand, when $M = 0$ we need $K \geq \underline{C}_{\mathbf{p}} \log N$ for Proposition 2.4.1 to apply. Hence at least in bounding the exponential moments of $|E(G, G')|$, the value $\overline{C}_{\mathbf{p}} = \max\left(\underline{C}_{\mathbf{p}}, \frac{1}{\log(1/p_{\mathtt{max}})}\right)$ arises from the need to apply Proposition 2.4.1 for all sizes of block B_x appearing in the partition (2.5.3).

2.6 Proof of the Mixing Time Lower Bound

In this section we take $K = \lfloor (C_{\mathbf{p}} - \varepsilon)\log(N) \rfloor$ and show that almost no total-variation mixing occurs after K shuffles. First, when $K \leq (\widetilde{C}_{\mathbf{p}} - \varepsilon)\log(N)$ we previously argued at the start of Subsection 2.2.1 that the total variation distance from uniform is $1 - o(1)$. Hence we may assume that $\widetilde{C}_{\mathbf{p}} < C_{\mathbf{p}}$ holds, else there is nothing to prove. By taking ε small enough, we may further assume

$$K \geq (\widetilde{C}_{\mathbf{p}} + \varepsilon)\log N. \tag{2.6.1}$$

For a set $H \subseteq \mathbb{Z}$, its boundary $\partial H \subseteq H$ is defined by

$$\partial H \equiv \{h \in H : h - 1 \notin H \text{ or } h + 1 \notin H\}.$$

Its edge set $E(H)$ is the set of edges with both endpoints in H, i.e. we identify H with the corresponding induced subgraph of G. We will verify the following criterion from [Lal00] for non-mixing. The idea of the proof is to use the number of ascents of $\sigma \in \mathfrak{S}_N$ within H to distinguish the uniform distribution $\sigma = \pi$ from the shuffled distribution $\sigma = \pi^G$.

Proposition 2.6.1 ([Lal00, Proposition 2])**.** *Let $(K_N)_{N\geq 1}$ be a deterministic sequence of positive integers. Suppose there exist deterministic subsets $H = H_N \subseteq [N]$ such that for some $\delta = \delta(\mathbf{p}, \varepsilon)$ the following properties hold as $N \to \infty$, where G is the shuffle graph for a deck of N cards undergoing K_N $\mathbf{p}$-shuffles:*

$$|H| \to \infty \tag{2.6.2}$$

$$|\partial H| = O(|H|^{1/2}) \tag{2.6.3}$$

$$\mathbb{P}\left[|E(G) \cap E(H)| \geq |H|^{\frac{1}{2}+\delta}\right] \to 1. \tag{2.6.4}$$

Then asymptotically no total-variation mixing occurs after K_N shuffles, i.e.

$$\lim_{N\to\infty} d_N^{\mathrm{TV}}(K_N) = 1.$$

Remark 2.6.1. By using AM-GM or Cauchy–Schwarz similarly to the proof of Lemma 2.5.3, the conditions of Proposition 2.6.1 imply

$$\begin{aligned}\mathbb{E}[|E(G, G')|] &\geq \frac{\mathbb{E}[|E(G) \cap E(H)|]^2}{|E(H)|} \cdot (1 - o(1)) \\ &\geq \Omega(|H|^{2\delta}) \\ &\gg 1.\end{aligned}$$

However it does **not** follow from what we show that $K = (\underline{C}_{\mathbf{p}} \pm o(1)) \log N$ is always the cutoff point where the expected number $\mathbb{E}[|E(G, G')|]$ of shared edges in G and G' transitions from superconstant to subconstant. This is because the analysis of this section assumes inequality (2.6.1).

2.6.1 Preparation and Proof Idea

Define $\alpha_{\mathtt{tot}} \log(N) = \left\lfloor \frac{1-\delta}{2I(\mathbf{p}, \mathbf{p}^{\theta_{\mathbf{p}}})} \log(N) \right\rfloor$, where as usual $\delta = \delta(\mathbf{p}, \varepsilon)$ is sufficiently small. Choose (via some rounding procedure) positive integers $\alpha_0 \log(N), \ldots, \alpha_{k-1} \log(N)$ satisfying

$$\sum_{i=0}^{k-1} \alpha_i = \alpha_{\mathtt{tot}} \quad \text{and} \quad \left| \alpha_i \log(N) - \frac{\alpha_{\mathtt{tot}} \log(N) p_i^{\theta_{\mathbf{p}}}}{\phi_{\mathbf{p}}(\theta_{\mathbf{p}})} \right| \leq 1. \tag{2.6.5}$$

Note that $\alpha_{\mathtt{tot}} \leq \frac{3\underline{C}_{\mathbf{p}}}{4}$; indeed we showed in Proposition 2.1.1 that $\theta_{\mathbf{p}} \leq 4$, hence

$$\begin{aligned} \alpha_{\mathtt{tot}} + O(\delta) = \frac{1}{2I(\mathbf{p}, \mathbf{p}^{\theta_{\mathbf{p}}})} &= \frac{\theta_{\mathbf{p}}}{2(H(\mathbf{p}^{\theta_{\mathbf{p}}}) + \psi_{\mathbf{p}}(\theta_{\mathbf{p}}))} \\ &< \frac{\theta_{\mathbf{p}}}{2\psi_{\mathbf{p}}(\theta_{\mathbf{p}})} < \frac{3 + \theta_{\mathbf{p}}}{3\psi_{\mathbf{p}}(\theta_{\mathbf{p}})} \\ &= \frac{2C_{\mathbf{p}}}{3} \leq \frac{2\underline{C}_{\mathbf{p}}}{3}. \end{aligned}$$

We may therefore take $\beta_{\mathtt{tot}} \log(N) = K - \alpha_{\mathtt{tot}} \log(N) \geq \Omega(\log N)$ and choose positive integers $(\beta_i \log(N))_{i \in [k]_0}$ with

$$\sum_{i=0}^{k-1} \beta_i = \beta_{\mathtt{tot}} \quad \text{and} \quad \left| \beta_i \log(N) - \frac{\beta_{\mathtt{tot}} \log(N) p_i^2}{\phi_{\mathbf{p}}(2)} \right| \leq 1.$$

The numbers just constructed satisfy

$$\sum_{i=0}^{k-1} \alpha_i \log(N) + \sum_{i=0}^{k-1} \beta_i \log(N) = \alpha_{\mathtt{tot}} \log(N) + \beta_{\mathtt{tot}} \log(N) = K.$$

We will consider G-edges coming from strings with $\alpha_i \log(N)$ digits i in the first $\alpha_{\mathtt{tot}} \log(N)$ digits, and $\beta_i \log(N)$ digits i in the last $\beta_{\mathtt{tot}} \log(N)$ digits, for each $i \in [k]_0$. This is essentially a two-part digit profile. Let us point out that strings with many leading 0 or $(k-1)$ digits will **not** require special care in this part.

Definition 2.6.2. *The length $\alpha_{\mathtt{tot}} \log(N)$ string $x \in [k]_0^M$ is a **collision-likely prefix** (we write $x \in \mathtt{Pre}_{\mathtt{CL}}$) if x contains $\alpha_i \log(N)$ digits of i for each $i \in [k]_0$.*

Definition 2.6.3. *The string $s \in [k]_0^K$ is **collision-likely** (we write $s \in \mathtt{CL}$) if s satisfies the following properties.*

- *With* $M = \alpha_{\mathtt{tot}} \log(N)$*, the first* M *digits of* s *form a collision-likely prefix.*
- $s[M+1] = 0$*,* $s[M+2] = 1$*.*
- *The* $\beta_{\mathtt{tot}} \log(N)$ *digits* $s[M+1], s[M+2], \ldots, s[K]$ *consist of* $\beta_i \log(N)$ *digits of* i *for each* $i \in [k]_0$*.*

Recall from (2.2.3) the definition $J_x = [t_x, t_x + \lambda_x)$ and set

$$H \equiv \mathbb{Z} \cap \left(\bigcup_{x \in \mathtt{Pre}_{\mathtt{CL}}} N J_x \right).$$

That is, H consists of the "expected locations" of collision-likely prefixes. The set H is essentially the same as in the lower bound of [Lal00]. Our analysis differs from Lalley's in the last part of Definition 2.6.3 where we consider strings whose later digits have empirical distribution $\mathbf{p}^2$.

Before proceeding into more technical details, let us give some intuition both for the definitions above and the remainder of the proof. Based on Subsection 2.4.5, we expect that the bulk of the edges in $E(G, G')$ come from the blocks B_x with digit profile

$$c_i(x) \approx c_i^* = \frac{1}{2I(\mathbf{p}, \mathbf{p}^{\theta_{\mathbf{P}}})} \cdot \frac{\mathbf{p}_i^{\theta_{\mathbf{P}}}}{\phi_{\mathbf{P}}(\theta_{\mathbf{P}})}.$$

Therefore we took $\alpha_i \approx c_i^*$ and defined H so that

$$H \approx \bigcup_{x \in \mathtt{Pre}_{\mathtt{CL}}} \mathcal{I}(B_x).$$

The main difficulty in applying Proposition 2.6.1 is to verify the last condition by lower-bounding the number of G-edges appearing in blocks B_x for $x \in \mathtt{Pre}_{\mathtt{CL}}$. Intuitively, to count these edges one should simply count pairs of strings in B_x as in Lemma 2.4.12. However this will overestimate the number of G-edges for strings that appear many times. Hence one would like to also control for example the number of equal triples $s_i = s_{i+1} = s_{i+2} = s$ — this is relevant for obtaining the correct first moment and also for controlling the variance. Such a strategy was carried out in [Lal00, Lemmas 8 and 9]. However for this approach to work, $\mathbf{p}$ must be close to a uniform distribution so that the expected number of triples does not overwhelm the expected number of pairs.

Instead of counting pairs of equal strings $s_i = s_j$, we consider for each $s \in \mathtt{CL}$ the event Y_s that $s_i = s_{i+1} = s$ holds for **at least** one $i \in [N]$. Because of the "extra margin" afforded by the second property of $\mathtt{CL}$ in Definition 2.6.3, it follows that with high probability, all $i \in [N]$ with $s_i \in \mathtt{CL}$

satisfy $i \in H$ (see Lemma 2.6.6). Under this event, we have

$$|E(G) \cap E(H)| \geq \sum_{s \in \mathtt{CL}} 1_{Y_s}. \tag{2.6.6}$$

The sum $\sum_{s\in\mathtt{CL}} 1_{Y_s}$ turns out to concentrate nicely while retaining almost the same expectated value. Indeed the indicator functions 1_{Y_s} are pairwise anti-correlated as $s \in \mathtt{CL}$ varies. Therefore whenever the expected value $\mathbb{E}\left[\sum_{s\in\mathtt{CL}} 1_{Y_s}\right] \geq N^{\Omega(1)}$ is large, Chebychev's inequality immediately implies a high-probability lower bound of the same order.

Since $\mathbb{P}[Y_s]$ is a function of the digit profile of s, it suffices to focus on a single digit profile, keeping in mind that the prefix should be collision-likely. Restricting the sums above to $s \in \mathtt{CL}$ exactly corresponds to such a choice of digit profile. The reason to choose $\mathbf{p}^2$ for the distribution of the later digits in the definition of $\mathtt{CL}$ is that conditioned on two $\mathbf{p}$-random digits being equal, the distribution of this shared digit is $\mathbf{p}^2$. Thus we expect most collisions inside a block B_x to have digit distribution $\mathbf{p}^2$ in the later $K - M$ digits.

In summary, the lower bound (2.6.6) essentially involves two separate truncation steps. The first step, truncating $|E(G) \cap E(H)|$ to a sum of indicators 1_{Y_s}, is important to obtain control of the second moment. The second step, restricting this sum to collision-likely strings $s \in \mathtt{CL}$, is simply a convenient way to isolate the dominant contribution to the sum over all strings s with collision-likely prefixes $s[1]\dots s[\alpha_{\mathtt{tot}}\log(N)] = x \in \mathtt{Pre}_{\mathtt{CL}}$.

We conclude this subsection with two lemmas, the second of which verifies the "easy" parts of Proposition 2.6.1.

Lemma 2.6.4. *For sufficiently large N,*

$$\sum_i \alpha_i \log(p_i) = \frac{-1+\delta}{2} \pm o(1).$$

Proof. By the definition of α_i in (2.6.5) and of $I(\mathbf{p}, \mathbf{p}^{\theta_{\mathtt{P}}})$ in (2.4.4),

$$\begin{aligned}\sum_i \alpha_i \log(p_i) &\geq \frac{(1-\delta)}{2I(\mathbf{p}, \mathbf{p}^{\theta_{\mathtt{P}}})} \cdot \sum_i \frac{p_i^{\theta_{\mathtt{P}}} \log(p_i)}{\phi_{\mathbf{p}}(\theta_{\mathbf{p}})} - O\left(\frac{1}{\log N}\right) \\ &= \frac{-1+\delta}{2} - O\left(\frac{1}{\log N}\right). \end{aligned} \tag{2.6.7}$$

□

Proposition 2.6.5. *As $N \to \infty$ we have $|H| \to \infty$ and $|\partial H| = O(|H|^{\frac{1}{2}})$. More precisely*

$$|H| = N^{1+\sum_{i=0}^{k-1} \alpha_i \log(p_i) + \alpha_{\mathtt{tot}} H(\alpha_0,\dots,\alpha_{k-1}) + o(1)}.$$

Proof. For each $x \in \mathtt{Pre}_{\mathtt{CL}}$, Lemma 2.6.4 shows

$$\lambda_x = N^{\sum_{i=0}^{k-1} \alpha_i \log(p_i) \pm o(1)} = N^{\frac{-1+\delta}{2} \pm o(1)}.$$

Moreover it is easy to see that

$$\lfloor N\lambda_x \rfloor \le |\mathbb{Z} \cap NJ_x| \le \lceil N\lambda_x \rceil. \tag{2.6.8}$$

This immediately implies $|H| \to \infty$ as $\mathtt{Pre}_{\mathtt{CL}}$ is non-empty. For the precise asymptotics, Proposition 2.4.2 implies

$$|\mathtt{Pre}_{\mathtt{CL}}| = \binom{\alpha_{\mathtt{tot}} \log(N)}{\alpha_0 \log(N), \ldots, \alpha_{k-1} \log(N)} = N^{\alpha_{\mathtt{tot}} H(\alpha_0, \ldots, \alpha_{k-1}) + o(1)}.$$

As the discrete sets $(\mathbb{Z} \cap NJ_x)_{x \in \mathtt{Pre}_{\mathtt{CL}}}$ are disjoint, they have total size at most N. By (2.6.8) these sets individually have size $N^{\frac{1+\delta}{2}+o(1)}$, and so $|\mathtt{Pre}_{\mathtt{CL}}| \le N^{\frac{1-\delta}{2}+o(1)}$. This means the number of connected components of H is smaller than the size of each component, hence $|\partial H| = O(|H|^{\frac{1}{2}})$. □

2.6.2 Lower Bounding the Number of G-Edges Inside H

It remains to show that H contains many G-edges with high probability. The next lemma shows that with high probability, all appearances of collision-likely strings are inside H, so that it suffices to simply count edges $(i, i+1)$ with $s_i = s_{i+1} \in \mathtt{CL}$. The reason is simply that the requirements $s[M+1] = 0$ and $s[M+2] = 1$ effectively refine collision-likely prefixes $x \in \mathtt{Pre}_{\mathtt{CL}}$ to $x01$. B_{x01} is deep enough inside B_x to overcome the small fluctuations of $\mathcal{I}(B_x)$ vs NJ_x.

Lemma 2.6.6. *With probability $1 - o(1)$, all $i \in [N]$ with $s_i \in \mathtt{CL}$ satisfy $i \in H$.*

Proof. The Dvoretzky–Kiefer–Wolfowitz–Massart inequality [DKW56, Mas90] implies that with probability $1 - o(1)$, all $y \in [k]_0^M$ for $0 \le M \le K$ simultaneously satisfy

$$|\iota(y) - Nt_y| \le N^{\frac{1}{2}+\frac{\delta}{10}}, \qquad |\tau(y) - N(t_y + \lambda_x)| \le N^{\frac{1}{2}+\frac{\delta}{10}}. \tag{2.6.9}$$

We assume the inequalities (2.6.9) hold for all y and show the conclusion under this assumption. Fixing a collision-likely string s with collision-likely prefix x, we apply (2.6.9) with $y = x$ and $y = x01$. Here $x01$ denotes concatenation. Using (2.6.7), we obtain

$$\min(\lambda_x, \lambda_{x01}, \lambda_{x1}) \ge \lambda_x p_{\mathtt{min}}^2 \ge \Omega\left(N^{\frac{-1+\delta}{2}}\right).$$

Therefore

$$N(t_{x01} - t_x) = N\lambda_{x01} \geq \Omega\left(N^{\frac{1+\delta}{2}}\right),$$
$$N\big(t_x + \lambda_x - t_{x01} - \lambda_{x01}\big) = N\lambda_{x1} \geq \Omega\left(N^{\frac{1+\delta}{2}}\right).$$

By the triangle inequality,

$$\begin{aligned}\iota(x01) &\geq Nt_x + N(t_{x01} - t_x) - |\iota(x01) - t_{x01}| \\ &\geq Nt_x + \Omega\left(N^{\frac{1+\delta}{2}}\right) - N^{\frac{1}{2}+\frac{\delta}{10}} \\ &\geq Nt_x\end{aligned}$$

and

$$\begin{aligned}\tau(x01) &\leq N\big(t_x + \lambda_x\big) + N\big(t_{x01} + \lambda_{x01} - t_x - \lambda_x\big) + |\tau(x01) - t_{x01} - \lambda_{x01}| \\ &\leq N\big(t_x + \lambda_x\big) - \Omega\left(N^{\frac{1+\delta}{2}}\right) - N^{\frac{1}{2}+\frac{\delta}{10}} \\ &\leq N\big(t_x + \lambda_x\big).\end{aligned}$$

Altogether if (2.6.9) holds for all y, then all $x \in \mathtt{Pre}_{\mathtt{CL}}$ satisfy

$$Nt_x \leq \iota(x01) \leq \tau(x01) \leq N(t_x + \lambda_x).$$

Therefore $s_i \in B_{x01}$ implies $i \in H$, which completes the proof. □

Define the constant

$$\gamma \equiv 2 + 2\sum_{i=0}^{k-1}(\alpha_i + \beta_i)\log(p_i) + \alpha_{\mathtt{tot}}H\left(\alpha_0, \ldots, \alpha_{k-1}\right) + \beta_{\mathtt{tot}}H(\beta_0, \ldots, \beta_{k-1}).$$

We next give another important numerical lemma, which up to $O(\delta)$ terms will ensure that the number N^γ of edges in H is large enough for Proposition 2.6.1 to apply. (It is only important that $\frac{\psi_{\mathbf{P}}(2)}{2}\varepsilon$ is positive below.)

Lemma 2.6.7. *With α_i, β_i and γ as defined above,*

$$\gamma \geq \frac{1}{2}\left(1 + \sum_{i=0}^{k-1}\alpha_i\log(p_i) + \alpha_{\mathtt{tot}}H\left(\alpha_0, \ldots, \alpha_{k-1}\right)\right) + \frac{\psi_{\mathbf{p}}(2)}{2}\varepsilon. \tag{2.6.10}$$

Proof of Lemma 2.6.7. Recall the following definitions and identities.

- $\psi_{\mathbf{p}}(t) = -\log \phi_{\mathbf{p}}(t) = -\log\left(\sum_{i=0}^{k-1} p_i^t\right) > 0$ for any $t > 1$.
- $\psi_{\mathbf{p}}(\theta_{\mathbf{p}}) = 2\psi_{\mathbf{p}}(2)$.
- $C_{\mathbf{p}} = \frac{3+\theta_{\mathbf{p}}}{4\psi_{\mathbf{p}}(2)} = \frac{3+\theta_{\mathbf{p}}}{2\psi_{\mathbf{p}}(\theta_{\mathbf{p}})}$.
- $I(\mathbf{p}, \mathbf{p}^t) = -\sum_i \frac{p_i^t \log(p_i)}{\phi_{\mathbf{p}}(t)}$.
- $H(\mathbf{p}^t) = tI(\mathbf{p}, \mathbf{p}^t) - \psi_{\mathbf{p}}(t)$ for any $t > 0$.
- $\alpha_{\mathtt{tot}} = \frac{1-\delta}{2I(\mathbf{p}, \mathbf{p}^{\theta_{\mathbf{p}}})} \pm o(1)$.
- $\alpha_{\mathtt{tot}} + \beta_{\mathtt{tot}} \leq C_{\mathbf{p}} - \varepsilon$.
- $\alpha_i = (\mathbf{p}^{\theta_{\mathbf{p}}})_i \cdot \alpha_{\mathtt{tot}} \pm o(1)$
- $\beta_i = (\mathbf{p}^2)_i \cdot \beta_{\mathtt{tot}} \pm o(1)$

After rearranging (2.6.10) and multiplying by 2, it suffices to show

$$3 + \sum_{i=0}^{k-1} (3\alpha_i + 4\beta_i) \log(p_i) + \alpha_{\mathtt{tot}} H(\alpha_0, \ldots, \alpha_{k-1}) + 2\beta_{\mathtt{tot}} H(\beta_0, \ldots, \beta_{k-1}) \stackrel{?}{\geq} \psi_{\mathbf{p}}(2)\varepsilon.$$

First, replacing both entropy terms using $H(\mathbf{p}^t) = tI(\mathbf{p}, \mathbf{p}^t) - \psi_{\mathbf{p}}(t)$ and then substituting $\psi_{\mathbf{p}}(\theta_{\mathbf{p}}) = 2\psi_{\mathbf{p}}(2)$ reduces us to showing

$$3 + \sum_{i=0}^{k-1} (3\alpha_i + 4\beta_i) \log(p_i) + \alpha_{\mathtt{tot}}(\theta_{\mathbf{p}} I(\mathbf{p}, \mathbf{p}^{\theta_{\mathbf{p}}}) - 2\psi_{\mathbf{p}}(2)) + 2\beta_{\mathtt{tot}}(2I(\mathbf{p}, \mathbf{p}^2) - \psi_{\mathbf{p}}(2)) \stackrel{?}{\geq} \psi_{\mathbf{p}}(2)\varepsilon.$$

Using $\alpha_{\mathtt{tot}} + \beta_{\mathtt{tot}} = \frac{K}{\log N} \leq C_{\mathbf{p}} - \varepsilon$, it remains to prove

$$3 + \sum_{i=0}^{k-1} (3\alpha_i + 4\beta_i) \log(p_i) + \theta_{\mathbf{p}} \alpha_{\mathtt{tot}} I(\mathbf{p}, \mathbf{p}^{\theta_{\mathbf{p}}}) + 4\beta_{\mathtt{tot}} I(\mathbf{p}, \mathbf{p}^2) - 2\psi_{\mathbf{p}}(2) C_{\mathbf{p}} \stackrel{?}{\geq} -\psi_{\mathbf{p}}(2)\varepsilon.$$

Substituting $C_{\mathbf{p}} = \frac{3+\theta_{\mathbf{p}}}{4\psi_{\mathbf{p}}(2)}$ and $\alpha_{\mathtt{tot}} = \frac{1-\delta}{2I(\mathbf{p}, \mathbf{p}^{\theta_{\mathbf{p}}})} + o(1)$ we are reduced to showing

$$\frac{3}{2} + \sum_{i=0}^{k-1} (3\alpha_i + 4\beta_i) \log(p_i) + 4\beta_{\mathtt{tot}} I(\mathbf{p}, \mathbf{p}^2) \stackrel{?}{\geq} -\psi_{\mathbf{p}}(2)\varepsilon + O(\delta) + o(1). \tag{2.6.11}$$

Now, using $I(\mathbf{p}, \mathbf{p}^{\theta_{\mathbf{P}}}) = -\sum_i \frac{p_i^{\theta_{\mathbf{P}}} \log(p_i)}{\phi_{\mathbf{p}}(\theta_{\mathbf{p}})}$ allows us to simplify

$$\sum_i \alpha_i \log(p_i) = \alpha_{\mathtt{tot}} \sum_i \frac{p_i^{\theta_{\mathbf{P}}} \log(p_i)}{\phi_{\mathbf{p}}(\theta_{\mathbf{p}})} + o(1) = -\frac{1-\delta}{2} + o(1).$$

Furthermore,

$$\beta_{\mathtt{tot}} I(\mathbf{p}, \mathbf{p}^2) = -\beta_{\mathtt{tot}} \sum_{i=0}^{k-1} \frac{p_i^2 \log(p_i)}{\phi_{\mathbf{p}}(2)} = -\sum_i \beta_i \log(p_i) + o(1).$$

Substituting these near-equalities into (2.6.11), it suffices to show

$$0 \overset{?}{\geq} -\psi_{\mathbf{p}}(2)\varepsilon + O(\delta) + o(1).$$

Recalling that $\delta = \delta(\mathbf{p}, \varepsilon)$ was chosen sufficiently small completes the proof. □

Lemma 2.6.8. *With probability* $1 - o(1)$*, at least* $N^{\gamma-\delta}$ *distinct* $s \in \mathtt{CL}$ *appear* 2 *or more times in the* $\mathbf{p}$*-random sequence* $S = (s_1, \ldots, s_N) \in \mathcal{S}$*.*

Proof. By Proposition 2.4.2, there are

$$|\mathtt{CL}| = N^{\alpha_{\mathtt{tot}} H(\alpha_0, \ldots, \alpha_{k-1}) + \beta_{\mathtt{tot}} H(\beta_0, \ldots, \beta_{k-1}) + o(1)}$$

collision-likely strings, each of which occurs $\mathtt{Bin}\left(N, N^{\sum_{i=0}^{k-1}(\alpha_i+\beta_i)\log(p_i)}\right)$ times in S. Because $(\widetilde{C}_{\mathbf{p}} + \varepsilon)\log N \leq K$ holds (recall (2.6.1)) and $\log(p_i) \leq \log(p_{\mathtt{max}}) < 0$ for all i, we obtain

$$\begin{aligned}\sum_{i=0}^{k-1}(\alpha_i + \beta_i)\log(p_i) &\leq \frac{K \log(p_{\mathtt{max}})}{\log N} \\ &\leq (\widetilde{C}_{\mathbf{p}} + \varepsilon)\log(p_{\mathtt{max}}) \\ &\leq -1 - \delta\end{aligned}$$

for $\delta = \delta(\mathbf{p}, \varepsilon)$ sufficiently small. This implies

$$\left(1 - N^{\sum_{i=0}^{k-1}(\alpha_i+\beta_i)\log(p_i)}\right)^N = \Omega(1).$$

Next for each $s \in \mathtt{CL}$, let Y_s denote the event that s appears at least twice in S. By the binomial distribution formula, each $s \in \mathtt{CL}$ satisfies

$$\mathbb{P}[Y_s] \geq \binom{N}{2} N^{2\sum_{i=0}^{k-1}(\alpha_i+\beta_i)\log(p_i)} \cdot \Omega(1) = N^{2+2\sum_{i=0}^{k-1}(\alpha_i+\beta_i)\log(p_i)+o(1)}.$$

Letting $Y_{\mathtt{tot}} = \sum_{s \in \mathtt{CL}} 1_{Y_s}$ and estimating $|\mathtt{CL}|$ with Proposition 2.4.2, we get

$$\mathbb{E}[Y_{\mathtt{tot}}] \geq N^{\gamma - o(1)}.$$

We claim that the Bernoulli random variables $(1_{Y_s})_{s \in \mathtt{CL}}$ are pairwise non-positively correlated, i.e.

$$\mathbb{P}[Y_s \cap Y_{s'}] \leq \mathbb{P}[Y_s] \cdot \mathbb{P}[Y_{s'}], \quad s \neq s'.$$

Indeed for any collision-likely strings $s \neq s'$, set $n_{s'} \in \mathbb{Z}_{\geq 0}$ to be the number of i such that $s_i = s'$. It is easy to see that $\mathbb{P}[Y_s | n_{s'}]$ is decreasing in $n_{s'}$, which implies the claim.

From Lemmas 2.6.4 and 2.6.7 it follows that $\gamma > \frac{1}{4}$ for N large enough. Therefore for large N, we have

$$\mathbb{E}[Y_{\mathtt{tot}}] \geq \Omega(N^{1/4}).$$

As argued just above, $Y_{\mathtt{tot}}$ is a sum of Bernoulli random variables Y_s with pairwise non-positive correlations, which implies that $Y_{\mathtt{tot}}$ has smaller variance than expectation. In particular

$$\mathrm{Var}(Y_{\mathtt{tot}}) \leq \mathbb{E}[Y_{\mathtt{tot}}] \leq O\left(\frac{\mathbb{E}[Y_{\mathtt{tot}}]^2}{N^{1/4}}\right).$$

Chebychev's inequality now completes the proof as

$$\begin{aligned}
\mathbb{P}\left[Y_{\mathtt{tot}} \geq N^{\gamma-\delta}\right] &\geq \mathbb{P}\left[Y_{\mathtt{tot}} \geq \frac{1}{2} \cdot \mathbb{E}\left[Y_{\mathtt{tot}}\right]\right] \\
&\geq 1 - \frac{4 \cdot \mathrm{Var}(Y_{\mathtt{tot}})}{\mathbb{E}[Y_{\mathtt{tot}}]^2} \\
&\geq 1 - O(N^{-1/4}).
\end{aligned}$$

□

We are finally ready to establish the mixing time lower bound (2.1.2) in Theorem 8.

Proof of (2.1.2). By Lemmas 2.6.7 and 2.6.8, with probability $1 - o(1)$ at least $N^{\gamma-\delta} \geq |H|^{\frac{1}{2}+\delta}$ strings $s \in \mathtt{CL}$ appear at least twice in S. Each such s by definition results in an edge $(i, i+1) \in E(G)$ with $s_i = s_{i+1} = s$. Moreover Lemma 2.6.6 implies that with probability $1 - o(1)$, all of these edges appear inside H. Then by Lemma 2.6.7,

$$|E(G) \cap E(H)| \geq |H|^{\frac{1}{2}+\Omega_{\mathbf{p}}(\varepsilon)} \geq |H|^{\frac{1}{2}+\delta}$$

also holds with probability $1 - o(1)$. Combined with Proposition 2.6.5, it follows that H satisfies the conditions of Proposition 2.6.1. This completes the proof. □

Chapter 3

Algorithmic Stochastic Localization for the Sherrington-Kirkpatrick Model

3.1 Introduction

This Sherrington-Kirkpatrick (SK) Gibbs measure is the probability distribution on $\Sigma_N = \{-1, +1\}^N$ given by

$$\mu_{\boldsymbol{A}}(\boldsymbol{x}) = \frac{1}{Z(\beta, \boldsymbol{A})} \exp\Big\{\frac{\beta}{2}\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle\Big\}, \tag{3.1.1}$$

where $\beta \geq 0$ is an inverse temperature parameter and $\boldsymbol{A} \sim \mathsf{GOE}(N)$; i.e., $\boldsymbol{A}$ is symmetric. This means that $A_{ij} \sim \mathcal{N}(0, 1/N)$ are i.i.d. for $1 \leq i \leq j \leq N$, and the diagonal entries $A_{ii} \sim \mathcal{N}(0, 2/N)$ are i.i.d. for $1 \leq i \leq N$. The parameter β is fixed and we will leave implicit the dependence of μ upon β, unless mentioned otherwise.

In this chapter, we consider the problem of efficiently sampling from the measure (3.1.1). Namely, we seek a randomized algorithm that accepts as input $\boldsymbol{A}$ and generates $\boldsymbol{x}^{\mathrm{alg}} \sim \mu_{\boldsymbol{A}}^{\mathrm{alg}}$, such that:

1. The algorithm runs in polynomial time for any $\boldsymbol{A}$.
2. The distribution $\mu_{\boldsymbol{A}}^{\mathrm{alg}}$ is close to $\mu_{\boldsymbol{A}}$ for typical realizations of $\boldsymbol{A}$. Given a bounded distance $\mathsf{dist}(\mu, \nu)$ between probability distributions μ, ν, this can be formalized by requiring

$$\mathbb{E}[\mathsf{dist}(\mu_{\boldsymbol{A}}, \mu_{\boldsymbol{A}}^{\mathrm{alg}})] = o_N(1).$$

Gibbs sampling (also known in this context as Glauber dynamics) provides an algorithm to approximately sample from $\mu_{\boldsymbol{A}}$. However, standard techniques to bound its mixing time (e.g., Dobrushin condition [AH87]) only imply polynomial mixing for a vanishing interval of temperatures $\beta = O(N^{-1/2})$. By contrast, physicists [SZ81, MPV87] predict fast convergence to equilibrium (at least for certain observables) for all $\beta < 1$.

Significant progress on this question was achieved only recently. In [BB19], Bauerschmidt and Bodineau showed that, for $\beta < 1/4$, the measure $\mu_{\boldsymbol{A}}$ can be decomposed into a log-concave mixture of product measures. They use this decomposition to prove that $\mu_{\boldsymbol{A}}$ satisfies a log-Sobolev inequality, although not for the Dirichlet form of Glauber dynamics[1]. Eldan, Koehler, Zeitouni [EKZ21] prove that, in the same region $\beta < 1/4$, $\mu_{\boldsymbol{A}}$ satisfies a Poincaré inequality for the Dirichlet form of Glauber dynamics. Hence Glauber dynamics mixes in $O(N^2)$ spin flips in total variation distance. This mixing time estimate was improved to $O(N \log N)$ by [AJK$^+$21] using a modified log Sobolev inequality, see also [CE22, Corollary 51]. The aforementioned results apply deterministically to any matrix $\boldsymbol{A}$ satisfying $\beta(\lambda_{\max}(\boldsymbol{A}) - \lambda_{\min}(\boldsymbol{A})) \le 1 - \varepsilon$ (for some constant $\varepsilon > 0$).

For *spherical* spin glasses, it is shown in [GJ19] that Langevin dynamics have a polynomial spectral gap at high temperature. Meanwhile [BAJ18] proves that at sufficiently low temperature and under an overlap gap condition, the mixing times of Glauber and Langevin dynamics are exponentially large in Ising and spherical spin glasses, respectively.

In this chapter we develop a different approach which is not based on a Monte Carlo Markov Chain strategy. We build on the well known remark that approximate sampling can be reduced to approximate computation of expectations of the measure $\mu_{\boldsymbol{A}}$, and of a family of measures obtained from $\mu_{\boldsymbol{A}}$. One well known method to achieve this reduction is via sequential sampling [JVV86, CDHL05, BD11]. A sequential sampling approach to $\mu_{\boldsymbol{A}}$ would proceed as follows. Order the variables $x_1, \dots, x_N \in \{-1, +1\}$ arbitrarily. At step i compute the marginal distribution of x_i, conditional to $x_1, \dots, x_{i-1}$ taking the previously chosen values: $p_s^{(i)} := \mu_{\boldsymbol{A}}(x_i = s | x_1, \dots, x_{i-1})$, $s \in \{-1, +1\}$. Fix $x_i = +1$ with probability $p_{+1}^{(i)}$ and $x_i = -1$ with probability $p_{-1}^{(i)}$.

We follow a different route, which is similar in spirit, but that we find more convenient technically, and of potential practical interest. Our approach is motivated by the stochastic localization process [Eld20]. Given any probability measure μ on $\mathbb{R}^N$ with finite second moment, positive time $t > 0$, and vector $\boldsymbol{y} \in \mathbb{R}^N$, define the tilted measure

$$\mu_{\boldsymbol{y},t}(\mathrm{d}\boldsymbol{x}) := \frac{1}{Z(\boldsymbol{y})} e^{\langle \boldsymbol{y}, \boldsymbol{x} \rangle - \frac{t}{2}\|\boldsymbol{x}\|_2^2} \, \mu(\mathrm{d}\boldsymbol{x}) \,, \tag{3.1.2}$$

and let its mean vector be

$$\mathbf{m}(\boldsymbol{y}, t) := \int_{\mathbb{R}^N} \boldsymbol{x} \, \mu_{\boldsymbol{y},t}(\mathrm{d}\boldsymbol{x}) \,. \tag{3.1.3}$$

[1] We note in passing that their result immediately suggests a sampling algorithm: sample from the log-concave mixture using Langevin dynamics, and then sample from the corresponding component using the product form.

Consider the stochastic differential equation[2] (SDE)

$$\mathrm{d}\boldsymbol{y}(t) = \mathbf{m}(\boldsymbol{y}(t), t)\mathrm{d}t + \mathrm{d}\boldsymbol{B}(t), \qquad \boldsymbol{y}(0) = 0\,, \tag{3.1.4}$$

where $(\boldsymbol{B}(t))_{t\geq 0}$ is a standard Brownian motion in $\mathbb{R}^N$. Then, the measure-valued process $(\mu_{\boldsymbol{y}(t),t})_{t\geq 0}$ is a martingale and (almost surely) $\mu_{\boldsymbol{y}(t),t} \Rightarrow \delta_{\boldsymbol{x}^*}$ as $t \to \infty$, for some random $\boldsymbol{x}^*$ (i.e. the measure localizes). As a consequence of the martingale property, $\mathbb{E}[\int \varphi(\boldsymbol{x})\mu_{\boldsymbol{y}(t),t}(\mathrm{d}\boldsymbol{x})]$ is a constant for any bounded continuous function φ, whence $\mathbb{E}[\varphi(\boldsymbol{x}^*)] = \int \varphi(\boldsymbol{x})\mu(\mathrm{d}\boldsymbol{x})$. In other words, $\boldsymbol{x}^*$ is a sample from μ. For further information on this process, we refer to Section 3.3.

In order to use this process as an algorithm to sample from the SK measure $\mu = \mu_{\boldsymbol{A}}$, we need to overcome two problems:

- *Discretization.* We need to discretize the SDE (3.1.4) in time, and still guarantee that the discretization closely tracks the original process. This is of course possible only if the map $\boldsymbol{y} \mapsto \mathbf{m}(\boldsymbol{y}, t)$ is sufficiently regular.
- *Mean computation.* We need to be able to compute the mean vector $\mathbf{m}(\boldsymbol{y}, t)$ efficiently. To this end, we use an approximate message passing (AMP) algorithm for which we can leverage earlier work [DAM17] to establish that $\|\mathbf{m}(\boldsymbol{y}) - \hat{\boldsymbol{m}}_{\mathsf{AMP}}(\boldsymbol{y})\|_2^2/N = o_N(1)$ along the algorithm trajectory. (Note that the SK measure is supported on vectors with $\|\boldsymbol{x}\|_2^2 = N$, and hence the quadratic component of the tilt in Eq. (3.1.2) drops out. We will therefore write $\mathbf{m}(\boldsymbol{y})$ or $\mathbf{m}(\boldsymbol{A}, \boldsymbol{y})$ instead $\mathbf{m}(\boldsymbol{y}, t)$ for the mean of the Gibbs measure.)

To our knowledge, ours is the first algorithmic implementation of the stochastic localization process, although a recent paper by Nam, Sly and Zhang [NSZ22] uses this process (without naming it as such) to show that the Ising measure on the infinite regular tree is a factor of IID process up to a constant factor away from the Kesten–Stigum, or "reconstruction", threshold. Their construction can easily be transformed into a sampling algorithm.

In order to state our results, we define the normalized 2-Wasserstein distance between two probability measures μ, ν on $\mathbb{R}^N$ with finite second moments as

$$W_{2,N}(\mu,\nu)^2 = \inf_{\pi\in\mathcal{C}(\mu,\nu)} \frac{1}{N}\mathop{\mathbb{E}}_{\pi}\left[\left\|\boldsymbol{X}-\boldsymbol{Y}\right\|_2^2\right], \tag{3.1.5}$$

where the infimum is over all couplings $(\boldsymbol{X}, \boldsymbol{Y}) \sim \pi$ with marginals $\boldsymbol{X} \sim \mu$ and $\boldsymbol{Y} \sim \nu$.

In this chapter, we establish two main results.

Sampling algorithm for $\beta < 1/2$. We prove that the strategy outlined above yields an algorithm with complexity $O(N^2)$, which samples from a distribution $\mu_{\boldsymbol{A}}^{\mathrm{alg}}$ with $W_{2,N}(\mu_{\boldsymbol{A}}^{\mathrm{alg}}, \mu_{\boldsymbol{A}}) = o_{N,\mathbb{P}}(1)$.

[2] If μ is has finite variance, then $\boldsymbol{y} \to \mathbf{m}(\boldsymbol{y}, t)$ is Lipschitz and so this SDE is well posed with unique strong solution.

Hardness for stable algorithms, for $\beta > 1$. We prove that no algorithm satisfying a certain *stability* property can sample from the SK measure (under the same criterion $W_{2,N}(\mu_{\boldsymbol{A}}^{\mathrm{alg}}, \mu_{\boldsymbol{A}}) = o_{N,\mathbb{P}}(1)$) for $\beta > 1$, i.e., when replica symmetry is broken. Roughly speaking, stability formalizes the notion that the algorithm output behaves continuously with respect to $\boldsymbol{A}$.

It is worth pointing out that we expect our algorithm to be successful (in the sense described above) for all $\beta < 1$ and that closing the gap between $\beta = 1/2$ and $\beta = 1$ should be within reach of existing techniques, at the price of a longer technical argument. We expound on this point in Remark 3.2.1 further below, and in Section 3.7.

The hardness results for $\beta > 1$ are proven using the notion of disorder chaos, in a similar spirit to the use of the *overlap gap property* for random optimization, estimation, and constraint satisfaction problems [GS14, RV17a, GS17, CGPR19, GJ21, GJW20a, Wei22, GK21a, BH21, GJW21, HS22]. While the overlap gap property has been used to rule out stable algorithms for this class of problems, and variants have been used to rule out efficient sampling by specific Markov chain algorithms, to the best of our knowledge we are the first to rule out stable sampling algorithms using these ideas. In sampling there is no hidden solution or set of solutions to be found, and therefore no notion of an overlap gap in the most natural sense. Instead, we argue directly that the distribution to be sampled from is unstable in a $W_{2,N}$ sense at low temperature, and hence cannot be approximated by any stable algorithm.

The rest of the chapter is organized as follows. In Section 3.2 we formally state our results. In Section 3.3 we collect some useful properties of the stochastic localization process, and we present the analysis of our algorithm in Section 3.4. Finally, the proof of hardness under stability is given in Section 3.9.

3.2 Main Results

3.2.1 Sampling algorithm for $\beta < 1/2$

In this section we describe the sampling algorithm, and formally state the result of our analysis. As pointed out in the introduction, a main component is the computation of the mean of the tilted SK measure:

$$\mu_{\boldsymbol{A},\boldsymbol{y}}(\boldsymbol{x}) := \frac{1}{Z(\boldsymbol{A},\boldsymbol{y})} \exp\Big\{\frac{\beta}{2}\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle + \langle \boldsymbol{y}, \boldsymbol{x}\rangle\Big\}, \quad \boldsymbol{x} \in \{-1,+1\}^N. \tag{3.2.1}$$

We describe the algorithm to approximate this mean in Section 3.2.1, the overall sampling procedure (which uses this estimator as a subroutine) in Section 11, and our Wasserstein-distance guarantee in Section 10.

Approximating the mean of the Gibbs measure

Algorithm 1: MEAN OF THE TILTED GIBBS MEASURE

Input: Data $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, $\boldsymbol{y} \in \mathbb{R}^N$, parameters $\beta, \eta > 0$, $q \in (0,1)$, iteration numbers K_{AMP}, K_{NGD}.

$\hat{\boldsymbol{m}}^{-1} = \boldsymbol{z}^0 = 0$,
for $k = 0, \cdots, K_{\mathsf{AMP}} - 1$ **do**
  $\hat{\boldsymbol{m}}^k = \tanh(\boldsymbol{z}^k)$, $\quad \mathsf{b}_k = \frac{\beta^2}{N}\sum_{i=1}^N \left(1 - \tanh^2(z_i^k)\right)$,
  $\boldsymbol{z}^{k+1} = \beta \boldsymbol{A}\hat{\boldsymbol{m}}^k + \boldsymbol{y} - \mathsf{b}_k \hat{\boldsymbol{m}}^{k-1}$,
end
$\boldsymbol{u}^0 = \boldsymbol{z}^{K_{\mathsf{AMP}}}$,
for $k = 0, \cdots, K_{\mathsf{NGD}} - 1$ **do**
  $\boldsymbol{u}^{k+1} = \boldsymbol{u}^k - \eta \cdot \nabla \mathscr{F}_{\mathrm{TAP}}(\hat{\boldsymbol{m}}^{+,k}; \boldsymbol{y}, q)$,
  $\hat{\boldsymbol{m}}^{+,k+1} = \tanh(\boldsymbol{u}^{k+1})$,
end
return $\hat{\boldsymbol{m}}^{+,K_{\mathsf{NGD}}}$

We will denote our approximation of the mean of the Gibbs measure $\mu_{\boldsymbol{A},\boldsymbol{y}}$ by $\hat{\boldsymbol{m}}(\boldsymbol{A}, \boldsymbol{y})$, while the actual mean will be $\mathbf{m}(\boldsymbol{A}, \boldsymbol{y})$.

The algorithm to compute $\hat{\boldsymbol{m}}(\boldsymbol{A}, \boldsymbol{y})$ is given in Algorithm 1, and is composed of two phases:

1. An Approximate Message Passing (AMP) algorithm is run for K_{AMP} iterations and constructs a first estimate of the mean. We denote by $\mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; k)$ the estimate produced after k AMP iterations
$$\mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; k) := \hat{\boldsymbol{m}}^k \,. \tag{3.2.2}$$

2. Natural gradient descent (NGD) is run for K_{NGD} iterations with initialization given by vector computed at the end of the first phase. This phase attempts to minimize the following version of the TAP free energy (for a specific value of q):
$$\mathscr{F}_{\mathrm{TAP}}(\mathbf{m}; \boldsymbol{y}, q) := -\frac{\beta}{2}\langle \mathbf{m}, \boldsymbol{A}\mathbf{m}\rangle - \langle \boldsymbol{y}, \mathbf{m}\rangle - \sum_{i=1}^N h(m_i) - \frac{N\beta^2(1-q)(1+q-2Q(\mathbf{m}))}{4}\,, \tag{3.2.3}$$
$$Q(\mathbf{m}) = \frac{1}{N}\|\mathbf{m}\|^2, \qquad h(m) = -\frac{1+m}{2}\log\left(\frac{1+m}{2}\right) - \frac{1-m}{2}\log\left(\frac{1-m}{2}\right)\,. \tag{3.2.4}$$

The second stage is motivated by the TAP (Thouless-Anderson-Palmer) equations for the Gibbs mean of a high-temperature spin glass [MPV87, Tal10]. Essentially by construction, stationary points for the function $\mathscr{F}_{\mathrm{TAP}}(\mathbf{m}; \boldsymbol{y}, q)$ satisfy the TAP equations, and we show in Lemma 3.7.2 that the first stage above constructs an approximate stationary point for $\mathscr{F}_{\mathrm{TAP}}(\mathbf{m}; \boldsymbol{y}, q)$. The effect of

the second stage is therefore numerically small, but it turns out to reduce the error incurred by discretizing time in line 6 of Algorithm 2.

Let us emphasize that this two-stage construction is considered for technical reasons. Indeed a simpler algorithm, that runs AMP for a larger number of iteration, and does not run NGD at all, is expected to work but our arguments do not go through. The hybrid algorithm above allows us to exploit known properties of AMP (precise analysis via state evolution) and of $\mathscr{F}_{\text{TAP}}(\mathbf{m}; \boldsymbol{y}, q)$ (Lipschitz continuity of the minimizer in $\boldsymbol{y}$).

Sampling via stochastic localization

Algorithm 2: Approximate sampling from the SK Gibbs measure

Input: Data $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, parameters $(\beta, \eta, K_{\mathsf{AMP}}, K_{\mathsf{NGD}}, L, \delta)$

1 $\hat{\boldsymbol{y}}_0 = 0$,
2 **for** $\ell = 0, \cdots, L-1$ **do**
3 Draw $\boldsymbol{w}_{\ell+1} \sim \mathcal{N}(0, \boldsymbol{I}_N)$ independent of everything so far;
4 Set $q = q_*(\beta, t = \ell\delta)$;
5 Set $\hat{\boldsymbol{m}}(\boldsymbol{A}, \hat{\boldsymbol{y}}_\ell)$ the output of Algorithm 1, with parameters $(\beta, \eta, q, K_{\mathsf{AMP}}, K_{\mathsf{NGD}})$;
6 Update $\hat{\boldsymbol{y}}_{\ell+1} = \hat{\boldsymbol{y}}_\ell + \hat{\boldsymbol{m}}(\boldsymbol{A}, \hat{\boldsymbol{y}}_\ell)\,\delta + \sqrt{\delta}\,\boldsymbol{w}_{\ell+1}$
7 **end**
8 Set $\hat{\boldsymbol{m}}(\boldsymbol{A}, \hat{\boldsymbol{y}}_L)$ the output of Algorithm 1, with parameters $(\eta, q, K_{\mathsf{AMP}}, K_{\mathsf{NGD}})$;
9 Draw $\{x_i^{\text{alg}}\}_{i\le N}$ conditionally independent with $\mathbb{E}[x_i^{\text{alg}} | \boldsymbol{y}, \{\boldsymbol{w}_\ell\}] = \widehat{m}_i(\boldsymbol{A}, \hat{\boldsymbol{y}}_L)$
10 **return** $\boldsymbol{x}^{\text{alg}}$

Our sampling algorithm is presented as Algorithm 2. The algorithm makes uses of constants $q_k := q_k(\beta, t)$. With $W \sim \mathcal{N}(0,1)$ a standard Gaussian, these constants are defined for $k, \beta, t \ge 0$ by the recursion

$$q_{k+1} = \mathbb{E}\left[\tanh\left(\beta^2 q_k + t + \sqrt{\beta^2 q_k + t}\, W\right)^2\right], \quad q_0 = 0\,, \quad q_* = \lim_{k\to\infty} q_k\,. \tag{3.2.5}$$

This iteration can be implemented via a one-dimensional integral, and the limit q_* is approached exponentially fast in k (see Lemma 3.6.3 below). The values $q_*(\beta, t = \ell\delta)$ for $\ell \in \{0, \dots, L\}$ can be precomputed and are independent of the input $\boldsymbol{A}$. For the sake of simplicity, we will neglect errors in this calculation.

The core of the sampling procedure is step 6, which is a standard Euler discretization of the SDE (3.1.4), with step size δ, over the time interval $[0, T]$, $T = L\delta$. The mean of the Gibbs measure $\mathbf{m}(\boldsymbol{A}, \boldsymbol{y})$ is replaced by the output of Algorithm 1 which we recall is denoted by $\hat{\boldsymbol{m}}(\boldsymbol{A}, \boldsymbol{y})$. We reproduce the Euler iteration here for future reference

$$\hat{\boldsymbol{y}}_{\ell+1} = \hat{\boldsymbol{y}}_\ell + \hat{\boldsymbol{m}}(\boldsymbol{A}, \hat{\boldsymbol{y}}_\ell)\,\delta + \sqrt{\delta}\,\boldsymbol{w}_{\ell+1}\,. \tag{3.2.6}$$

The output of the iteration is $\hat{\boldsymbol{m}}(\boldsymbol{A}, \hat{\boldsymbol{y}}_L)$, which should be thought of as an approximation of $\mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(T))$, $T = L\delta$, that is the mean of $\mu_{\boldsymbol{A},\boldsymbol{y}(T)}$. According to the discussion in the introduction, for large T, $\mu_{\boldsymbol{A},\boldsymbol{y}(T)}$ concentrates around $\boldsymbol{x}^* \sim \mu_{\boldsymbol{A}}$. In other words, $\mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(T))$ is close to the corner $\boldsymbol{x}^*$ of the hypercube. We round its coordinates independently to produce the output $\boldsymbol{x}^{\rm alg}$.

Theoretical guarantee

Our main positive result is the following.

Theorem 9. *For any $\varepsilon > 0$ and $\beta_0 < 1/2$ there exist $\eta, K_{\mathsf{AMP}}, K_{\mathsf{NGD}}, L, \delta$ independent of N, so that the following holds for all $\beta \le \beta_0$. The sampling algorithm 2 takes as input $\boldsymbol{A}$ and parameters $(\eta, K_{\mathsf{AMP}}, K_{\mathsf{NGD}}, L, \delta)$ and outputs a random point $\boldsymbol{x}^{\rm alg} \in \{-1, +1\}^N$ with law $\mu_{\boldsymbol{A}}^{\rm alg}$ such that with probability $1 - o_N(1)$ over $\boldsymbol{A} \sim \mathsf{GOE}(N)$,*

$$W_{2,N}(\mu_{\boldsymbol{A}}^{\rm alg}, \mu_{\boldsymbol{A}}) \le \varepsilon\,. \tag{3.2.7}$$

The total complexity of this algorithm is $O(N^2)$.

Remark 3.2.1. The condition $\beta < 1/2$ arises because our proof requires the Hessian of the TAP free energy to be positive definite at its minimizer. A simple calculation yields

$$\nabla^2 \mathscr{F}_{\rm TAP}(\mathbf{m}; \boldsymbol{y}, q) = -\beta \boldsymbol{A} + \boldsymbol{D}(\mathbf{m}) + \beta^2(1-q)\,\boldsymbol{I}_N\,, \qquad \boldsymbol{D}(\mathbf{m}) := \mathrm{diag}\big(\{(1-m_i^2)^{-1}\}_{i\le N}\big). \tag{3.2.8}$$

A crude bound yields $\nabla^2 \mathscr{F}_{\rm TAP}(\mathbf{m}; \boldsymbol{y}, q) \succeq -\beta\boldsymbol{A} + \boldsymbol{I}_N \succeq (1 - \beta\lambda_{\max}(\boldsymbol{A}))\boldsymbol{I}_N$. Since $\text{p-lim}_{N\to\infty} \lambda_{\max}(\boldsymbol{A}) = 2$ the desired condition holds trivially for $\beta < 1/2$. However, we expect that a more careful treatment will reveal that the Hessian is locally positive in a neighborhood of the minimizer for all $\beta < 1$.

3.2.2 Hardness for stable algorithms, for $\beta > 1$

The sampling algorithm 2 enjoys stability properties with respect to changes in the inverse temperature β and the matrix $\boldsymbol{A}$ which are shared by many natural efficient algorithms. We will use the fact that the actual Gibbs measure does not enjoy this stability property for $\beta > 1$ to conclude that sampling is hard for all stable algorithms.

Throughout this section, we denote the Gibbs and algorithmic output distributions by $\mu_{\boldsymbol{A},\beta}$ and $\mu_{\boldsymbol{A},\beta}^{\rm alg}$ respectively to emphasize the dependence on β.

Definition 3.2.1. *Let $\{\mathsf{ALG}_N\}_{N\ge1}$ be a family of randomized sampling algorithms, i.e., measurable maps*

$$\mathsf{ALG}_N : (\boldsymbol{A}, \beta, \omega) \mapsto \mathsf{ALG}_N(\boldsymbol{A}, \beta, \omega) \in [-1, 1]^N\,,$$

where ω is a random seed (a point in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$). Let $\boldsymbol{A}'$ and $\boldsymbol{A} \sim \mathsf{GOE}(N)$ be independent copies of the coupling matrix, and consider perturbations $\boldsymbol{A}_s = \sqrt{1-s^2}\boldsymbol{A} + s\boldsymbol{A}'$ for $s \in [0,1]$. Finally, denote by $\mu^{\mathrm{alg}}_{\boldsymbol{A}_s,\beta}$ the law of the algorithm output, i.e., the distribution of $\mathsf{ALG}_N(\boldsymbol{A}_s, \beta, \omega)$ when $\omega \sim \mathbb{P}$ independent of $\boldsymbol{A}_s, \beta$ which are fixed.

We say ALG_N is stable with respect to disorder*, at inverse temperature β, if*

$$\lim_{s\to 0} \operatorname*{p-lim}_{N\to\infty} W_{2,N}(\mu^{\mathrm{alg}}_{\boldsymbol{A},\beta}, \mu^{\mathrm{alg}}_{\boldsymbol{A}_s,\beta}) = 0\,. \tag{3.2.9}$$

We say ALG_N is stable with respect to temperature *at inverse temperature β, if*

$$\lim_{\beta'\to\beta} \operatorname*{p-lim}_{N\to\infty} W_{2,N}(\mu^{\mathrm{alg}}_{\boldsymbol{A},\beta}, \mu^{\mathrm{alg}}_{\boldsymbol{A},\beta'}) = 0\,. \tag{3.2.10}$$

We begin by establishing the stability of the proposed sampling algorithm.

Theorem 10 (Stability of the sampling Algorithm 2)**.** *For any $\beta \in (0,\infty)$ and fixed parameters (η, K_{AMP}, K_{NGD}, L, δ), Algorithm 2 is stable with respect to disorder and with respect to temperature.*

This theorem is proved in Section 3.9.1. As a consequence, the Gibbs measures $\mu_{\boldsymbol{A},\beta}$ enjoy similar stability properties for $\beta < 1/2$, which amount (as discussed below) to the absence of chaos in both temperature and disorder:

Corollary 3.2.2. *For any $\beta < 1/2$, the following properties hold for the Gibbs measure $\mu_{\boldsymbol{A},\beta}$ of the Sherrington-Kirkpatrick model, cf. Eq.* (3.1.1)*:*

1. $\lim_{s\to 0} \operatorname{p-lim}_{N\to\infty} W_{2,N}(\mu_{\boldsymbol{A},\beta}, \mu_{\boldsymbol{A}_s,\beta}) = 0$.
2. $\lim_{\beta'\to\beta} \operatorname{p-lim}_{N\to\infty} W_{2,N}(\mu_{\boldsymbol{A},\beta}, \mu_{\boldsymbol{A},\beta'}) = 0$.

Proof. Take $\varepsilon > 0$ arbitrarily small and choose parameters $(\eta, K_{\mathsf{AMP}}, K_{\mathsf{NGD}}, L, \delta)$ of Algorithm 2 with the desired tolerance ε so that Theorem 37 holds. Combining with Theorem 10 using the same parameters $(\eta, K_{\mathsf{AMP}}, K_{\mathsf{NGD}}, L, \delta)$ implies the result since ε is arbitrarily small. (Recall that $(\eta, K_{\mathsf{AMP}}, K_{\mathsf{NGD}}, L, \delta)$ can be chosen independent of β for $\beta \le \beta_0 < 1/2$.) □

Remark 3.2.2. We emphasize that Corollary 3.2.2 makes no reference to the sampling algorithm, and is instead a purely structural property of the Gibbs measure. The sampling algorithm, however, is the key tool of our proof.

Stability is related to chaos, which is a well studied and important property of spin glasses, see e.g. [Cha09, Che13a, Cha14, CHHS15, CP18]. In particular, "disorder chaos" refers to the following

phenomenon. Draw $\boldsymbol{x}^0 \sim \mu_{\boldsymbol{A},\beta}$ independently of $\boldsymbol{x}^s \sim \mu_{\boldsymbol{A}_s,\beta}$, and denote by $\mu_{\boldsymbol{A},\beta}^{(0,s)} := \mu_{\boldsymbol{A},\beta} \otimes \mu_{\boldsymbol{A}^s,\beta}$ their joint distribution. Disorder chaos holds at inverse temperature β if

$$\lim_{s\to 0}\lim_{N\to\infty} \mathbb{E}\,\mu_{\boldsymbol{A},\beta}^{(0,s)}\Big\{\Big(\frac{1}{N}\langle \boldsymbol{x}^0, \boldsymbol{x}^s\rangle\Big)^2\Big\} = 0\,. \tag{3.2.11}$$

Note that disorder chaos is not necessarily a surprising property. For instance when $\beta = 0$, the distribution $\mu_{\boldsymbol{A}_s,\beta}$ is simply the uniform measure over the hypercube $\{-1,+1\}^N$ for all s, and this example exhibits disorder chaos in the sense of Eq. (3.2.11). In fact, the SK Gibbs measure exhibits disorder chaos at all $\beta \in [0,\infty)$ [Cha09]. However, for $\beta > 1$, Eq. (3.2.11) leads to a stronger conclusion.

Theorem 11 (Disorder chaos in $W_{2,N}$ distance)**.** *For all $\beta > 1$,*

$$\inf_{s\in(0,1)} \liminf_{N\to\infty} \mathbb{E}\left[W_{2,N}(\mu_{\boldsymbol{A},\beta}, \mu_{\boldsymbol{A}_s,\beta})\right] > 0\,.$$

Finally, we obtain the desired hardness result by reversing the implication in Corollary 3.2.2: no stable algorithm which can approximately sample from the measure $\mu_{\boldsymbol{A},\beta}$ in the $W_{2,N}$ sense for $\beta > 1$.

Theorem 12. *Fix $\beta > 1$, and let $\{\mathsf{ALG}_N\}_{N\geq 1}$ be a family of randomized algorithms which is stable with respect to disorder as per Definition 3.2.1 at inverse temperature β. Let $\mu_{\boldsymbol{A},\beta}^{\mathrm{alg}}$ be the law of the output $\mathsf{ALG}_N(\boldsymbol{A},\beta,\omega)$ conditional on $\boldsymbol{A}$. Then*

$$\liminf_{N\to\infty} \mathbb{E}\left[W_{2,N}(\mu_{\boldsymbol{A},\beta}^{\mathrm{alg}},\ \mu_{\boldsymbol{A},\beta})\right] > 0\,.$$

We refer the reader to Section 3.9.2 for the proof of this theorem.

3.2.3 Notations

We use $o_N(1)$ to indicate a quantity tending to 0 as $N \to \infty$. We use $o_{N,\mathbb{P}}(1)$ for a quantity tending to 0 in probability. If X is a random variable, then $\mathcal{L}(X)$ indicates its law. The quantity $C(\beta)$ refers to a constant depending on β. For $\boldsymbol{x} \in \mathbb{R}^N$ and $\rho \in \mathbb{R}_{\geq 0}$, we denote the open ball of center $\boldsymbol{x}$ and radius ρ by $B(\boldsymbol{x},\rho) := \{\boldsymbol{y} \in \mathbb{R}^N : \|\boldsymbol{y} - \boldsymbol{x}\|_2 < \rho\}$. The uniform distribution on the interval $[a,b]$ is denoted by $\mathsf{Unif}([a,b])$. The set of probability distributions over a measurable space $(\Omega,\mathcal{F})$ is denoted by $\mathscr{P}(\Omega)$.

3.3 Properties of Stochastic Localization

We collect in this section the main properties of the stochastic localization process needed for our analysis. To be definite, we will focus on the stochastic localization process for the Gibbs measure (3.1.1), although most of what we will say generalizes to other probability measures in $\mathbb{R}^N$, under suitable tail conditions. Throughout this section, the matrix $\boldsymbol{A}$ is viewed as fixed.

Recalling the tilted measure $\mu_{\boldsymbol{A},\boldsymbol{y}}$ of Eq. (3.1.2), and the SDE of Eq. (3.1.4), we introduce the shorthand

$$\mu_t = \mu_{\boldsymbol{A},\boldsymbol{y}(t)}\,.$$

The following properties are well known. See for instance [ES22, Propositions 9, 10] or [Eld20]. We provide proofs for the reader's convenience.

Lemma 3.3.1. *For all $t \geq 0$ and all $\boldsymbol{x} \in \{-1,+1\}^N$,*

$$d\mu_t(\boldsymbol{x}) = \mu_t(\boldsymbol{x})\langle \boldsymbol{x} - \mathbf{m}_{\boldsymbol{A},\boldsymbol{y}(t)}, d\boldsymbol{B}(t)\rangle\,. \tag{3.3.1}$$

As a consequence, for any function $\varphi : \mathbb{R}^N \to \mathbb{R}^m$, the process $\big(\mathbb{E}_{\boldsymbol{x}\sim\mu_t}\big[\varphi(\boldsymbol{x})\big]\big)_{t\geq 0}$ is a martingale.

Proof. Let us evaluate the differential of $\log \mu_t$. By writing Z_t for the normalization constant $Z(\boldsymbol{y}(t))$ of Eq. (3.1.2), we get

$$\mathrm{d}\log \mu_t(\boldsymbol{x}) = \langle \mathrm{d}\boldsymbol{y}(t), \boldsymbol{x}\rangle - \mathrm{d}\log Z_t\,. \tag{3.3.2}$$

Using Itô's formula for Z_t we have

$$\begin{aligned}
\mathrm{d}Z_t &= \mathrm{d}\sum_{\boldsymbol{x}\in\{-1,+1\}^N} e^{(\beta/2)\langle \boldsymbol{x},\boldsymbol{A}\boldsymbol{x}\rangle + \langle \boldsymbol{y}(t),\boldsymbol{x}\rangle} \\
&= \sum_{\boldsymbol{x}\in\{-1,+1\}^N} \big(\langle \mathrm{d}\boldsymbol{y}(t),\boldsymbol{x}\rangle + \frac{1}{2}\|\boldsymbol{x}\|^2\mathrm{d}t\big) e^{(\beta/2)\langle \boldsymbol{x},\boldsymbol{A}\boldsymbol{x}\rangle + \langle \boldsymbol{y}(t),\boldsymbol{x}\rangle}\,.
\end{aligned}$$

Therefore, denoting by $[Z]_t$ the quadratic variation process associated to Z_t,

$$\begin{aligned}
\mathrm{d}\log Z_t &= \frac{\mathrm{d}Z_t}{Z_t} - \frac{1}{2}\frac{\mathrm{d}[Z]_t}{Z_t^2} \\
&= \langle \mathrm{d}\boldsymbol{y}(t), \mathbf{m}_{\boldsymbol{A},\boldsymbol{y}(t)}\rangle + \frac{1}{2}\mathop{\mathbb{E}}_{\mu_t}[\|\boldsymbol{x}\|^2]\mathrm{d}t - \frac{1}{2}\|\mathbf{m}_{\boldsymbol{A},\boldsymbol{y}(t)}\|^2\mathrm{d}t \\
&= \langle \mathrm{d}\boldsymbol{y}(t), \mathbf{m}_{\boldsymbol{A},\boldsymbol{y}(t)}\rangle + \frac{N}{2} - \frac{1}{2}\|\mathbf{m}_{\boldsymbol{A},\boldsymbol{y}(t)}\|^2\mathrm{d}t\,.
\end{aligned}$$

Substituting in (3.3.2) we obtain

$$\begin{aligned}\mathrm{d}\log\mu_t(\boldsymbol{x}) &= \langle \mathrm{d}\boldsymbol{y}(t), \boldsymbol{x}-\mathbf{m}_{\boldsymbol{A},\boldsymbol{y}(t)}\rangle - \frac{N}{2}\mathrm{d}t + \frac{1}{2}\|\mathbf{m}_{\boldsymbol{A},\boldsymbol{y}(t)}\|^2\mathrm{d}t\\ &= \langle \mathrm{d}\boldsymbol{B}_t, \boldsymbol{x}-\mathbf{m}_{\boldsymbol{A},\boldsymbol{y}(t)}\rangle - \frac{1}{2}\|\boldsymbol{x}-\mathbf{m}_{\boldsymbol{A},\boldsymbol{y}(t)}\|^2\mathrm{d}t\,.\end{aligned}$$

Applying Itô's formula to $e^{\log\mu_t(\boldsymbol{x})}$ yields the desired result.

Finally, Eq. (3.3.1) implies that $\mu_t(\boldsymbol{x})$ is a martingale for every $\boldsymbol{x}\in\{-1,+1\}^N$. Since $\mathbb{E}_{\boldsymbol{x}\sim\mu_t}\left[\varphi(\boldsymbol{x})\right]$ is a linear combination of martingales, it is itself a martingale. □

Lemma 3.3.2 ([Eld20])**.** *For all $t>0$,*

$$\mathbb{E}\,\mathrm{Cov}(\mu_t) \preceq \frac{1}{t}\boldsymbol{I}_N\,. \tag{3.3.3}$$

Lemma 3.3.3. *For all $t>0$,*

$$W_{2,N}\big(\mu_{\boldsymbol{A}}, \mathcal{L}(\mathbf{m}_{\boldsymbol{A},\boldsymbol{y}(t)})\big)^2 \le \frac{1}{t}\,. \tag{3.3.4}$$

In particular, the mean vector $\mathbf{m}_{\boldsymbol{A},\boldsymbol{y}(t)}$ converges in distribution to a random vector $\boldsymbol{x}^*\sim\mu_{\boldsymbol{A}}$ as $t\to\infty$.

Proof. By Lemma 3.3.2,

$$\mathbb{E}\big[\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mu_t}[\|\boldsymbol{x}-\mathbf{m}_{\boldsymbol{A},\boldsymbol{y}(t)}\|^2]\big] \le \frac{N}{t}\,,$$

therefore

$$\mathbb{E}\left[W_{2,N}\big(\mu_t,\delta_{\mathbf{m}_{\boldsymbol{A},\boldsymbol{y}(t)}}\big)^2\right] \le \frac{1}{t}\,.$$

Notice that $(\mu,\nu)\mapsto W_{2,N}^2(\mu,\nu)$ is jointly convex. Since $\mu_{\boldsymbol{A}}=\mathbb{E}[\mu_t]$, this implies

$$W_{2,N}\big(\mu_{\boldsymbol{A}}, \mathcal{L}(\mathbf{m}_{\boldsymbol{A},\boldsymbol{y}(t)})\big)^2 \le \mathbb{E}\left[W_{2,N}\big(\mu_t,\delta_{\mathbf{m}_{\boldsymbol{A},\boldsymbol{y}(t)}}\big)^2\right] \le \frac{1}{t}\,.$$

□

3.4 Analysis of Algorithm 2 and proof of Theorem 37

This section is devoted to the analysis of Algorithm 2 described in the previous section. An important simplification is obtained by reducing ourselves to working with a corresponding *planted* model. This approach has two advantages: (*i*) The joint distribution of the matrix $\boldsymbol{A}$ and the process $(\boldsymbol{y}(t))_{t\ge0}$ in (3.1.4) is significantly simpler in the planted model; (*ii*) Analysis in the planted model can be cast as a statistical estimation problem. In the latter, Bayes-optimality considerations can be exploited to relate the output of the AMP algorithm $\mathsf{AMP}(\boldsymbol{A},\boldsymbol{y};k)$ to the true mean vector $\mathbf{m}(\boldsymbol{A},\boldsymbol{y})$.

This section is organized as follows. Section 3.5 introduces the planted model and its relation to the original model. We then analyze the AMP component of our algorithm in Section 3.6, and the NGD component in Section 3.7. Finally, Section 3.8 puts the various elements together and proves Theorem 37.

3.5 The planted model and contiguity

Let $\overline{\nu}$ be the uniform distribution over $\{-1,+1\}^N$ and consider the joint distribution of pairs $(\boldsymbol{x},\boldsymbol{A}) \in \{-1,+1\}^N \times \mathbb{R}^{N\times N}_{\rm sym}$,

$$\mu_{\rm pl}(\mathrm{d}\boldsymbol{x},\mathrm{d}\boldsymbol{A}) = \frac{1}{Z_{\rm pl}}\,\exp\Big\{-\frac{N}{4}\Big\|\boldsymbol{A}-\frac{\beta\boldsymbol{x}\boldsymbol{x}^{\top}}{N}\Big\|_F^2\Big\}\,\overline{\nu}(\mathrm{d}\boldsymbol{x})\,\mathrm{d}\boldsymbol{A}\,, \tag{3.5.1}$$

where $\mathrm{d}\boldsymbol{A}$ is the Lebesgue measure over the space of symmetric matrices $\mathbb{R}^{N\times N}_{\rm sym}$, and the normalizing constant

$$Z_{\rm pl} := \int \exp\Big\{-\frac{N}{4}\Big\|\boldsymbol{A}-\frac{\beta\boldsymbol{x}\boldsymbol{x}^{\top}}{N}\Big\|_F^2\Big\}\,\mathrm{d}\boldsymbol{A} \tag{3.5.2}$$

is independent of $\boldsymbol{x}\in\{-1,+1\}^N$. It is easy to see by construction that the marginal distribution of $\boldsymbol{x}$ under $\mu_{\rm pl}$ is $\overline{\nu}$, and the conditional law $\mu_{\rm pl}(\,\cdot\,|\boldsymbol{x})$ is a rank-one spiked GOE model with spike $\beta\boldsymbol{x}\boldsymbol{x}^{\top}/N$. Namely, under $\mu_{\rm pl}(\,\cdot\,|\boldsymbol{x})$, we have

$$\boldsymbol{A} = \frac{\beta}{N}\boldsymbol{x}\boldsymbol{x}^{\top}+\boldsymbol{W}\,,\quad \boldsymbol{W}\sim\mathsf{GOE}(N)\,. \tag{3.5.3}$$

On the other hand, $\mu_{\rm pl}(\,\cdot\,|\boldsymbol{A})$ is the SK measure $\mu_{\boldsymbol{A}}$.

The marginal of $\boldsymbol{A}$ under $\mu_{\rm pl}$ is not the $\mathsf{GOE}(N)$ distribution μ_{GOE} but takes the form

$$\mu_{\rm pl}(\mathrm{d}\boldsymbol{A}) = \frac{1}{Z_{\rm pl}}\,e^{-\frac{N}{4}\|\boldsymbol{A}\|_F^2}\,Z_{\mathsf{SK}}(\boldsymbol{A})\,\mathrm{d}\boldsymbol{A} \tag{3.5.4}$$

$$= \mu_{\mathsf{GOE}}(\mathrm{d}\boldsymbol{A})\,Z_{\mathsf{SK}}(\boldsymbol{A})\,, \tag{3.5.5}$$

where $Z_{\mathsf{SK}}(\boldsymbol{A})$ is the (rescaled) partition function of the SK measure

$$Z_{\mathsf{SK}}(\boldsymbol{A}) = 2^{-n}\sum_{\boldsymbol{x}\in\{-1,+1\}^N}\exp\Big\{\frac{\beta}{2}\langle\boldsymbol{x},\boldsymbol{A}\boldsymbol{x}\rangle-\frac{\beta^2 N}{4}\Big\}\,. \tag{3.5.6}$$

By a classical result [ALR87], $Z_{\mathsf{SK}}(\boldsymbol{A})$ has log-normal fluctuations for all $\beta<1$:

Theorem 13 ([ALR87]). *Let $\beta < 1$, $\boldsymbol{A} \sim \mu_{\mathsf{GOE}}$ and $\sigma^2 = \frac{1}{4}(-\log(1-\beta^2) - \beta^2)$. Then*

$$Z_{\mathsf{SK}}(\boldsymbol{A}) \xrightarrow[n\to\infty]{d} \exp(W)\,, \tag{3.5.7}$$

where $W \sim \mathcal{N}\big(-\sigma^2, 2\sigma^2\big)$.

Therefore, by Le Cam's first lemma [VdV98, Lemma 6.4], $\mu_{\mathrm{pl}}(\mathrm{d}\boldsymbol{A})$ and $\mu_{\mathsf{GOE}}(\mathrm{d}\boldsymbol{A})$ are mutually contiguous for all $\beta < 1$. For the purpose of our analysis we will need a stronger result about the joint distributions of $(\boldsymbol{A}, \boldsymbol{y})$ under our "random" model and a planted model which we now introduce.

Recall that $\mathbf{m}(\boldsymbol{A}, \boldsymbol{y})$ denotes the mean of the Gibbs measure $\mu_{\boldsymbol{A},\boldsymbol{y}}$ in Eq. (3.1.2). For a fixed $T \geq 0$, we define two Borel distributions $\mathbb{P}$ and $\mathbb{Q}$ on $(\boldsymbol{A}, \boldsymbol{y}) \in \mathbb{R}^{N\times N}_{\mathrm{sym}} \times C([0,T], \mathbb{R}^N)$ as follows:

$$\mathbb{Q} \;:\; \begin{cases} \boldsymbol{A} & \sim & \mu_{\mathsf{GOE}}\,, \\ \boldsymbol{y}(t) & = & \displaystyle\int_0^t \mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(s))\,\mathrm{d}s + \boldsymbol{B}(t)\,, \quad t \in [0,T]\,, \end{cases} \qquad \text{(random)} \tag{3.5.8}$$

$$\mathbb{P} \;:\; \begin{cases} \boldsymbol{x}_0 & \sim & \overline{\nu}\,, \\ \boldsymbol{A} & \sim & \mu_{\mathrm{pl}}(\,\cdot\,|\,\boldsymbol{x}_0)\,, \\ \boldsymbol{y}(t) & = & t\boldsymbol{x}_0 + \boldsymbol{B}(t)\,, \quad t \in [0,T] \end{cases} \qquad \text{(planted)} \tag{3.5.9}$$

where $(\boldsymbol{B}(t))_{t\geq 0}$ is a standard Brownian motion in $\mathbb{R}^N$ independent of everything else. Note the SDE defining the process $\boldsymbol{y} = (\boldsymbol{y}(t))_{t\in[0,T]}$ in Eq. (3.5.8) is a restatement of the stochastic localization equation (3.1.4) applied to the SK measure $\mu_{\boldsymbol{A}}$.

Proposition 3.5.1. *For all $T \geq 0$ and $\beta \geq 0$, $\mathbb{P}$ absolutely continuous with respect to $\mathbb{Q}$ and for all $(\boldsymbol{A}, \boldsymbol{y}) \in \mathbb{R}^{N\times N}_{\mathrm{sym}} \times C([0,T], \mathbb{R}^N)$,*

$$\frac{d\mathbb{P}}{d\mathbb{Q}}(\boldsymbol{A}, \boldsymbol{y}) = Z_{\mathsf{SK}}(\boldsymbol{A})\,.$$

Therefore, for all $\beta < 1$, $\mathbb{P}$ and $\mathbb{Q}$ are mutually contiguous. (Namely, for a sequence of events $\mathcal{E}_N$, $\lim_{N\to\infty} \mathbb{P}(\mathcal{E}_N) = 0$ if and only if $\lim_{N\to\infty} \mathbb{Q}(\mathcal{E}_N) = 0$.)

Proof. Fix $\boldsymbol{x}_0 \in \mathbb{R}^N$. We first calculate the density of the process $\boldsymbol{y}(t) = t\boldsymbol{x}_0 + \boldsymbol{B}(t)$ with respect to Brownian motion. Let $\boldsymbol{W}$ be the Wiener measure on $C([0,T], \mathbb{R}^N)$. We obtain by Girsanov's theorem that

$$\frac{\mathrm{d}\mathbb{P}(\,\cdot\,|\boldsymbol{x}_0)}{\mathrm{d}\boldsymbol{W}}(\boldsymbol{y}) = e^{\langle \boldsymbol{x}_0, \boldsymbol{y}(T)\rangle - T\|\boldsymbol{x}_0\|^2/2}\,. \tag{3.5.10}$$

Notice that the above density only depends on the endpoint $\boldsymbol{y}(T)$ of the process $\boldsymbol{y}$. From this, we

obtain an explicit formula for the density of $\mathbb{P}$ with respect to $(\mathrm{d}\boldsymbol{A}) \times \boldsymbol{W}$:

$$\mathbb{P}(\mathrm{d}\boldsymbol{A}, \mathrm{d}\boldsymbol{y}) = \frac{1}{Z_{\mathrm{pl}}}\Big(\int \exp\Big\{ -\frac{N}{4}\Big\|\boldsymbol{A} - \frac{\beta \boldsymbol{x}_0 \boldsymbol{x}_0^\top}{N}\Big\|_F^2 + \langle \boldsymbol{x}_0, \boldsymbol{y}(T)\rangle - \frac{T}{2}\|\boldsymbol{x}_0\|^2\Big\} \overline{\nu}(\mathrm{d}\boldsymbol{x}_0)\Big)\, \mathrm{d}\boldsymbol{A}\, \boldsymbol{W}(\mathrm{d}\boldsymbol{y})\,, \tag{3.5.11}$$

where $Z_{\mathrm{pl}} = \int e^{-n\|\boldsymbol{A}\|_F^2/4}\mathrm{d}\boldsymbol{A}$ is given in Eq. (3.5.2).

Next we derive a similar formula for $\mathbb{Q}$. Fix a matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}_{\mathrm{sym}}$ and let $\boldsymbol{y}$ be the solution to the SDE in (3.5.8). Let $(\bar{\boldsymbol{B}}(t))_{t\ge 0}$ be another standard Brownian motion in $\mathbb{R}^N$, and consider the process $\bar{\boldsymbol{y}} = (\bar{\boldsymbol{y}}(t))_{t\in[0,T]}$ defined by

$$\bar{\boldsymbol{y}}(t) = t\boldsymbol{x} + \bar{\boldsymbol{B}}(t) \quad \text{ where } \ \boldsymbol{x} \sim \mu_{\boldsymbol{A}} \text{ independently of } \bar{\boldsymbol{B}}\,. \tag{3.5.12}$$

Then, there exists another Brownian motion $(\boldsymbol{W}(t))_{t\ge 0}$ adapted to the filtration $(\boldsymbol{F}_t = \sigma(\bar{\boldsymbol{y}}(s) : s \le t))_{t\in[0,T]}$ such that $\mathrm{d}\bar{\boldsymbol{y}}(t) = \mathbf{m}_{\boldsymbol{A},\bar{\boldsymbol{y}}(t)}\mathrm{d}t + \mathrm{d}\boldsymbol{W}(t)$ for all $t \in [0,T]$. This is stated as Theorem 7.12 of [LS77], and can be proved directly applying Levy's characterization of Brownian motion to the process $\bar{\boldsymbol{y}}(t) - \int_0^t \mathbf{m}(\boldsymbol{A}, \bar{\boldsymbol{y}}(s))\mathrm{d}s$.

Therefore, the processes $\bar{\boldsymbol{y}}$ and $\boldsymbol{y}$ share the same law conditional on $\boldsymbol{A}$. Since we computed the law of $\bar{\boldsymbol{y}}$ in (3.5.10), we obtain

$$\mathbb{Q}(\mathrm{d}\boldsymbol{A}, \mathrm{d}\boldsymbol{y}) = \frac{1}{Z_{\mathrm{pl}}}\Big(\int \exp\Big\{ -\frac{N}{4}\|\boldsymbol{A}\|_F^2 + \langle \boldsymbol{x}, \boldsymbol{y}(T)\rangle - \frac{T}{2}\|\boldsymbol{x}\|^2\Big\} \mu_{\boldsymbol{A}}(\mathrm{d}\boldsymbol{x})\Big)\, \mathrm{d}\boldsymbol{A}\, \boldsymbol{W}(\mathrm{d}\boldsymbol{y})\,, \tag{3.5.13}$$

where Z_{pl} is as above. Since $\mu_{\boldsymbol{A}}(\mathrm{d}\boldsymbol{x}) = Z_{\mathsf{SK}}(\boldsymbol{A})^{-1} e^{\beta\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle/2 - \beta^2 N/4}\overline{\nu}(\mathrm{d}\boldsymbol{x})$, we obtain after simplification

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\,\mathbb{Q}}(\boldsymbol{A}, \boldsymbol{y}) = Z_{\mathsf{SK}}(\boldsymbol{A})\,. \tag{3.5.14}$$

Mutual contiguity follows from Theorem 13 and Le Cam's first lemma. □

Therefore, for the remainder of the proof of Theorem 37, we work under the "planted" distribution $\mathbb{P}$. All results proven under $\mathbb{P}$ transfer to $\mathbb{Q}$ by contiguity.

3.6 Approximate Message Passing

In this section we analyze the AMP iteration of Algorithm 1, which we copy here for the reader's convenience

$$\begin{aligned}
&\hat{\boldsymbol{m}}^{-1} = \boldsymbol{z}^0 = 0,\\
&\hat{\boldsymbol{m}}^k = \tanh(\boldsymbol{z}^k), \qquad \mathsf{b}_k = \frac{\beta^2}{N}\sum_{i=1}^N \big(1-\tanh^2(z_i^k)\big) \qquad \forall k \ge 0\,, \\
&\boldsymbol{z}^{k+1} = \beta \boldsymbol{A}\hat{\boldsymbol{m}}^k + \boldsymbol{y} - \mathsf{b}_k \hat{\boldsymbol{m}}^{k-1}\,.
\end{aligned} \tag{3.6.1}$$

When needed, we will specify the dependence on $\boldsymbol{A}, \boldsymbol{y}$ by writing $\hat{\boldsymbol{m}}^k = \hat{\boldsymbol{m}}^k(\boldsymbol{A},\boldsymbol{y}) = \mathsf{AMP}(\boldsymbol{A},\boldsymbol{y};k)$ and $\boldsymbol{z}^k = \boldsymbol{z}^k(\boldsymbol{A},\boldsymbol{y})$. Throughout this section $(\boldsymbol{A},\boldsymbol{y}) \sim \mathbb{P}$ will be distributed according to the planted model introduced above.

Our analysis will be based on the general state evolution result of [BM11a, JM13], which implies the following asymptotic characterization for the iterates. Set $\gamma_0(\beta,t) = 0, \Sigma_{0,i}(\beta,t) = 0$ and recursively define

$$\gamma_{k+1}(\beta,t) = \beta^2 \cdot \mathbb{E}\left[\tanh\left(\gamma_k(\beta,t) + t + G_k\right)\right], \tag{3.6.2}$$

$$\Sigma_{k+1,j+1}(\beta,t) = \beta^2 \cdot \mathbb{E}\left[\tanh\left(\gamma_k(\beta,t) + t + G_k\right)\tanh\left(\gamma_j(\beta,t) + t + G_j\right)\right], \tag{3.6.3}$$

where $(\boldsymbol{G}_j)_{j\le k}$ are jointly Gaussian, with zero mean and covariance $\boldsymbol{\Sigma}_{\le k} + t\mathbf{1}\mathbf{1}^\top$, $\boldsymbol{\Sigma}_{\le k} := (\Sigma_{ij})_{i,j\le k}$.

Proposition 3.6.1 (Theorem 1 of [BM11a]). *For $(\boldsymbol{A},\boldsymbol{y}) \sim \mathbb{P}$ and any $k \in \mathbb{Z}_{\ge 0}$, the empirical distribution of the coordinate of the AMP iterates converges almost surely in $W_2(\mathbb{R}^{k+2})$ as follows:*

$$\frac{1}{N}\sum_{i=1}^N \delta_{(z_i^1,\cdots,z_i^k,x_i,y_i)} \xrightarrow[n\to\infty]{W_2} \mathcal{L}\left(\boldsymbol{\gamma}_{\le k}(\beta,t)X + \boldsymbol{G} + Y\mathbf{1}, X, Y\right), \tag{3.6.4}$$

$$\boldsymbol{\gamma}_{\le k}(\beta,t) = \big(\gamma_1(\beta,t),\dots,\gamma_k(\beta,t)\big), \qquad \boldsymbol{G} \sim \mathcal{N}(0,\boldsymbol{\Sigma}_{\le k}). \tag{3.6.5}$$

On the right-hand side, X is uniformly random in $\{-1,+1\}$, $Y = tX + \sqrt{t}W$ where $W \sim \mathcal{N}(0,1)$ and $X, \boldsymbol{G}, W$ are mutually independent.

Remark 3.6.1. This specific statement follows from [BM11a, Theorem 1] by a change of variables, as in [DAM17] or [MV21].

As in [DAM17, Eqs. (69,70)] we argue that the state evolution equations (3.6.2), (3.6.3) take a simple form thanks to our specific choice of AMP non-linearity $\tanh(\cdot)$. It will be convenient to use

the notations

$$\widetilde{\gamma}_k(\beta,t) = \gamma_k(\beta,t) + t\,,$$
$$\widetilde{\Sigma}_{k,j}(\beta,t) = \Sigma_{k,j}(\beta,t) + t\,.$$

Proposition 3.6.2. *For any $t \in \mathbb{R}_{\geq 0}$ and $k, j \in \mathbb{Z}_{\geq 0}$,*

$$\Sigma_{k,j}(\beta,t) = \gamma_{k\wedge j}(\beta,t)\,, \quad \textit{and} \quad \widetilde{\Sigma}_{k,j}(\beta,t) = \widetilde{\gamma}_{k\wedge j}(\beta,t)\,.$$

Proof. The two claims are equivalent and we proceed by induction. The base case $k = 0$ holds by definition, so we may assume $\Sigma_{i,j}(\beta,t) = \gamma_{i\wedge j}(\beta,t)$ for $i,j \leq k-1$. Set $Z_j = \gamma_j X + \widetilde{G}_j$ where $\widetilde{\boldsymbol{G}} \sim \mathcal{N}(0, \widetilde{\boldsymbol{\Sigma}}_{\leq k-1})$. Note that, by the induction hypothesis, Z_{k-1} is a sufficient statistic for X given $(Z_j)_{j\leq k-1}$. Using Bayes' rule, and writing $\widetilde{\sigma}^2_{k-1} := \widetilde{\Sigma}_{k-1,k-1}$, one easily computes

$$\mathbb{E}[X|Z_{k-1}] = \frac{e^{\widetilde{\gamma}_{k-1}Z_{k-1}/\widetilde{\sigma}^2_{k-1}} - e^{-\widetilde{\gamma}_{k-1}Z_{k-1}/\widetilde{\sigma}^2_{k-1}}}{e^{\widetilde{\gamma}_{k-1}Z/\widetilde{\sigma}^2_{k-1}} + e^{-\widetilde{\gamma}_{k-1}Z/\widetilde{\sigma}^2_{k-1}}} = \tanh(Z)\,.$$

Therefore using Eq. (3.6.2), the fact that tanh is an odd function and $WX \stackrel{\mathrm{d}}{=} W$,

$$\begin{aligned}
\widetilde{\Sigma}_{k,j} &= \mathbb{E}\left[\mathbb{E}[X|Z_{k-1}]\,\mathbb{E}[X|Z_{j-1}]\right] \\
&\stackrel{(a)}{=} \mathbb{E}\left[X\,\mathbb{E}[X|Z_{j-1}]\right] \\
&= \mathbb{E}\left[X\tanh(\widetilde{\gamma}_{j-1}X + \widetilde{\sigma}^2_{j-1}W)\right] \\
&= \mathbb{E}\left[\tanh(\widetilde{\gamma}_{j-1} + \widetilde{\sigma}^2_{j-1}W)\right] = \gamma_j\,,
\end{aligned}$$

where in step (a) we crucially used the sufficient statistic property. This completes the inductive step and hence the proof. □

Define the function $\mathsf{mmse} : \mathbb{R} \to \mathbb{R}$ given by

$$\mathsf{mmse}(\gamma) \equiv 1 - \mathbb{E}\left[\tanh(\gamma + \sqrt{\gamma}W)^2\right] = 1 - \mathbb{E}\left[\mathbb{E}[X|\gamma X + \sqrt{\gamma}W]^2\right].$$

It follows from Proposition 3.6.2 that (3.6.2) and (3.6.3) can be expressed just in terms of the sequence $\gamma_k(\beta,t)$ defined by $\gamma_0(t) = 0$ and the recursion

$$\gamma_{k+1}(\beta,t) = \beta^2\big(1 - \mathsf{mmse}(\gamma_k(\beta,t) + t)\big). \tag{3.6.6}$$

Note that $\gamma_k(\beta,t)$ depends also on β, which is usually treated as constant. The following result details some useful properties of mmse.

Lemma 3.6.3 (Lemma 6.1 of [DAM17])**.** *The following properties hold, where $\{\gamma_k(\beta,t)\}_{k\geq 1}$ is as*

defined by (3.6.6).

(a) mmse *is differentiable, strictly decreasing, and convex in* $\gamma \in \mathbb{R}_{\geq 0}$.

(b) $\mathsf{mmse}(0) = 1$, $\mathsf{mmse}'(0) = -1$ *and* $\lim_{\gamma\to\infty} \mathsf{mmse}(\gamma) = 0$.

(c) For $t \geq 0$ *there exists a non-negative solution* $\gamma_* = \gamma_*(\beta, t)$ *to the fixed point equation*

$$\gamma_* = \beta^2(1 - \mathsf{mmse}(\gamma_* + t))\,. \tag{3.6.7}$$

The solution to this equation is unique for all $t > 0$.

(d) The function $(\beta, t) \mapsto \gamma_*(\beta, t)$ *is differentiable for* $t > 0$.

(e) For all $\beta < 1$ *and* $t > 0$,

$$1 - \beta^{2k} \leq \frac{\gamma_k(\beta, t)}{\gamma_*(\beta, t)} \leq 1\,. \tag{3.6.8}$$

(f) For $\beta < 1$ *and* $\mathsf{T} > 0$, *there exist constants* $c(\beta, \mathsf{T}), C(\beta, \mathsf{T}) \in (0, \infty)$ *such that, for all* $t \in (0, \mathsf{T}]$,

$$c(\beta, \mathsf{T}) \leq \frac{\gamma_*(\beta, t)}{t} \leq C(\beta, \mathsf{T})\,. \tag{3.6.9}$$

(g) For $\beta < 1$ *and any* $t_1, t_2 \in (0, \infty)$,

$$\gamma_*(\beta, t_1) - \gamma_*(\beta, t_2) \leq \frac{\beta^2}{1 - \beta^2}|t_1 - t_2|. \tag{3.6.10}$$

Proof. Lemma 6.1 in [DAM17] proves that $\gamma \mapsto \mathsf{mmse}(\gamma)$ is differentiable, strictly decreasing, and convex in $\gamma \in \mathbb{R}_{\geq 0}$ (Note that the statement of that Lemma does not claim differentiability, but this is actually proved there by a simple application of dominated convergence). This proves point (a).

Point (b) follows by a direct calculation, cf. [DAM17]. Indeed, by Stein's lemma (Gaussian integration by parts), with $Z = \gamma + W\sqrt{\gamma}$,

$$\begin{aligned} -\mathsf{mmse}'(\gamma) &= \frac{\mathrm{d}}{\mathrm{d}\gamma}\,\mathbb{E}[\tanh(\gamma + W\sqrt{\gamma})^2] \\ &= \mathbb{E}[2\tanh(Z)\tanh'(Z) + \tanh'(Z)^2 + \tanh(Z)\tanh''(Z)] \end{aligned}$$

Evaluating at $\gamma = 0$ shows

$$\mathsf{mmse}'(0) = -1.$$

Also, dominated convergence yields the desired limit values.

Point (c), namely existence and uniqueness of solutions of Eq. (3.6.7) follows from the above monotonicity and convexity properties. Point (d) follows from the implicit function theorem.

We are left with the task of proving (3.6.8), (3.6.9) and (3.6.10), which are not given in [DAM17].

Define

$$f_t(\gamma) \equiv \beta^2(1 - \mathsf{mmse}(\gamma + t))$$

so that $f_t(\gamma_k(\beta, t)) = \gamma_{k+1}(\beta, t)$. By point (b), $f_t(0) \geq 0$. By point (a), $f_t(\cdot)$ is increasing and concave. Combined with the computation above, we conclude that $f_t'(\gamma) \in [0, \beta^2]$ for all $\gamma \geq 0$. By the mean value theorem, it follows that for $\gamma < \gamma_*$,

$$0 \leq \gamma_*(\beta, t) - f_t(\gamma) = f(\gamma_*(\beta, t)) - f_t(\gamma) \leq \beta^2(\gamma_*(\beta, t) - \gamma)\,. \tag{3.6.11}$$

Setting $\gamma = \gamma_j(\beta, t)$, we obtain

$$0 \leq \frac{\gamma_*(\beta, t) - \gamma_{j+1}(\beta, t)}{\gamma_*(\beta, t) - \gamma_j(\beta, t)} \leq \beta^2.$$

Multiplying for $j \in \{0, \dots, k-1\}$, we find

$$0 \leq \frac{\gamma_*(\beta, t) - \gamma_k(\beta, t)}{\gamma_*(\beta, t)} \leq \beta^{2k}\,,$$

or,

$$\frac{\gamma_k(\beta, t)}{\gamma_*(\beta, t)} \in \left[1 - \beta^{2k}, 1\right]\,,$$

which proves (3.6.8).

To prove (3.6.9), note that we just showed

$$\frac{\gamma_1(\beta, t)}{\gamma_*(\beta, t)} \in \left[1 - \beta^2, 1\right].$$

Therefore it suffices to show that

$$c(\beta, \mathsf{T}) \leq \frac{\gamma_1(\beta, t)}{t} \leq C(\beta, \mathsf{T}), \quad t \in (0, \mathsf{T}]\,. \tag{3.6.12}$$

By definition, $\gamma_1(\beta, t) = \beta^2(1 - \mathsf{mmse}(t))$. Thus (3.6.12) follows from the fact that $\mathsf{mmse}(0) = 1$, $\mathsf{mmse}'(0) = -1$, and $\mathsf{mmse} : \mathbb{R}_{\geq 0} \to [0, 1]$ is convex and strictly decreasing. In fact we have $\gamma_1(\beta, \mathsf{T})/\mathsf{T} \leq \gamma_1(\beta, t)/t \leq \beta^2$ for all $t \in (0, \mathsf{T}]$.

Finally, we prove (3.6.10). Since $|\mathsf{mmse}'(t)| \leq 1$ for all $t \geq 0$ we find that for $t_1, t_2 \geq 0$,

$$\begin{aligned}|\gamma_*(\beta, t_1) - \gamma_*(\beta, t_2)| &= \beta^2|\mathsf{mmse}(\gamma_*(\beta, t_1) + t_1) - \mathsf{mmse}(\gamma_*(\beta, t_2) + t_2)| \\ &\leq \beta^2|\gamma_*(\beta, t_1) - \gamma_*(\beta, t_2)| + \beta^2|t_1 - t_2|.\end{aligned}$$

Rearranging, we obtain

$$\frac{|\gamma_*(\beta,t_1)-\gamma_*(\beta,t_2)|}{|t_1-t_2|} \le \frac{\beta^2}{1-\beta^2}.$$

□

For $(\boldsymbol{A},\boldsymbol{y}) \sim \mathbb{P}$ and $\boldsymbol{x} \sim \mu_{\boldsymbol{A},\boldsymbol{y}(t)}$, define

$$\mathrm{MSE}_{\mathsf{AMP}}(k;\beta,t) = \operatorname*{p-lim}_{N\to\infty} \frac{1}{N}\,\mathbb{E}\left\|\boldsymbol{x}-\hat{\boldsymbol{m}}^k(\boldsymbol{A},\boldsymbol{y}(t))\right\|_2^2, \quad \hat{\boldsymbol{m}}^k(\boldsymbol{A},\boldsymbol{y}(t)) := \mathsf{AMP}(\boldsymbol{A},\boldsymbol{y}(t);k)\,, \tag{3.6.13}$$

where the limit is guaranteed to exist by Proposition 3.6.1.

Lemma 3.6.4. *We have*

$$\mathrm{MSE}_{\mathsf{AMP}}(k;\beta,t) = 1 - \frac{\gamma_{k+1}(\beta,t)}{\beta^2}\,.$$

In particular,

$$\lim_{k\to\infty} \mathrm{MSE}_{\mathsf{AMP}}(k;\beta,t) = 1 - \frac{\gamma_*(\beta,t)}{\beta^2}\,.$$

Proof. By state evolution

$$\begin{aligned}
\mathrm{MSE}_{\mathsf{AMP}}(k;\beta,t) &= \operatorname*{p-lim}_{N\to\infty} \frac{1}{N}\,\mathbb{E}\left\|\hat{\boldsymbol{m}}^k(\boldsymbol{A},\boldsymbol{y}(t))-\boldsymbol{x}\right\|_2^2 \\
&= \mathbb{E}\left[\big(\tanh(\gamma_k X+\sigma_k W+Y)-X\big)^2\right] \\
&= \mathbb{E}\left[\big(\tanh(\widetilde{\gamma}_k X+\widetilde{\sigma}_k W)-X\big)^2\right] \\
&= 1-2\,\mathbb{E}[\tanh(\widetilde{\gamma}_k X+\widetilde{\sigma}_k W)X]+\mathbb{E}[\tanh(\widetilde{\gamma}_k X+\widetilde{\sigma}_k W)^2] \\
&= 1-2\gamma_{k+1}/\beta^2+\sigma_{k+1}^2/\beta^2 \\
&= 1-\gamma_{k+1}/\beta^2,
\end{aligned}$$

where the last line follows from Proposition 3.6.2. □

We next show that, for any $t>0$, the mean square error achieved by AMP is the same as the Bayes optimal error, i.e., the mean squared error achieved by the posterior expectation $\mathbf{m}(\boldsymbol{A},\boldsymbol{y}(t))$.

Proposition 3.6.5. *Fix $\beta<1$ and $t\ge 0$. We have*

$$\lim_{N\to\infty} \frac{1}{N}\,\mathbb{E}\left[\left\|\boldsymbol{x}-\mathbf{m}(\boldsymbol{A},\boldsymbol{y}(t))\right\|_2^2\right] = \frac{\gamma_*(\beta,t)}{\beta^2}\,. \tag{3.6.14}$$

Proof. The proof is an adaptation from [DAM17], which we will present succinctly.

Let $I(X;Y)$ denote the mutual information between random variables X, Y on the same probability space. Letting $X \sim \mathsf{Unif}(\{-1,+1\})$ independent of $W \sim \mathcal{N}(0,1)$, define the function

$$\mathsf{I}(\gamma) := I\big(X; \gamma X + \sqrt{\gamma} W\big) \tag{3.6.15}$$

$$= \gamma - \mathbb{E} \log \cosh\big(\gamma + \sqrt{\gamma} W\big)\,. \tag{3.6.16}$$

We also define the function

$$\Psi(\gamma; \beta, t) := \frac{\beta^2}{4} + \frac{\gamma^2}{4\beta^2} - \frac{\gamma}{2} + \mathsf{I}(\gamma + t)\,. \tag{3.6.17}$$

As in [DAM17], it is easy to check that $\partial_\gamma \Psi(\gamma_*(\beta,t);\beta,t) = 0$ and, using the continuity of $(\beta,t) \mapsto \gamma_*(\beta,t)$,

$$\frac{\mathrm{d}}{\mathrm{d}(\beta^2)} \Psi(\gamma_*(\beta,t);\beta,t) = \frac{1}{4}\Big(1 - \frac{\gamma_*(\beta,t)^2}{\beta^4}\Big)\,, \tag{3.6.18}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} \Psi(\gamma_*(\beta,t);\beta,t) = \frac{1}{2}\Big(1 - \frac{\gamma_*(\beta,t)}{\beta^2}\Big)\,. \tag{3.6.19}$$

We further note that by the de Brujin identity (also known as I-MMSE relation [GSV05])

$$\frac{\mathrm{d}}{\mathrm{d}(\beta^2)} I(\boldsymbol{x}; \boldsymbol{A}(\beta), \boldsymbol{y}(t)) = \frac{1}{4n}\, \mathbb{E}\left[\big\|\boldsymbol{x}\boldsymbol{x}^\top - \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top | \boldsymbol{A}(\beta), \boldsymbol{y}(t)]\big\|_F^2\right]\,, \tag{3.6.20}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} I(\boldsymbol{x}; \boldsymbol{A}(\beta), \boldsymbol{y}(t)) = \frac{1}{2}\, \mathbb{E}\left[\big\|\boldsymbol{x} - \mathbb{E}[\boldsymbol{x} | \boldsymbol{A}(\beta), \boldsymbol{y}(t)]\big\|_2^2\right]\,. \tag{3.6.21}$$

Here we write $\boldsymbol{A} = \boldsymbol{A}(\beta)$ to emphasize the dependence upon β. Using Eqs. (3.6.18) and (3.6.20), we have

$$\begin{aligned}
\log 2 - \mathsf{I}(t) &= \lim_{\beta\to\infty} \lim_{N\to\infty} \frac{1}{N}\big[I(\boldsymbol{x}; \boldsymbol{A}(\beta), \boldsymbol{y}(t)) - I(\boldsymbol{x}; \boldsymbol{A}(0), \boldsymbol{y}(t))\big] \\
&= \lim_{N\to\infty} \int_0^\infty \frac{1}{4n}\, \mathbb{E}\left[\big\|\boldsymbol{x}\boldsymbol{x}^\top - \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top | \boldsymbol{A}(\beta), \boldsymbol{y}(t)]\big\|_F^2\right] \mathrm{d}\beta^2 \\
&\le \lim_{k\to\infty} \lim_{N\to\infty} \int_0^\infty \frac{1}{4n}\, \mathbb{E}\left[\big\|\boldsymbol{x}\boldsymbol{x}^\top - \hat{\boldsymbol{m}}^k(\boldsymbol{A}(\beta), \boldsymbol{y}(t))\hat{\boldsymbol{m}}^k(\boldsymbol{A}(\beta), \boldsymbol{y}(t))^\top\big\|_F^2\right] \mathrm{d}\beta^2 \\
&= \lim_{k\to\infty} \int_0^\infty \frac{1}{4}\Big(1 - \frac{\gamma_k(\beta,t)^2}{\beta^4}\Big)\, \mathrm{d}\beta^2 \\
&= \int_0^\infty \frac{1}{4}\Big(1 - \frac{\gamma_*(\beta,t)^2}{\beta^4}\Big)\, \mathrm{d}\beta^2 \\
&= \lim_{\beta\to\infty} \big[\Psi(\gamma_*(\beta,t);\beta,t) - \Psi(\gamma_*(0,t);0,t)\big]\,.
\end{aligned}$$

(The exchanges of limits are justified by dominated convergence.)

Finally, a direct calculation reveals that $\lim_{\beta\to\infty} \big[\Psi(\gamma_*(\beta,t);\beta,t) - \Psi(\gamma_*(0,t);0,t)\big] = \log(2) - \mathsf{I}(t)$

and therefore equality holds at each of the steps above. We deduce that $\lim_{N\to\infty} n^{-1} I(\boldsymbol{x}; \boldsymbol{A}(\beta), \boldsymbol{y}(t)) = \Psi(\gamma_*(\beta,t);\beta,t)$.

Using this fact, together with Eqs. (3.6.19), (3.6.21) and the fact that the right hand sides of these equations are monotone decreasing in t, we get that the following holds for almost every $t > 0$:

$$\lim_{N\to\infty} \frac{1}{N} \mathbb{E}\left[\left\|\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}|\boldsymbol{A}(\beta), \boldsymbol{y}(t)]\right\|_2^2\right] = 1 - \frac{\gamma_*(\beta,t)}{\beta^2}\,. \tag{3.6.22}$$

This coincides with the claim (3.6.14), and actually holds for every $t > 0$ since the right-hand side of Eq. (3.6.14) is continuous in $t > 0$ by Lemma 3.6.3. □

It follows that AMP approximately computes the posterior mean $\mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(t))$ in the following sense.

Proposition 3.6.6. *Fix $\beta < 1$, $\mathsf{T} > 0$ and let $t \in (0, \mathsf{T}]$. Recalling that $\hat{\boldsymbol{m}}^k(\boldsymbol{A}, \boldsymbol{y}(t)) := \mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}(t); k)$ denotes the AMP estimate afte k iterations, and that $\boldsymbol{z}^k$ is defined by Eq.* (3.6.1)*, we have*

$$\lim_{k\to\infty} \sup_{t\in(0,\mathsf{T})} \operatorname*{p-lim}_{N\to\infty} \frac{\|\mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(t)) - \hat{\boldsymbol{m}}^k(\boldsymbol{A}, \boldsymbol{y}(t))\|_2}{\|\mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(t))\|_2} = 0\,. \tag{3.6.23}$$

Moreover

$$\lim_{k\to\infty} \sup_{t\in(0,\mathsf{T})} \operatorname*{p-lim}_{N\to\infty} \frac{\|\boldsymbol{z}^{k+1} - \boldsymbol{z}^k\|}{\|\boldsymbol{z}^k\|} = 0\,. \tag{3.6.24}$$

Remark 3.6.2. A somewhat similar result has recently been proved by Chen and Tang [CT21] where the external field vector $\boldsymbol{y}(t)$ is replaced by a multiple of the all-ones vector $h\mathbf{1}$, for any pair (β, h) for which a certain condition of uniform concentration of the overlap between two independent draws from the measure $\mu_{\boldsymbol{A},h\mathbf{1}}$ holds. In our setting, we are concerned with a different family of external fields, namely the ones generated by the stochastic localization process (3.1.4). The argument, which proceeds via the planted model, does not require the uniform concentration condition.

Proof. Throughout this proof we write $\boldsymbol{y}$ instead of $\boldsymbol{y}(t)$ for ease of notation. To show Eq. (3.6.23), observe that the bias-variance decomposition yields (recalling the definition $\mathrm{MSE}_{\mathsf{AMP}}(\,\cdot\,)$ in Eq. (3.6.13))

$$\mathrm{MSE}_{\mathsf{AMP}}(k;\beta,t) = \operatorname*{p-lim}_{N\to\infty} \left\{ \frac{1}{N} \mathbb{E}\left[\left\|\hat{\boldsymbol{m}}^k(\boldsymbol{A}, \boldsymbol{y}) - \mathbf{m}(\boldsymbol{A}, \boldsymbol{y})\right\|_2^2\right] + \frac{1}{N} \mathbb{E}\left[\left\|\boldsymbol{x} - \mathbf{m}(\boldsymbol{A}, \boldsymbol{y})\right\|_2^2\right] \right\}.$$

Using Lemma 3.6.4 for the left-hand side and Proposition 3.6.5 for the second step the second term on the right-hand side, we get

$$\operatorname*{p-lim}_{N\to\infty} \frac{1}{N} \mathbb{E}\left[\left\|\hat{\boldsymbol{m}}^k(\boldsymbol{A}, \boldsymbol{y}) - \mathbf{m}(\boldsymbol{A}, \boldsymbol{y})\right\|_2^2\right] = \frac{\gamma_*(\beta,t) - \gamma_{k+1}(\beta,t)}{\beta^2}\,. \tag{3.6.25}$$

Claim (3.6.23) now follows by combining Eq. (3.6.25) with Eqs. (3.6.8) and (3.6.9) of Lemma 3.6.3.

Finally, Eq. (3.6.24) is an immediate consequence of Proposition 3.6.1 and Proposition 3.6.2. Indeed, by Proposition 3.6.1, we have

$$\begin{aligned}
\operatorname*{p-lim}_{N\to\infty} \frac{1}{N}\|\boldsymbol{z}^k\|_2^2 &= \mathbb{E}\left[(\gamma_k X + G_k + Y)^2\right] = (\gamma_k + t)^2 + \gamma_k + t\,,\\
\operatorname*{p-lim}_{N\to\infty} \frac{1}{N}\|\boldsymbol{z}^{k+1} - \boldsymbol{z}^k\|_2^2 &= \mathbb{E}\left[\left((\gamma_{k+1} - \gamma_k)X + G_{k+1} - G_k\right)^2\right]\\
&= (\gamma_{k+1} - \gamma_k)^2 + (\Sigma_{k+1,k+1} - 2\Sigma_{k,k+1} + \Sigma_{k,k})\\
&= (\gamma_{k+1} - \gamma_k)^2 + (\gamma_{k+1} - \gamma_k)\,,
\end{aligned}$$

where in the last step we used Proposition 3.6.2. We therefore obtained we have

$$\operatorname*{p-lim}_{N\to\infty} \frac{\|\boldsymbol{z}^{k+1} - \boldsymbol{z}^k\|_2^2}{\|\boldsymbol{z}^k\|_2^2} = \frac{(\gamma_{k+1} - \gamma_k)^2 + (\gamma_{k+1} - \gamma_k)}{(\gamma_k + t)^2 + \gamma_k + t}\,.$$

Hence Eq. (3.6.24) also follows from Eq. (3.6.8). □

We conclude this subsection with a lemma controlling the regularity of the posterior path $t \mapsto \mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(t))$, which will be useful later.

Lemma 3.6.7. *Fix $\beta < 1$ and $0 \le t_1 < t_2 \le \mathsf{T}$. Then*

$$\begin{aligned}
\operatorname*{p-lim}_{N\to\infty} \sup_{t\in[t_1,t_2]} \frac{1}{N}\left\|\mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(t)) - \mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(t_1))\right\|_2^2 &= \operatorname*{p-lim}_{N\to\infty} \frac{1}{N}\left\|\mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(t_2)) - \mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(t_1))\right\|_2^2\\
&= \frac{\gamma_*(\beta, t_2) - \gamma_*(\beta, t_1)}{\beta^2}\,. \qquad (3.6.26)
\end{aligned}$$

Proof. We will exploit the fact that $(\mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(t)))_{t\ge 0}$ is a martingale, as a consequence of Lemma 3.3.1 (with $\varphi : \mathbb{R}^N \to \mathbb{R}^N$ given by $\varphi(\boldsymbol{x}) = \boldsymbol{x}$).

Using Proposition 3.6.5, we obtain, for any $t_1 < t_2$

$$\begin{aligned}
\lim_{N\to\infty} \frac{1}{N}\,\mathbb{E}\left[\left\|\mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(t_2)) - \mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(t_1))\right\|_2^2\right] &= \operatorname*{p-lim}_{N\to\infty} \frac{\mathbb{E}\left[\left\|\boldsymbol{x} - \mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(t_1))\right\|_2^2\right] - \mathbb{E}\left[\left\|\boldsymbol{x} - \mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(t_1))\right\|_2^2\right]}{N}\\
&= \frac{\gamma_*(\beta, t_2) - \gamma_*(\beta, t_1)}{\beta^2}\,,
\end{aligned}$$

where the first equality uses the fact that $\mathbb{E}[\mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(t_2))|\boldsymbol{A}, \boldsymbol{y}(t_1)] = \mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(t_1))$. By Lemma 3.6.6, we have, with high probability, $\|\mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(t_i)) - \hat{\boldsymbol{m}}^k(\boldsymbol{A}, \boldsymbol{y}(t))\|_2^2/N \le \varepsilon_k$, for some deterministic

constants ε_k so that $\varepsilon_k \to 0$ as $k \to \infty$. As a consequence

$$\operatorname*{p-lim}_{N\to\infty} \frac{1}{N}\big\|\mathbf{m}(\boldsymbol{A},\boldsymbol{y}(t_2)) - \mathbf{m}(\boldsymbol{A},\boldsymbol{y}(t_1))\big\|_2^2 = \frac{\gamma_*(\beta,t_2)-\gamma_*(\beta,t_1)}{\beta^2}\,. \tag{3.6.27}$$

Now, since $t \to \mathbf{m}_{\boldsymbol{A},\boldsymbol{y}(t)}$ is a bounded martingale, it follows that, for any fixed constant c, the process

$$Y_{N,t} := \big|M_{N,t} - c\big|\,, \quad \text{where} \quad M_{N,t} := \frac{1}{\sqrt{N}}\big\|\mathbf{m}(\boldsymbol{A},\boldsymbol{y}(t)) - \mathbf{m}(\boldsymbol{A},\boldsymbol{y}(t_1))\big\|_2\,, \tag{3.6.28}$$

is a positive bounded submartingale for $t \ge t_1$. Therefore by Doob's maximal inequality [Dur19],

$$\mathbb{P}\Big(\sup_{t\in[t_1,t_2]} Y_{N,t} \ge a\Big) \le \frac{1}{a}\,\mathbb{E}\left[Y_{N,t_2}\right] \le \frac{1}{a}\,\mathbb{E}\left[Y_{N,t_2}^2\right]^{1/2}, \tag{3.6.29}$$

for any $a > 0$. We choose $c = \sqrt{\gamma_*(\beta,t_2)-\gamma_*(\beta,t_1)}/\beta$. By (3.6.27), we have

$$\operatorname*{p-lim}_{N\to\infty} M_{N,t_2}^2 = \frac{\gamma_*(t_2)-\gamma_*(t_1)}{\beta^2} = c^2\,,$$

and therefore, since $M_{N,t}$ is bounded, for any fixed $a > 0$

$$\begin{aligned}\lim_{N\to\infty} \mathbb{P}\Big(\sup_{t\in[t_1,t_2]} M_{N,t} \ge c + a\Big) &\le \lim_{N\to\infty} \mathbb{P}\Big(\sup_{t\in[t_1,t_2]} Y_{N,t} \ge a\Big)\\ &\le \frac{1}{a}\lim_{N\to\infty} \mathbb{E}\left[(M_{N,t_2}-c)^2\right]^{1/2} = 0\,.\end{aligned}$$

Together with Eq. (3.6.27), this yields

$$\operatorname*{p-lim}_{N\to\infty} \sup_{t\in[t_1,t_2]} M_{N,t}^2 = \frac{\gamma_*(t_2)-\gamma_*(t_1)}{\beta^2}\,,$$

which coincides with the claim (3.6.26). □

3.7 Natural Gradient Descent

The main objective of this section is to show that $\mathscr{F}_{\text{TAP}}(\mathbf{m};\boldsymbol{y},q)$ behaves well for $q = q_*(\beta,t)$ and for $\mathbf{m}$ in a neighborhood of $\hat{\boldsymbol{m}}^{K_{\mathsf{AMP}}}$. Namely it has a unique local minimum $\mathbf{m}_* = \mathbf{m}_*(\boldsymbol{A},\boldsymbol{y})$ in such a neighborhood, and NGD approximates $\mathbf{m}_*$ well for large number of iterations K. Crucially, the map $\boldsymbol{y} \mapsto \mathbf{m}_*$ will be Lipschitz. For reference, we reproduce the NGD algorithm as Algorithm 3. This corresponds to lines 6-11 of Algorithm 1.

Lemma 3.7.1. *Let $\beta < \frac{1}{2}$, $c \in (0, 1-2\beta)$, and $\mathsf{T} > 0$ be fixed. Then there exists $\varepsilon_0 = \varepsilon_0(\beta,\mathsf{T})$*

Algorithm 3: NATURAL GRADIENT DESCENT ON $\mathscr{F}_{\mathrm{TAP}}(\,\cdot\,;\boldsymbol{y},q)$

Input: Initialization $\boldsymbol{u}^0 \in \mathbb{R}^N$, data $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, $\hat{\boldsymbol{y}} \in \mathbb{R}^N$, step size $\eta > 0$, $q \in (0,1)$, integer $K > 0$.

$\hat{\boldsymbol{m}}^{+,0} = \tanh(\boldsymbol{u}^0)$.
for $k = 0, \cdots, K-1$ **do**
  $\boldsymbol{u}^{k+1} \leftarrow \boldsymbol{u}^k - \eta \cdot \nabla \mathscr{F}_{\mathrm{TAP}}(\hat{\boldsymbol{m}}^{+,k}; \boldsymbol{y}, q)$,
  $\hat{\boldsymbol{m}}^{+,k+1} = \tanh(\boldsymbol{u}^{+,k+1})$,
end
return $\hat{\boldsymbol{m}}^{+,K}$

such that, for all $\varepsilon \in (0,\varepsilon_0)$ there exists $K_{\mathsf{AMP}} = K_{\mathsf{AMP}}(\beta,\mathsf{T},\varepsilon)$ and $\rho_0 = \rho_0(\beta,\mathsf{T},\varepsilon)$ such that for all $\rho \in (0,\rho_0)$ there exists $K_{\mathsf{NGD}} = K_{\mathsf{NGD}}(\beta,\mathsf{T},\varepsilon,\rho)$, such that the following holds.

Let $\hat{\boldsymbol{m}}^{\mathsf{AMP}} = \mathsf{AMP}(\boldsymbol{A},\boldsymbol{y}(t);K_{\mathsf{AMP}})$ be the output of the AMP after K_{AMP} iterations, when applied to $\boldsymbol{y}(t)$. Fix $K \ge K_{\mathsf{AMP}}$. With probability $1-o_N(1)$ over $(\boldsymbol{A},\boldsymbol{y}) \sim \mathbb{P}$, for all $t \in (0,\mathsf{T}]$ and all $\hat{\boldsymbol{y}} \in B\left(\boldsymbol{y}(t), c\sqrt{\varepsilon t N}/4\right)$, setting $q_* := q_*(\beta,t)$:

1. *The function*
$$\mathbf{m} \mapsto \mathscr{F}_{\mathrm{TAP}}(\mathbf{m};\hat{\boldsymbol{y}},q_*)$$
restricted to $B\left(\hat{\boldsymbol{m}}^{\mathsf{AMP}}, \sqrt{\varepsilon t N}\right) \cap (-1,1)^N$ has a unique stationary point
$$\mathbf{m}_*(\boldsymbol{A},\hat{\boldsymbol{y}}) \in B\left(\hat{\boldsymbol{m}}^{\mathsf{AMP}}, \sqrt{\varepsilon t N}/2\right) \cap (-1,1)^N$$
which is also a local minimum. In the case $\hat{\boldsymbol{y}} = \boldsymbol{y}(t)$, $\mathbf{m}_*(\boldsymbol{A},\boldsymbol{y}(t))$ also satisfies
$$\mathbf{m}_*(\boldsymbol{A},\boldsymbol{y}) \in B\left(\hat{\boldsymbol{m}}^{k'}, \sqrt{\varepsilon t N}/2\right) \cap (-1,1)^N$$
for all $k' \in [K_{\mathsf{AMP}}, K]$, where $\hat{\boldsymbol{m}}^{k'} = \mathsf{AMP}(\boldsymbol{A},\boldsymbol{y}(t);k')$.

2. *The stationary point $\mathbf{m}_*(\boldsymbol{A},\hat{\boldsymbol{y}})$ satisfies (recall that $\mathbf{m}(\boldsymbol{A},\boldsymbol{y})$ denotes the mean of the Gibbs measure)*
$$\left\|\mathbf{m}(\boldsymbol{A},\boldsymbol{y}) - \mathbf{m}_*(\boldsymbol{A},\boldsymbol{y})\right\|_2 \le \rho\sqrt{tN}\,.$$

3. *The stationary point $\mathbf{m}_*$ obeys the following Lipschitz property for all $\hat{\boldsymbol{y}}, \hat{\boldsymbol{y}}' \in B\left(\boldsymbol{y}(t), c\sqrt{\varepsilon t N}/4\right)$:*
$$\left\|\mathbf{m}_*(\boldsymbol{A},\hat{\boldsymbol{y}}) - \mathbf{m}_*(\boldsymbol{A},\hat{\boldsymbol{y}}')\right\| \le c^{-1}\|\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}'\|\,. \tag{3.7.1}$$

4. *There exists a learning rate $\eta = \eta(\beta,\mathsf{T},\varepsilon)$ such that the following holds. Let $\hat{\boldsymbol{m}}^{\mathsf{NGD}}(\boldsymbol{A},\hat{\boldsymbol{y}})$ be the output of NGD (Algorithm 3), when run for K_{NGD} iterations with parameter q_*, $\hat{\boldsymbol{y}}$, η. Assume*

that the initialization $\boldsymbol{u}^0$ *satisfies*

$$\left\|\boldsymbol{u}^0 - \operatorname{arctanh}(\hat{\boldsymbol{m}}^{\mathsf{AMP}})\right\| \le \frac{c\sqrt{\varepsilon t N}}{200}\,. \tag{3.7.2}$$

Then the algorithm output satisfies

$$\left\|\hat{\boldsymbol{m}}^{\mathsf{NGD}}(\boldsymbol{A}, \hat{\boldsymbol{y}}) - \mathbf{m}_*(\boldsymbol{A}, \hat{\boldsymbol{y}})\right\| \le \rho\sqrt{tN}\,. \tag{3.7.3}$$

The proof of this lemma is deferred to the appendix. Here we will prove the two key elements: first that $\hat{\boldsymbol{m}}^{\mathsf{AMP}}$ is an approximate stationary point of $\mathscr{F}_{\mathrm{TAP}}(\,\cdot\,; \boldsymbol{y}(t), q_*)$ (Lemma 3.7.2), and second that $\mathscr{F}_{\mathrm{TAP}}(\,\cdot\,; \hat{\boldsymbol{y}}, q_*)$ is strongly convex in a neighborhood of $\hat{\boldsymbol{m}}^{\mathsf{AMP}}$ (Lemma 3.7.3). Let us point out that, in the local convexity guarantee, it is important that the neighborhood has radius $\Theta(\sqrt{tN})$ as $t \to 0$.

We recall below the expressions for the gradient and Hessian of $\mathscr{F}_{\mathrm{TAP}}(\,\cdot\,; \boldsymbol{y}, q)$ at $\mathbf{m} \in (-1,1)^N$:

$$\nabla \mathscr{F}_{\mathrm{TAP}}(\mathbf{m}; \boldsymbol{y}, q) = -\beta \boldsymbol{A}\mathbf{m} - \boldsymbol{y} + \operatorname{arctanh}(\mathbf{m}) + \beta^2 (1-q)\,\mathbf{m} \tag{3.7.4}$$

$$\nabla^2 \mathscr{F}_{\mathrm{TAP}}(\mathbf{m}; \boldsymbol{y}, q) = -\beta \boldsymbol{A} + \boldsymbol{D}(\mathbf{m}) + \beta^2 (1-q)\,\boldsymbol{I}_N\,, \quad \boldsymbol{D}(\mathbf{m}) := \operatorname{diag}\big(\{(1-m_i^2)^{-1}\}_{i\le N}\big). \tag{3.7.5}$$

In (3.7.4), arctanh is applied coordinate-wise to $\mathbf{m} \in (-1,1)^N$.

For $t > 0, k \ge 0$ we let $\hat{\boldsymbol{m}}^k = \mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}(t); k)$ and define the quantities

$$q_k(\beta, t) := \frac{\gamma_{k+1}(\beta, t)}{\beta^2}\,, \qquad q_*(\beta, t) := \frac{\gamma_*(\beta, t)}{\beta^2}\,. \tag{3.7.6}$$

Note that, by Lemma 3.6.4, we have

$$q_k(\beta, t) = \operatorname*{p-lim}_{N\to\infty} \frac{\left\|\hat{\boldsymbol{m}}^k\right\|^2}{N}\,, \qquad q_*(\beta, t) = \lim_{k\to\infty} q_k(\beta, t)\,. \tag{3.7.7}$$

We will use the bounds (3.6.8), (3.6.9) in Lemma 3.6.3 several times below, which ensures that $(q_k(\beta,t)/t) \in [c, C]$ holds for constants $c, C > 0$ independent of $t \in (0, \mathsf{T}]$ and $k \ge 1$.

Lemma 3.7.2. *Let* $\hat{\boldsymbol{m}}^k = \hat{\boldsymbol{m}}^k(\boldsymbol{A}, \boldsymbol{y}(t))$ *denote the AMP iterates on input* $\boldsymbol{A}, \boldsymbol{y}(t)$*. Then for any* $\mathsf{T} > 0$*,*

$$\lim_{k\to\infty} \sup_{t\in(0,\mathsf{T}]} \sup_{q\in[q_k(\beta,t), q_*(\beta,t)]} \operatorname*{p-lim}_{N\to\infty} \frac{\left\|\nabla \mathscr{F}_{\mathrm{TAP}}(\hat{\boldsymbol{m}}^k; \boldsymbol{y}(t), q)\right\|}{\sqrt{tN}} = 0.$$

Proof. As in Algorithm 1, let

$$\boldsymbol{z}^{k+1} = \operatorname{arctanh}(\hat{\boldsymbol{m}}^{k+1}) = \beta \boldsymbol{A}\hat{\boldsymbol{m}}^k + \boldsymbol{y} - \beta^2 \left(1 - \frac{1}{N}\left\|\hat{\boldsymbol{m}}^k\right\|^2\right) \hat{\boldsymbol{m}}^{k-1}.$$

Let $q \in [q_k(\beta,t), q_*(\beta,t)]$. Combining the above with Eqs. (3.7.4) and (3.7.7) yields

$$\begin{aligned}
\frac{1}{\sqrt{N}}\|\nabla \mathscr{F}_{\text{TAP}}(\mathsf{AMP}(\boldsymbol{A},\boldsymbol{y};k);\boldsymbol{y},q)\| &= \frac{1}{\sqrt{N}}\Big\|-\beta \boldsymbol{A}\hat{\boldsymbol{m}}^k - \boldsymbol{y} + \operatorname{arctanh}(\hat{\boldsymbol{m}}^k) + \beta^2(1-q)\hat{\boldsymbol{m}}^k\Big\| \\
&= \frac{1}{\sqrt{N}}\Big\|\boldsymbol{z}^k - \beta \boldsymbol{A}\hat{\boldsymbol{m}}^k - \boldsymbol{y} + \beta^2\,(1-q)\,\hat{\boldsymbol{m}}^k\Big\| \\
&\le \frac{1}{\sqrt{N}}\|\boldsymbol{z}^{k+1}-\boldsymbol{z}^k\| + \frac{1}{\sqrt{N}}\Big\|\boldsymbol{z}^{k+1} - \beta \boldsymbol{A}\hat{\boldsymbol{m}}^k - \boldsymbol{y} + \beta^2\,(1-q)\,\hat{\boldsymbol{m}}^k\Big\| \\
&= \frac{1}{\sqrt{N}}\big\|\boldsymbol{z}^{k+1}-\boldsymbol{z}^k\big\| + \frac{\beta^2}{\sqrt{N}}\Big\|\big(1-\big\|\hat{\boldsymbol{m}}^k\big\|^2/N\big)\hat{\boldsymbol{m}}^{k-1} - (1-q)\,\hat{\boldsymbol{m}}^k\Big\| \\
&\le \frac{1}{\sqrt{N}}\big\|\boldsymbol{z}^{k+1}-\boldsymbol{z}^k\big\| + \frac{\beta^2}{\sqrt{N}}\big\|\hat{\boldsymbol{m}}^{k-1} - \hat{\boldsymbol{m}}^k\big\| \\
&\quad + \beta^2(q_*(\beta,t) - q_k(\beta,t)) + o_{N,\mathbb{P}}(1).
\end{aligned}$$

Here $o_{N,\mathbb{P}}(1)$ denotes terms which converge to 0 in probability as $N \to \infty$. By (3.6.24), (3.7.7) and the bound $(q_k(\beta,t)/t) \in [c,C]$

$$\lim_{k\to\infty} \sup_{t\in(0,T)} \operatorname*{p-lim}_{N\to\infty} \frac{\|\boldsymbol{z}^{k+1}-\boldsymbol{z}^k\|}{\sqrt{tN}} = 0.$$

Moreover, $\|\hat{\boldsymbol{m}}^{k-1} - \hat{\boldsymbol{m}}^k\| \le \|\boldsymbol{z}^{k-1} - \boldsymbol{z}^k\|$ since the function $x \mapsto \tanh(x)$ is 1-Lipschitz. Finally (3.6.8) and (3.6.9) of Lemma 3.6.3 imply

$$\lim_{k\to\infty} \sup_{t\in(0,\mathsf{T}]} \frac{q_*(\beta,t) - q_k(\beta,t)}{\sqrt{t}} = 0\,.$$

Combining the above statements concludes the proof. □

We next control on the Hessian $\nabla^2 \mathscr{F}_{\text{TAP}}(\,\cdot\,;\boldsymbol{y},q)$. As anticipated in Remark 3.2.1, this is the only part of our proof that requires $\beta < 1/2$ instead of $\beta < 1$.

Lemma 3.7.3. *Let $\beta > 0$, $\boldsymbol{y} \in \mathbb{R}^N$ and $q \in [0,1]$. Then for all $\mathbf{m} \in (-1,1)^N$,*

$$\big(1-\beta\|\boldsymbol{A}\|_{\text{op}}\big)\boldsymbol{D}(\mathbf{m}) \preceq \nabla^2 \mathscr{F}_{\text{TAP}}(\mathbf{m};\boldsymbol{y},q) \preceq \big(1+\beta^2+\beta\|\boldsymbol{A}\|_{\text{op}}\big)\boldsymbol{D}(\mathbf{m})\,. \tag{3.7.8}$$

In particular if $\beta \le \frac{1}{2} - c$, for $c > 0$, then with probability $1 - o_N(1)$, for all $\mathbf{m} \in (-1,1)^N$,

$$c\,\boldsymbol{D}(\mathbf{m}) \preceq \nabla^2 \mathscr{F}_{\text{TAP}}(\mathbf{m};\boldsymbol{y},q) \preceq 2\boldsymbol{D}(\mathbf{m})\,. \tag{3.7.9}$$

Proof. The upper and lower bounds in Eq. (3.7.8) are obtained from (3.7.5) using the fact that $\boldsymbol{D}(\mathbf{m}) \succeq \boldsymbol{I}_N$ for all $\mathbf{m} \in (-1,1)^N$. Further, we use the fact that $\|\boldsymbol{A}\|_{\text{op}} \le 2 + o_N(1)$ with probability $1 - o_N(1)$. Therefore, Eq. (3.7.9) follows from the assumption $\beta \le \frac{1}{2} - c$. □

As mentioned above, our convergence analysis of NGD, and proof of Lemma 3.7.1 are given in Appendix 3.10. The key insight is that the main iterative step in line 3 of Algorithm 3 can be expressed as a version of mirror descent. Define the concave function $h(\mathbf{m}) = \sum_{i=1}^{N} h(m_i)$ for $\mathbf{m} \in (-1,1)^N$ (recall that $h(x) := -((1+x)/2)\log((1+x)/2) - ((1-x)/2)\log((1-x)/2)$). Following [LFN18], we define for $\mathbf{m}, \boldsymbol{n} \in (-1,1)^N$ the Bregman divergence

$$D_{-\boldsymbol{h}}(\mathbf{m}, \boldsymbol{n}) = -\boldsymbol{h}(\mathbf{m}) + \boldsymbol{h}(\boldsymbol{n}) + \langle \nabla \boldsymbol{h}(\boldsymbol{n}), \mathbf{m} - \boldsymbol{n} \rangle \ . \tag{3.7.10}$$

Then with $L = 1/\eta$, the update in line 3 admits the alternate description

$$\hat{\boldsymbol{m}}^{+,k+1} = \arg\min_{\boldsymbol{x} \in (-1,1)^N} \left\langle \nabla \mathscr{F}_{\text{TAP}}(\hat{\boldsymbol{m}}^{+,k}; \boldsymbol{y}, q), \boldsymbol{x} - \hat{\boldsymbol{m}}^{+,k} \right\rangle + L \cdot D_{-\boldsymbol{h}}(\boldsymbol{x}, \hat{\boldsymbol{m}}^{+,k}) \, . \tag{3.7.11}$$

We will use this description to prove convergence.

Remark 3.7.1. If the Hessian $\nabla^2 \mathscr{F}_{\text{TAP}}$ were bounded above and below by constant multiples of the **identity** matrix instead of $\boldsymbol{D}(\mathbf{m})$, then we could use simple gradient descent instead of NGD in Algorithm 1. This would also simplify the proof. However, $\nabla^2 \mathscr{F}_{\text{TAP}}$ is not bounded above near the boundaries of $(-1,+1)^N$. The use of NGD to minimize TAP free energy was introduced in [CFM21], which however considered a different regime in the planted model.

Remark 3.7.2. Our proof of Lemma 3.7.1 does not require $\nabla^2 \mathscr{F}_{\text{TAP}}$ to be globally convex. Instead, we only use the fact that, with probability $1 - o_N(1)$,

$$\nabla^2 \mathscr{F}_{\text{TAP}}(\mathbf{m}; \boldsymbol{y}, q) \succeq c\boldsymbol{D}(\mathbf{m}), \qquad \forall \mathbf{m} \in B\left(\hat{\boldsymbol{m}}^{\mathsf{AMP}}, \sqrt{\varepsilon t N}\right) \cap (-1,1).$$

For $\beta \in [1/2, 1)$ we expect only this weaker guarantee to hold. We believe the technique of [CFM21] could be used to prove such local strong convexity in the full regime $\beta \in [0,1)$.

3.8 Continuous limit and proof of Theorem 37

We fix (β, T) and choose constants $K_{\mathsf{AMP}} = K_{\mathsf{AMP}}(\beta, \mathsf{T}, \varepsilon)$, $\rho_0 = \rho_0(\beta, \mathsf{T}, \varepsilon, K_{\mathsf{AMP}})$, $\rho \in (0, \rho_0)$ and $K_{\mathsf{NGD}} = K_{\mathsf{NGD}}(\beta, \mathsf{T}, \varepsilon, \rho)$ so that Lemma 3.7.1 holds.

We couple the discretized process $(\hat{\boldsymbol{y}}_\ell)_{\ell \geq 0}$ defined in Eq. (3.2.6) (line 6 of Algorithm 2) to the continuous time process $(\boldsymbol{y}(t))_{t \in \mathbb{R}_{\geq 0}}$ (cf. Eq. (3.5.8)) via the driving noise, as follows:

$$\boldsymbol{w}_{\ell+1} = \frac{1}{\sqrt{\delta}} \int_{\ell\delta}^{(\ell+1)\delta} \mathrm{d}\boldsymbol{B}(t) \, . \tag{3.8.1}$$

We denote by $\hat{\boldsymbol{m}}(\boldsymbol{A}, \boldsymbol{y})$ the output of the mean estimation algorithm 1 on input $\boldsymbol{A}, \boldsymbol{y}$. By Lemma 3.7.1,

which ensures that, for any $t \in (0, \mathsf{T}]$, with probability $1 - o_N(1)$,

$$\left\|\hat{\boldsymbol{m}}(\boldsymbol{A}, \boldsymbol{y}(t)) - \mathbf{m}_*(\boldsymbol{A}, \boldsymbol{y}(t); q_*(\beta, t))\right\| \le \rho\sqrt{tN}\,. \tag{3.8.2}$$

Here and below we note explicitly the dependence of $\mathbf{m}_*$ on t via q_*. The next lemma provides a crude estimate on the Lipschitz continuity of AMP with respect to its input.

Lemma 3.8.1. *Recall that* $\mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; k) \in \mathbb{R}^N$ *denotes the output of the AMP algorithm on input* $(\boldsymbol{A}, \boldsymbol{y})$*, after* k *iterations, cf. Eq.* (3.2.2)*. If* $\|\boldsymbol{A}\|_{\mathrm{op}} \le 3$*, then, for any* $\boldsymbol{y}, \hat{\boldsymbol{y}} \in \mathbb{R}^N$*,*

$$\left\| \operatorname{arctanh}\big(\mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; k)\big) - \operatorname{arctanh}\big(\mathsf{AMP}(\boldsymbol{A}, \hat{\boldsymbol{y}}; k)\big)\right\|_2 \le k6^k \left\|\boldsymbol{y} - \hat{\boldsymbol{y}}\right\|_2 . \tag{3.8.3}$$

Proof. For $0 \le j \le k$, set:

$$\mathbf{m}^j = \mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; j), \qquad \boldsymbol{z}^j = \operatorname{arctanh}(\mathbf{m}^j), \qquad \mathsf{b}_j = \frac{\beta^2}{N}\sum_{i=1}^N \big(1 - \tanh^2(z_i^j)\big),$$
$$\hat{\boldsymbol{m}}^j = \mathsf{AMP}(\boldsymbol{A}, \hat{\boldsymbol{y}}; j), \qquad \hat{\boldsymbol{z}}^j = \operatorname{arctanh}(\hat{\boldsymbol{m}}^j), \qquad \hat{\mathsf{b}}_j = \frac{\beta^2}{N}\sum_{i=1}^N \big(1 - \tanh^2(\hat{z}_i^j)\big).$$

Using the AMP update equation (line 4 of Algorithm 1) and the fact that $\tanh(\,\cdot\,)$ is 1-Lipschitz, we obtain

$$\begin{aligned}\|\boldsymbol{z}^{j+1} - \hat{\boldsymbol{z}}^{j+1}\| &\le \|\beta\boldsymbol{A}(\mathbf{m}^j - \hat{\boldsymbol{m}}^j)\| + \|\boldsymbol{y} - \hat{\boldsymbol{y}}\| + \|\mathsf{b}_j\mathbf{m}^{j-1} - \mathsf{b}_j\hat{\boldsymbol{m}}^{j-1}\| + \|\mathsf{b}_j\hat{\boldsymbol{m}}^{j-1} - \hat{\mathsf{b}}_j\hat{\boldsymbol{m}}^{j-1}\| \\ &\le 3\beta\|\boldsymbol{z}^j - \hat{\boldsymbol{z}}^j\| + \|\boldsymbol{y} - \hat{\boldsymbol{y}}\| + \mathsf{b}_j\|\boldsymbol{z}^{j-1} - \hat{\boldsymbol{z}}^{j-1}\| + |\mathsf{b}_j - \hat{\mathsf{b}}_j|\sqrt{N}\,.\end{aligned}$$

Note that $|1 - \tanh^2(x)| \le 1$ for all $x \in \mathbb{R}$ and $|b_j| \le \beta^2$. Setting $E_j = \max_{i \le j}\|\boldsymbol{z}^{i+1} - \hat{\boldsymbol{z}}^{i+1}\|$, we find

$$\begin{aligned}E_{j+1} &\le (3\beta^2 + 3\beta)E_j + \|\boldsymbol{y} - \hat{\boldsymbol{y}}\| \\ &\le 6E_j + \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|\,.\end{aligned}$$

It follows by induction that

$$E_j \le j6^j\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|\,.$$

Setting $j = k$ concludes the proof. □

Define the random approximation errors

$$A_\ell := \frac{1}{\sqrt{N}} \big\| \hat{\boldsymbol{y}}_\ell - \boldsymbol{y}(\ell\delta) \big\| , \tag{3.8.4}$$

$$B_\ell := \frac{1}{\sqrt{N}} \big\| \hat{\boldsymbol{m}}(\boldsymbol{A}, \hat{\boldsymbol{y}}_\ell) - \mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(\ell\delta)) \big\| . \tag{3.8.5}$$

Note that $A_0 = B_0 = 0$. In the next lemma we bound the above quantities:

Lemma 3.8.2. *For $\beta < 1/2$ and $\mathsf{T} > 0$, there exists a constant $C = C(\beta) < \infty$, and a deterministic non-negative sequence $\alpha(N)$ with $\lim_{N\to\infty} \alpha(N) = 0$ such that the following holds with probability $1 - o_N(1)$. For every $\ell \geq 0$, $\delta \in (0,1)$ such that $\ell\delta \leq \mathsf{T}$,*

$$A_\ell \leq C e^{C\ell\delta} \ell\delta \big(\rho\sqrt{\ell\delta} + \sqrt{\delta}\big) + \alpha(N) , \tag{3.8.6}$$

$$B_\ell \leq C e^{C\ell\delta} \ell\delta \big(\rho\sqrt{\ell\delta} + \sqrt{\delta}\big) + C\rho\sqrt{\ell\delta} + \alpha(N) . \tag{3.8.7}$$

Proof. Throughout the proof, we denote by $\alpha(N)$ a deterministic non-negative sequence $\alpha(N)$ with $\lim_{N\to\infty} \alpha(N) = 0$, which can change from line to line. Also, C will denote a generic constant that may depend on $\beta, \mathsf{T}, K_{\mathsf{AMP}}$.

The proof proceeds by induction on ℓ. As the base case is trivial, we assume the result holds for all $j \leq \ell$ and we prove it for $\ell + 1$. We first claim that with probability $1 - o_N(1)$,

$$A_{\ell+1} \leq A_\ell + \delta B_\ell + C\delta^{3/2}. \tag{3.8.8}$$

Indeed, using (3.8.1) we find

$$\begin{aligned}
A_{\ell+1} - A_\ell &\leq n^{-1/2} \int_{\ell\delta}^{(\ell+1)\delta} \big\| \hat{\boldsymbol{m}}(\boldsymbol{A}, \hat{\boldsymbol{y}}_\ell) - \mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(t)) \big\| \, \mathrm{d}t \\
&\leq \delta n^{-1/2} \Big(\big\| \hat{\boldsymbol{m}}(\boldsymbol{A}, \hat{\boldsymbol{y}}_\ell) - \mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(\ell\delta)) \big\| + \sup_{t \in [\ell\delta, (\ell+1)\delta]} \big\| \mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(t)) - \mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(\ell\delta)) \big\| \Big) \\
&\leq \delta B_\ell + \delta n^{-1/2} \cdot \sup_{t \in [\ell\delta, (\ell+1)\delta]} \big\| \mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(t)) - \mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(\ell\delta)) \big\| \\
&\leq \delta B_\ell + C(\beta)\delta^{3/2} + \alpha(N) ,
\end{aligned}$$

where the last line holds with high probability by Lemma 3.6.7 and Eq. (3.6.10) of Lemma 3.6.3. Using this bound together with the inductive hypothesis on A_ℓ and B_ℓ, we obtain

$$\begin{aligned}
A_{\ell+1} &\leq C e^{C(\ell+1)\delta} \ell\delta (\rho\sqrt{\ell\delta} + \sqrt{\delta}) + C\rho\delta\sqrt{\ell\delta} + C\delta^{3/2} + \alpha(N) \\
&\leq C e^{C(\ell+1)\delta} (\ell+1)\delta (\rho + \sqrt{\delta}) + \alpha(N) .
\end{aligned}$$

This implies Eq. (3.8.6) for $\ell + 1$.

We next show that Eq. (3.8.7) holds with ℓ replaced by $\ell+1$. By the bound (3.8.6) for $\ell+1$, taking $\delta \le \delta(\beta,\varepsilon,K_{\mathsf{AMP}},\mathsf{T})$ and $\rho \in (0,\rho_0)$ $\rho = \rho(\beta,\varepsilon,K_{\mathsf{AMP}},\mathsf{T})$ ensures that

$$A_{\ell+1} \le \frac{c\sqrt{\varepsilon\ell\delta}}{200K_{\mathsf{AMP}}6^{K_{\mathsf{AMP}}}}\,,$$

where ε can be chosen an arbitrarily small constant. So by Lemma 3.8.1, we have with probability $1-o_N(1)$,

$$\begin{aligned}\big\|\operatorname{arctanh}(\mathsf{AMP}(\boldsymbol{A},\boldsymbol{y}((\ell+1)\delta);K_{\mathsf{AMP}})) - \operatorname{arctanh}(\mathsf{AMP}(\boldsymbol{A},\hat{\boldsymbol{y}}_{\ell+1};K_{\mathsf{AMP}}))\big\|_2 &\le K_{\mathsf{AMP}}6^{K_{\mathsf{AMP}}}A_{\ell+1}\sqrt{N}\\ &\le \frac{c\sqrt{\varepsilon\ell\delta N}}{200}\,.\end{aligned}$$

By choosing $\varepsilon \le \varepsilon_0(\beta,\mathsf{T})$, we obtain that Lemma 3.7.1, part 4 applies. We thus find

$$\|\hat{\boldsymbol{m}}(\boldsymbol{A},\hat{\boldsymbol{y}}_{\ell+1}) - \mathbf{m}_*(\boldsymbol{A},\hat{\boldsymbol{y}}_{\ell+1})\| \le \rho\sqrt{\ell\delta N}\,.$$

Using parts 3 and 2 respectively of Lemma 3.7.1 on the other terms below, by triangle inequality we obtain (writing for simplicity $q_\ell := q_*(\beta,\ell\delta)$)

$$\begin{aligned}\|\hat{\boldsymbol{m}}(\boldsymbol{A},\hat{\boldsymbol{y}}_{\ell+1}) - \mathbf{m}(\boldsymbol{A},\boldsymbol{y}((\ell+1)\delta))\| &\le \|\hat{\boldsymbol{m}}(\boldsymbol{A},\hat{\boldsymbol{y}}_{\ell+1}) - \mathbf{m}_*(\boldsymbol{A},\hat{\boldsymbol{y}}_{\ell+1};q_{\ell+1})\|\\ &\quad + \|\mathbf{m}_*(\boldsymbol{A},\hat{\boldsymbol{y}}_{\ell+1};q_{\ell+1}) - \mathbf{m}_*(\boldsymbol{A},\boldsymbol{y}((\ell+1)\delta);q_{\ell+1})\|\\ &\quad + \|\mathbf{m}_*(\boldsymbol{A},\boldsymbol{y}((\ell+1)\delta);q_{\ell+1}) - \mathbf{m}(\boldsymbol{A},\boldsymbol{y}((\ell+1)\delta))\|\\ &\le \big(\rho\sqrt{\ell\delta} + c^{-1}A_{\ell+1} + \rho\sqrt{\ell\delta} + \alpha(N)\big)\sqrt{N}\,.\end{aligned} \tag{3.8.9}$$

In other words with probability $1-o_N(1)$,

$$B_{\ell+1} \le c^{-1}A_{\ell+1} + 2\rho\sqrt{\ell\delta} + \alpha(N)\,.$$

Using this together with the bound (3.8.6) for $\ell+1$ verifies the inductive step for (3.8.7) and concludes the proof. □

Finally we show that standard randomized rounding is continuous in $W_{2,N}$.

Lemma 3.8.3. *Suppose probability distributions μ_1,μ_2 on $[-1,1]^N$ are given. Sample $\mathbf{m}_1 \sim \mu_1$ and $\mathbf{m}_2 \sim \mu_2$ and let $\boldsymbol{x}_1,\boldsymbol{x}_2 \in \{-1,+1\}^N$ be standard randomized roundings, respectively of $\mathbf{m}_1$ and $\mathbf{m}_2$. (Namely, the coordinates of $\boldsymbol{x}_i$ are conditionally independent given $\mathbf{m}_i$, with $\mathbb{E}[\boldsymbol{x}_i|\mathbf{m}_i] = \mathbf{m}_i$.) Then*

$$W_{2,N}(\mathcal{L}(\boldsymbol{x}_1),\mathcal{L}(\boldsymbol{x}_2)) \le 2\sqrt{W_{2,N}(\mu_1,\mu_2)}\,.$$

Proof. Let $(\mathbf{m}_1,\mathbf{m}_2)$ be distributed according to a $W_{2,N}$-optimal coupling between μ_1,μ_2. Couple

the roundings $\boldsymbol{x}_1, \boldsymbol{x}_2$ by choosing i.i.d. uniform random variables $u_i \sim \mathsf{Unif}([0,1])$ for $i \in [n]$, and for $(i,j) \in [n] \times \{1,2\}$ setting

$$(\boldsymbol{x}_j)_i = \begin{cases} +1, & \text{if } u \le \frac{1+(\mathbf{m}_j)_i}{2}\,, \\ -1, & \text{else.} \end{cases}$$

Then it is not difficult to see that

$$\begin{aligned} \frac{1}{N}\,\mathbb{E}\left[\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2 \;|(\mathbf{m}_1, \mathbf{m}_2)\right] &= \frac{2}{N}\sum_{i=1}^{N} |(\mathbf{m}_1)_i - (\mathbf{m}_2)_i| \\ &\le 2\sqrt{\frac{1}{N}\|\mathbf{m}_1 - \mathbf{m}_2\|^2}. \end{aligned}$$

Averaging over the choice of $(\mathbf{m}_1, \mathbf{m}_2)$ implies the result. □

Proof of Theorem 37. Set $\ell = L = \mathsf{T}/\delta$ and $\rho = \sqrt{\delta}$ in Eq. (3.8.7). With all laws $\mathcal{L}(\,\cdot\,)$ conditional on $\boldsymbol{A}$ below, we find

$$\begin{aligned} \mathbb{E}\, W_{2,N}(\mu_{\boldsymbol{A}}, \mathcal{L}(\hat{\boldsymbol{m}}(\boldsymbol{A}, \hat{\boldsymbol{y}}_L))) &\le \mathbb{E}\, W_{2,N}(\mu_{\boldsymbol{A}}, \mathcal{L}(\mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(\mathsf{T})))) + \mathbb{E}\, W_{2,N}(\mathcal{L}(\mathbf{m}(\boldsymbol{A}, \boldsymbol{y}(\mathsf{T})))), \mathcal{L}(\hat{\boldsymbol{m}}(\boldsymbol{A}, \hat{\boldsymbol{y}}_L))) \\ &\le \mathsf{T}^{-1/2} + C(\beta, \mathsf{T})\sqrt{\delta} + o_N(1). \end{aligned}$$

Here the first term was bounded by Eq. (3.3.4) in Section 3.3 and the second by Eq. (3.8.7). Taking T sufficiently large, δ sufficiently small, and N sufficiently large, we may obtain

$$\mathbb{E}\, W_{2,N}\big(\mu_{\boldsymbol{A}}, \mathcal{L}(\hat{\mathbf{m}}_{\mathsf{NGD}}(\boldsymbol{A}, \hat{\boldsymbol{y}}_L))\big) \le \frac{\varepsilon^2}{4}$$

for any desired $\varepsilon > 0$. Applying Lemma 3.8.3 shows that

$$\mathbb{E}\, W_{2,N}(\mu_{\boldsymbol{A}}, \boldsymbol{x}^{\mathrm{alg}}) \le \varepsilon\,.$$

The Markov inequality now implies that (7.2.6) holds with probability $1 - o_N(1)$ as desired. □

3.9 Algorithmic stability and disorder chaos

In this section we prove Theorem 10 establishing that our sampling algorithm, Algorithm 2 is stable. Next, we prove that the Sherrington-Kirkpatrick measure $\mu_{\boldsymbol{A},\beta}$ exhibits W_2-disorder chaos for $\beta > 1$, proving Theorem 11 and deduce that no stable algorithm can sample in normalized W_2 distance for $\beta > 1$, see Theorem 12.

3.9.1 Algorithmic stability: Proof of Theorem 10

Recall Definition 3.2.1, defining sampling algorithms as measurable functions $\mathsf{ALG}_N : (\boldsymbol{A}, \beta, \omega) \mapsto \mathsf{ALG}_N(\boldsymbol{A}, \beta, \omega) \in [-1,1]^N$ where $\beta \geq 0$ and ω is an independent random variable taking values in some probability space.

Remark 3.9.1. In light of Lemma 3.8.3, we can always turn a stable sampling algorithm ALG with codomain $[-1,1]^N$ into a stable sampling algorithm with binary output:

$$\widetilde{\mathsf{ALG}}_N(\boldsymbol{A}, \beta, \widetilde{\omega}) \in \{-1,+1\}^N .$$

Indeed this is achieved by standard randomized rounding, i.e., drawing a (conditionally independent) random binary value with mean $\big(\widetilde{\mathsf{ALG}}(\boldsymbol{A}, \beta, \widetilde{\omega})\big)_i$ for each coordinate $1 \leq i \leq N$.

Recall the definition of the interpolating family $(\boldsymbol{A}_s)_{s\in[0,1]}$ whereby $\boldsymbol{A}_0, \boldsymbol{A}_1 \sim \mathsf{GOE}(N)$ i.i.d. and

$$\boldsymbol{A}_s = \sqrt{1-s^2}\boldsymbol{A}_0 + s\boldsymbol{A}_1 , \quad s \in [0,1] , \tag{3.9.1}$$

We take $\mu_{\boldsymbol{A}_s,\beta}(\boldsymbol{x}) \propto \exp\big\{(\beta/2)\langle \boldsymbol{x}, \boldsymbol{A}_s\, \boldsymbol{x}\rangle\big\}$ to be the corresponding Gibbs measure.

We start with the following simple estimate.

Lemma 3.9.1. *There exists an absolute constant $C > 0$ such that*

$$\inf_{s\in(0,1)} \mathbb{P}\Big(\big\|\boldsymbol{A}_0\boldsymbol{u} - \boldsymbol{A}_s\boldsymbol{v}\big\| \leq C(\|\boldsymbol{u}-\boldsymbol{v}\| + s\sqrt{N}) , \quad \forall\, \boldsymbol{u}, \boldsymbol{v} \in [-1,1]^N \Big) = 1 - o_N(1) . \tag{3.9.2}$$

Proof. We write

$$\begin{aligned} \big\|\boldsymbol{A}_0\boldsymbol{u} - \boldsymbol{A}_s\boldsymbol{v}\big\| &\leq \big\|\boldsymbol{A}_0\boldsymbol{u} - \boldsymbol{A}_0\boldsymbol{v}\big\| + \big\|\boldsymbol{A}_0\boldsymbol{v} - \boldsymbol{A}_s\boldsymbol{v}\big\| \\ &\leq \|\boldsymbol{A}_0\|_{\mathrm{op}} \big\|\boldsymbol{u}-\boldsymbol{v}\big\| + \big\|(1-\sqrt{1-s^2})\boldsymbol{A}_0 - s\boldsymbol{A}_1\big\|_{\mathrm{op}} \|\boldsymbol{v}\| . \end{aligned}$$

We note that $(1-\sqrt{1-s^2})\boldsymbol{A}_0 - s\boldsymbol{A}_1 \stackrel{\mathrm{d}}{=} \sqrt{2(1-\sqrt{1-s^2})}\boldsymbol{A}_0$ and $\sqrt{2(1-\sqrt{1-s^2})} \sim s$ for small s and this quantity is bounded above by a constant for any $s \in [0,1]$. The result follows since $\|\boldsymbol{A}_0\|_{\mathrm{op}} \leq 2.1$ with probability $1 - o_N(1)$. □

Proposition 3.9.2. *Suppose an algorithm ALG is given by an iterative procedure*

$$\begin{aligned} &\boldsymbol{z}^{k+1} = G_k\left((\boldsymbol{z}^j, \beta\boldsymbol{A}\mathbf{m}^j, \boldsymbol{A}\mathbf{m}^j, \beta^2\mathbf{m}^j, \boldsymbol{w}^j)_{0\leq j\leq k}\right), \quad 0 \leq k \leq K-1, \\ &\mathbf{m}^k = \rho_k(\boldsymbol{z}^k), \quad 0 \leq k \leq K-1, \\ &\mathsf{ALG}_N(\boldsymbol{A}, \beta, \omega) := \mathbf{m}^K , \end{aligned}$$

where the sequence $\omega = (\boldsymbol{w}^0, \dots, \boldsymbol{w}^{K-1}) \in (\mathbb{R}^N)^K$*, the initialization* $\boldsymbol{z}^0 \in \mathbb{R}^N$*, and* $\boldsymbol{A}$ *are mutually independent, and the functions* $G_k : (\mathbb{R}^N)^{5k+5} \to \mathbb{R}^N$ *and* $\rho_k : \mathbb{R}^N \to [-1,1]^N$ *are* L_0*-Lipschitz for* $L_0 \geq 0$ *an* N*-independent constant. Then* ALG *is both disorder-stable and temperature-stable.*

Proof. Let us generate iterates $\boldsymbol{z}^k = \boldsymbol{z}^k(\boldsymbol{A}_0, \beta)$ and $\widetilde{\boldsymbol{z}}^k = \boldsymbol{z}^k(\boldsymbol{A}_s, \widetilde{\beta})$ for $0 \leq k \leq K$ using the same initialization $\boldsymbol{z}^0 = \widetilde{\boldsymbol{z}}^0$ and external randomness $\omega = (\boldsymbol{w}^0, \dots, \boldsymbol{w}^{K-1})$, but with different Hamiltonians and inverse temperatures. Similarly let $\mathbf{m}^k = \rho_k(\boldsymbol{z}^k)$ and $\widetilde{\mathbf{m}}^k = \rho_k(\widetilde{\boldsymbol{z}}^k)$. We will allow C to vary from line to line in the proof below.

First by Lemma 3.9.1, with probability $1 - o_N(1)$,

$$\begin{aligned}\|\beta \boldsymbol{A}_0 \mathbf{m}^k - \widetilde{\beta} \boldsymbol{A}_s \widetilde{\mathbf{m}}^k\| &\leq \|\beta \boldsymbol{A}_0 \mathbf{m}^k - \beta \boldsymbol{A}_s \widetilde{\mathbf{m}}^k\| + \|\beta \boldsymbol{A}_s \widetilde{\mathbf{m}}^k - \widetilde{\beta} \boldsymbol{A}_s \widetilde{\mathbf{m}}^k\| \\ &\leq C\beta \|\mathbf{m}^k - \widetilde{\mathbf{m}}^k\| + C\beta s\sqrt{N} + |\beta - \widetilde{\beta}| \cdot \|\boldsymbol{A}_s \widetilde{\mathbf{m}}^k\| \\ &\leq C(\|\mathbf{m}^k - \widetilde{\mathbf{m}}^k\| + s\sqrt{N} + |\beta - \widetilde{\beta}|\sqrt{N})\,.\end{aligned}$$

Similarly as long as $\widetilde{\beta} \leq 2\beta$ so that $|\beta^2 - \widetilde{\beta}^2| \leq 3\beta|\beta - \widetilde{\beta}|$, we have

$$\begin{aligned}\|\beta^2 \mathbf{m}^k - \widetilde{\beta}^2 \widetilde{\mathbf{m}}^k\| &\leq \|\beta^2 \mathbf{m}^k - \beta^2 \widetilde{\mathbf{m}}^k\| + \|\beta^2 \widetilde{\mathbf{m}}^k - \widetilde{\beta}^2 \widetilde{\mathbf{m}}^k\| \\ &\leq \beta^2 \|\mathbf{m}^k - \widetilde{\mathbf{m}}^k\| + 3\beta|\beta - \widetilde{\beta}|\sqrt{N}.\end{aligned}$$

It follows that the error sequence

$$A_k = \frac{1}{\sqrt{N}} \max_{j \leq k} \|\boldsymbol{z}^{j+1}(\boldsymbol{A}_0, \beta) - \boldsymbol{z}^{j+1}(\boldsymbol{A}_s, \widetilde{\beta})\|$$

satisfies with probability $1 - o_N(1)$ the recursion

$$\begin{aligned}A_{k+1} &\leq L_0 k^{1/2} C(A_k + s + |\beta - \widetilde{\beta}|)\,, \\ A_0 &= 0\,,\end{aligned}$$

for a suitable $C = C(\beta)$. It follows that with probability $1 - o_N(1)$,

$$A_K \leq \sum_{k=1}^{K} (L_0 k^{1/2} C)^k (s + |\beta - \widetilde{\beta}|) \leq K(L_0 K C)^K (s + |\beta - \widetilde{\beta}|)\,. \tag{3.9.3}$$

Since $\|\mathbf{m}^K(\boldsymbol{A}_0) - \mathbf{m}^K(\boldsymbol{A}_s)\| \leq 2\sqrt{N}$ almost surely, we obtain for any $\eta > 0$

$$n^{-1}\,\mathbb{E}\left[\left\|\mathbf{m}^K(\boldsymbol{A}_0) - \mathbf{m}^K(\boldsymbol{A}_s)\right\|^2\right] \leq \left(L_0 K (L_0 K C)^K (s + |\beta - \widetilde{\beta}|)\right)^2 + \eta$$

if $N \geq N_0(\eta)$ is large enough so that Eq. (3.9.3) holds with probability at least $1 - \frac{\eta}{4}$. The stability of the algorithm follows. □

Proof of Theorem 10. We show that Algorithm 2 with N-independent parameters $(\beta, \eta, K_{\mathsf{AMP}}, K_{\mathsf{NGD}}, L, \delta)$ is of the form in Proposition 3.9.2 for a constant $L_0 = L_0(\beta, \eta, K_{\mathsf{AMP}}, K_{\mathsf{NGD}}, L, \delta)$. Indeed note that the algorithm goes through L iterations, indexed by $\ell \in \{0, \dots, L-1\}$.

During each of these iterations, two loops are run (here we modify the notation introduced in Algorithm 1 and Algorithm 2, to account for the dependence on ℓ, and to get closer to the notation of Proposition 3.9.2):

1. The AMP loop, whereby, for $k = 0, \cdots, K_{\mathsf{AMP}} - 1$,

$$\hat{\boldsymbol{m}}^{\ell,k} = \tanh(\boldsymbol{z}^{\ell,k}), \qquad \mathsf{b}(\hat{\boldsymbol{m}}^{\ell,k}) = \frac{\beta^2}{N}\sum_{i=1}^{N} \tanh'(z_i^{\ell,k})\,, \tag{3.9.4}$$

$$\boldsymbol{z}^{\ell,k+1} = \beta \boldsymbol{A}\hat{\boldsymbol{m}}^{\ell,k} + \hat{\boldsymbol{y}}_\ell - \mathsf{b}(\boldsymbol{z}^{\ell,k})\,\hat{\boldsymbol{m}}^{\ell,k-1}\,. \tag{3.9.5}$$

   (Here $\tanh'(x)$ denotes the first derivative of $\tanh(x)$.)

2. The NGD loop, whereby, for $k = K_{\mathsf{AMP}}, \cdots, K_{\mathsf{AMP}} + K_{\mathsf{NGD}} - 1$, setting $q_\ell = q_{K_{\mathsf{AMP}}}(\beta, t = \ell\delta)$,

$$\hat{\boldsymbol{m}}^{\ell,k} = \tanh(\boldsymbol{z}^{\ell,k})\,, \tag{3.9.6}$$

$$\boldsymbol{z}^{\ell,k+1} = \boldsymbol{z}^{\ell,k} + \eta\big[\beta \boldsymbol{A}\hat{\boldsymbol{m}}^{\ell,k} + \boldsymbol{y}_\ell - \boldsymbol{z}^{\ell,k} - \beta^2\,(1-q_\ell)\,\mathbf{m}^{\ell,k}\big]\,. \tag{3.9.7}$$

Further, recalling line 6 of Algorithm 2, $\hat{\boldsymbol{y}}_\ell$ is updated via

$$\hat{\boldsymbol{y}}_{\ell+1} = \hat{\boldsymbol{y}}_\ell + \hat{\boldsymbol{m}}^{\ell, K_{\mathsf{AMP}}+K_{\mathsf{NGD}}}\,\delta + \sqrt{\delta}\,\boldsymbol{w}_{\ell+1}\,. \tag{3.9.8}$$

These updates take the same form as in Proposition 3.9.2, with iterations indexed by (ℓ, k), $\omega = (\boldsymbol{w}_\ell)_{\ell \le L}$, $\rho_{\ell,k}(\boldsymbol{z}) = \tanh(\boldsymbol{z})$ for all ℓ, k, and

$$G_{\ell,k}\Big((\boldsymbol{z}^{\ell',j}, \beta\boldsymbol{A}\hat{\boldsymbol{m}}^{\ell',j}, \boldsymbol{A}\hat{\boldsymbol{m}}^{\ell',j}, \beta^2\hat{\boldsymbol{m}}^{\ell',j}, \boldsymbol{w}_{\ell'})_{\ell',j}\Big) = \beta\boldsymbol{A}\hat{\boldsymbol{m}}^{\ell,k} + \hat{\boldsymbol{y}}_\ell - \mathsf{b}(\boldsymbol{z}^{\ell,k})\hat{\boldsymbol{m}}^{\ell,k-1}\,, \quad 0 \le k \le K_{\mathsf{AMP}} - 1\,, \tag{3.9.9}$$

$$\begin{aligned} G_{\ell,k}\Big((\boldsymbol{z}^{\ell',j}, \beta\boldsymbol{A}\hat{\boldsymbol{m}}^{\ell',j}, \boldsymbol{A}\hat{\boldsymbol{m}}^{\ell',j}, \beta^2\hat{\boldsymbol{m}}^{\ell',j}, \boldsymbol{w}_{\ell'})_{\ell',j}\Big) = \\ = \boldsymbol{z}^{\ell,k} + \eta\big[\beta\boldsymbol{A}\hat{\boldsymbol{m}}^{\ell,k} + \boldsymbol{y}_\ell - \boldsymbol{z}^{\ell,k} - \beta^2\,(1-q_\ell)\,\mathbf{m}^{\ell,k}\big]\,, \quad K_{\mathsf{AMP}} \le k \le K_{\mathsf{AMP}} + K_{\mathsf{NGD}} - 1\,. \end{aligned} \tag{3.9.10}$$

Notice that these functions depend on previous iterates both explicitly, as noted, and implicitly through $\hat{\boldsymbol{y}}_\ell$. By summing up Eq. (3.9.8), we obtain

$$\hat{\boldsymbol{y}}_\ell = \sum_{j=0}^{\ell-1} \hat{\boldsymbol{m}}^{j, K_{\mathsf{AMP}}+K_{\mathsf{NGD}}}\,\delta + \sqrt{\delta}\sum_{j=1}^{\ell} \boldsymbol{w}_{\ell+1}\,, \tag{3.9.11}$$

which is Lipschitz in the previous iterates $(\mathbf{m}^{j,k})_{j\le \ell-1, k<K_{\mathsf{AMP}}+K_{\mathsf{NGD}}}$. Since both (3.9.9) and (3.9.10) depend linearly on $\hat{\boldsymbol{y}}_\ell$ (with N-independent coefficients), it is sufficient to consider the explicit dependence on previous iterates of $G_{\ell,k}$. Namely, it is sufficient to control the Lipschitz modulus of the following functions

$$\widetilde{G}_{\ell,k}\left(\boldsymbol{z}^{\ell,k}, \beta \boldsymbol{A}\hat{\boldsymbol{m}}^{\ell,k}, \hat{\boldsymbol{m}}^{\ell,k-1}\right) = \beta \boldsymbol{A}\hat{\boldsymbol{m}}^{\ell,k} - \mathsf{b}(\boldsymbol{z}^{\ell,k})\hat{\boldsymbol{m}}^{\ell,k-1}\,, \qquad k < K_{\mathsf{AMP}} \tag{3.9.12}$$

$$\widetilde{G}_{\ell,k}\left(\boldsymbol{z}^{\ell,k}, \beta \boldsymbol{A}\mathbf{m}^{\ell,k}, \beta^2\mathbf{m}^{\ell,k}\right) = \boldsymbol{z}^{\ell,k} + \eta\big[\beta \boldsymbol{A}\hat{\boldsymbol{m}}^{\ell,k} - \boldsymbol{z}^{\ell,k} - \beta^2\,(1-q_\ell)\,\hat{\boldsymbol{m}}^{\ell,k}\big]\,, \qquad k > K_{\mathsf{AMP}}\,. \tag{3.9.13}$$

Consider first Eq. (3.9.12). Since $|\tanh''(x)| \le 2$ for all $x \in \mathbb{R}$, it follows that

$$|\mathsf{b}(\boldsymbol{z}) - \mathsf{b}(\widetilde{\boldsymbol{z}})| \le \frac{2\beta^2}{N}\sum_{i=1}^N |z_i - \tilde{z}_i| \le \frac{2\beta^2}{\sqrt{N}}\|\boldsymbol{z} - \widetilde{\boldsymbol{z}}\|_2.$$

Therefore, that for any $(\boldsymbol{u}, \boldsymbol{v}, \beta, \widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}}, \widetilde{\beta})$ (Noting explicitly the dependence of b upon β):

$$\begin{aligned}
\|\mathsf{b}_\beta(\boldsymbol{u})\tanh(\boldsymbol{v}) - \mathsf{b}_{\widetilde{\beta}}(\widetilde{\boldsymbol{u}})\tanh(\widetilde{\boldsymbol{v}})\| &\le \|\mathsf{b}_\beta(\boldsymbol{u})\tanh(\boldsymbol{v}) - \mathsf{b}_\beta(\widetilde{\boldsymbol{u}})\tanh(\boldsymbol{v})\| + \|\mathsf{b}_\beta(\widetilde{\boldsymbol{u}})\tanh(\boldsymbol{v}) - \mathsf{b}_{\widetilde{\beta}}(\widetilde{\boldsymbol{u}})\tanh(\widetilde{\boldsymbol{v}})\| \\
&\le \frac{2\beta^2}{\sqrt{N}}\|\boldsymbol{u} - \widetilde{\boldsymbol{u}}\| \cdot \|\tanh(\boldsymbol{v})\| + \Big(\frac{1}{N}\sum_{i=1}^N \tanh'(\tilde{u}_i)\Big)\|\beta^2\tanh(\boldsymbol{v}) - \widetilde{\beta}^2\tanh(\widetilde{\boldsymbol{v}})\| \\
&\le 2\beta^2\|\boldsymbol{u} - \widetilde{\boldsymbol{u}}\| + \|\beta^2\tanh(\boldsymbol{v}) - \widetilde{\beta}^2\tanh(\widetilde{\boldsymbol{v}})\|.
\end{aligned}$$

Using this bound implies that the function $\widetilde{G}$ of Eq. (3.9.12) satisfies the Lipschitz assumption of Proposition 3.9.2.

Consider next Eq. (3.9.13). Since this function is linear in its arguments, with coefficients independent of N, it follows that it satisfies Lipschitz assumption of Proposition 3.9.2. This completes the proof. □

3.9.2 Hardness for stable algorithms: Proof of Theorems 11 and 12

Before proving Theorem 11 and Theorem 12 we recall a known result about disorder chaos, already stated in Eq. (3.2.11). Draw $\boldsymbol{x}^0 \sim \mu_{\boldsymbol{A},\beta}$ independently of $\boldsymbol{x}^s \sim \mu_{\boldsymbol{A}_s,\beta}$, and denote by $\mu_{\boldsymbol{A},\beta}^{(0,s)} := \mu_{\boldsymbol{A},\beta} \otimes \mu_{\boldsymbol{A}^s,\beta}$ their joint distribution. Then [Cha14, Theorem 1.11] implies that, for all $\beta \in (0,\infty)$,

$$\lim_{s\to 0}\lim_{N\to\infty} \mathbb{E}\,\mu_{\boldsymbol{A},\beta}^{(0,s)}\Big\{\Big(\frac{1}{N}\langle \boldsymbol{x}^0, \boldsymbol{x}^s\rangle\Big)^2\Big\} = 0\,. \tag{3.9.14}$$

The following simple estimate will be used in our proof.

Lemma 3.9.3. *Recall that $\mathscr{P}(\{-1,+1\}^N)$ denotes the space of probability distributions over $\{-1,+1\}^N$,*

and let the function $f : \mathscr{P}(\{-1,+1\}^N)^2 \to \mathbb{R}$ *be defined as*

$$f(\mu,\mu') = \mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{x}')\sim\mu\otimes\mu'} \Big\{ \frac{1}{N} |\langle \boldsymbol{x}, \boldsymbol{x}'\rangle| \Big\} . \tag{3.9.15}$$

Then, for all $\mu_1,\mu_2,\nu_1,\nu_2 \in \mathscr{P}(\{-1,+1\}^N)$*, we have*

$$\big| f(\mu_1,\nu_1) - f(\mu_2,\nu_2) \big| \le W_{2,N}(\mu_1,\mu_2) + W_{2,N}(\nu_1,\nu_2) .$$

Proof. Let the vector pairs $(\boldsymbol{x}^{\mu_1},\boldsymbol{x}^{\mu_2})$ and $(\boldsymbol{x}^{\nu_1},\boldsymbol{x}^{\nu_2})$ be independently drawn from the optimal $W_{2,N}$-couplings of the pairs (μ_1,μ_2) and (ν_1,ν_2), respectively. Then we have:

$$\begin{aligned}
\Big| \mathbb{E}\big[|\langle \boldsymbol{x}^{\mu_1}, \boldsymbol{x}^{\nu_1}\rangle|\big] - \mathbb{E}\big[|\langle \boldsymbol{x}^{\mu_2}, \boldsymbol{x}^{\nu_2}\rangle|\big] \Big| &\le \Big| \mathbb{E}\big[|\langle \boldsymbol{x}^{\mu_1}, \boldsymbol{x}^{\nu_1}\rangle| - |\langle \boldsymbol{x}^{\mu_2}, \boldsymbol{x}^{\nu_1}\rangle|\big] \Big| + \Big| \mathbb{E}\big[|\langle \boldsymbol{x}^{\mu_2}, \boldsymbol{x}^{\nu_1}\rangle| - |\langle \boldsymbol{x}^{\mu_2}, \boldsymbol{x}^{\nu_2}\rangle|\big] \Big| \\
&\le \sqrt{N} \Big(\mathbb{E} \big\| \boldsymbol{x}^{\mu_1} - \boldsymbol{x}^{\mu_2} \big\| + \mathbb{E} \big\| \boldsymbol{x}^{\nu_1} - \boldsymbol{x}^{\nu_2} \big\| \Big) \\
&\le \sqrt{N} \Big(\mathbb{E} \Big[\big\| \boldsymbol{x}^{\mu_1} - \boldsymbol{x}^{\mu_2} \big\|^2 \Big]^{1/2} + \mathbb{E} \Big[\big\| \boldsymbol{x}^{\nu_1} - \boldsymbol{x}^{\nu_2} \big\|^2 \Big]^{1/2} \Big) ,
\end{aligned}$$

where the second inequality follows from the fact that $\boldsymbol{x} \mapsto |\langle \boldsymbol{v}, \boldsymbol{x}\rangle|$ is Lipschitz continuous with Lipschitz constant $\|\boldsymbol{v}\|_2$. □

We are now in position to prove Theorem 11.

Proof of Theorem 11. Using the notations of the last lemma Eq. (3.9.14) implies that for all $s \in (0,1]$,

$$\lim_{N\to\infty} \mathbb{E}\, f(\mu_{\boldsymbol{A}_s,\beta}, \mu_{\boldsymbol{A}_0,\beta}) = 0 . \tag{3.9.16}$$

Therefore, Theorem 11 follows from Lemma 3.9.3 if we can show that $f(\mu_{\boldsymbol{A}_0,\beta}, \mu_{\boldsymbol{A}_0,\beta})$ remains bounded away from zero. This is in turn a well-known consequence of the Parisi formula, as we recall below.

Define the free energy density of the SK model as

$$F_N(\beta) = \frac{1}{N} \mathbb{E} \log \Big\{ \sum_{\boldsymbol{x}\in\{-1,+1\}^N} e^{\beta\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle/2} \Big\} . \tag{3.9.17}$$

The free energy F_N is convex in β and one obtains by Gaussian integration parts that

$$\frac{\mathrm{d}}{\mathrm{d}\beta} F_N(\beta) = \frac{\beta}{2} \Big(1 - \mathbb{E}\, \mu_{\boldsymbol{A}_0,\beta} \otimes \mu_{\boldsymbol{A}_0,\beta} \Big\{ \Big(\frac{1}{N} \langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle \Big)^2 \Big\} \Big) . \tag{3.9.18}$$

Moreover, the limit of $F_N(\beta)$ for large N is known to exist for all $\beta > 0$ and its value is given by

the Parisi formula [Tal06d]:

$$\lim_{N\to\infty} F_N(\beta) = \inf_{\zeta\in\mathscr{P}([0,1])} \mathsf{P}_\beta(\zeta)\,, \tag{3.9.19}$$

where $\mathscr{P}([0,1])$ denotes the set of Borel probability measures supported on $[0,1]$, and P_β is the Parisi functional at inverse temperature β; see for instance [Tal06d] or [Pan13b, Chapter 3] for definitions.

The following properties are known:

1. A unique minimizer $\zeta^*_\beta \in \mathscr{P}([0,1])$ of P_β exists for all β [AC15].
2. If $\beta > 1$, then ζ^*_β is not an atom on 0: $\zeta^*_\beta \neq \delta_0$. This follows from Toninelli's theorem [Ton02] that $\limsup_{N\to\infty} F_N(\beta) \le \log 2 + \beta^2/4 - \varepsilon(\beta)$ for some continuous $\varepsilon(\beta)$, with $\varepsilon(\beta) > 0$ when $\beta > 1$.
3. The function $\beta \mapsto \mathsf{P}_\beta(\zeta^*_\beta)$ is convex and differentiable at all $\beta > 0$, and

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\mathsf{P}_\beta(\zeta^*_\beta) = \frac{\beta}{2}\Big(1 - \int q^2\zeta^*_\beta(\mathrm{d}q)\Big)\,. \tag{3.9.20}$$

   See for instance [Pan13b, Theorem 3.7] or [Tal06c, Theorem 1.2] for a proof.

The convexity of F_N implies that for almost all $\beta > 0$, $\lim_{N\to\infty} F_N'(\beta) = \frac{\mathrm{d}}{\mathrm{d}\beta}\mathsf{P}_\beta(\zeta^*_\beta)$. Using Eq. (3.9.18) and Eq. (3.9.20) we obtain

$$\lim_{N\to\infty} \frac{\beta}{2}\Big(1 - \mathbb{E}\,\mu_{\boldsymbol{A}_0,\beta}\otimes\mu_{\boldsymbol{A}_0,\beta}\Big\{\Big(\frac{1}{N}\langle \boldsymbol{x}_1,\boldsymbol{x}_2\rangle\Big)^2\Big)\Big\} = \frac{\beta}{2}\Big(1 - \int q^2\zeta^*_\beta(\mathrm{d}q)\Big) < \frac{\beta}{2} - \varepsilon(\beta)\,, \tag{3.9.21}$$

where the last inequality holds for almost all $\beta > 1$ by Property 2 above. Since the both sides are non-decreasing and the right hand side is continuous, the inequality holds for all β. This is equivalent to

$$\lim_{N\to\infty} \mathbb{E}\, f(\mu_{\boldsymbol{A}_0,\beta}, \mu_{\boldsymbol{A}_0,\beta}) > 0\,. \tag{3.9.22}$$

Now, using Eq. (3.9.16) and Eq. (3.9.22), together with the continuity of f (Lemma 3.9.3) implies the claim of the theorem. □

We next prove that Theorem 12 is an immediate consequence of Theorem 11.

Proof of Theorem 12. Fix $s \in (0,1)$ and $\mu^{\mathrm{alg}}_{\boldsymbol{A}_s,\beta}$ be the law of $\mathsf{ALG}_N(\boldsymbol{A}_s,\beta,\omega)$ conditional on $\boldsymbol{A}_s$. By the triangle inequality,

$$W_{2,N}(\mu_{\boldsymbol{A}_s,\beta,s}, \mu_{\boldsymbol{A}_0,\beta}) \le W_{2,N}(\mu_{\boldsymbol{A}_s,\beta}, \mu^{\mathrm{alg}}_{\boldsymbol{A}_s,\beta}) + W_{2,N}(\mu^{\mathrm{alg}}_{\boldsymbol{A}_s,\beta}, \mu^{\mathrm{alg}}_{\boldsymbol{A}_0,\beta}) + W_{2,N}(\mu^{\mathrm{alg}}_{\boldsymbol{A}_0,\beta}, \mu_{\boldsymbol{A}_0,\beta,0})\,.$$

Taking expectations over $\boldsymbol{A}$ and $\boldsymbol{A}_s$, we have $\mathbb{E}\left[W_{2,N}(\mu_{\boldsymbol{A}_s,\beta}, \mu^{\mathrm{alg}}_{\boldsymbol{A}_s,\beta})\right] = \mathbb{E}\left[W_{2,N}(\mu^{\mathrm{alg}}_{\boldsymbol{A}_0,\beta}, \mu_{\boldsymbol{A}_0,\beta})\right]$. Further, by stability of the algorithm, $\mathbb{E}\left[W_{2,N}(\mu^{\mathrm{alg}}_{\boldsymbol{A}_s,\beta}, \mu^{\mathrm{alg}}_{\boldsymbol{A}_0,\beta})\right] \to 0$ when $N\to\infty$ followed by $s\to 0$.

Therefore, using Theorem 11 and choosing s sufficiently small, we obtain

$$\liminf_{N\to\infty} \mathbb{E}\left[W_{2,N}(\mu^{\rm alg}_{\boldsymbol{A}_0,\beta},\ \mu_{\boldsymbol{A}_0,\beta})\right] \geq W_* > 0\,.$$

□

3.10 Convergence analysis of Natural Gradient Descent

The main objective of this appendix is to prove Lemma 3.7.1, which we will do in Section 3.10.2, after some technical preparations in Section 3.10.1.

3.10.1 Technical preliminaries

Definition 3.10.1. *Let $Q \subseteq (-1,1)^N$ be a convex set. We say that a twice differentiable function $\boldsymbol{F} : Q \to \mathbb{R}$ is relatively c-strongly convex if it satisfies*

$$\nabla^2 \boldsymbol{F}(\mathbf{m}) \succeq c\boldsymbol{D}(\mathbf{m}) \quad \forall\, \mathbf{m} \in Q\,. \tag{3.10.1}$$

We say it is relatively C-smooth if it satisfies

$$\nabla^2 \boldsymbol{F}(\mathbf{m}) \preceq C\boldsymbol{D}(\mathbf{m}) \quad \forall\, \mathbf{m} \in Q\,. \tag{3.10.2}$$

As $\boldsymbol{D}(\mathbf{m}) = \nabla^2(-\boldsymbol{h}(\mathbf{m})) \succeq \boldsymbol{I}_N$, it follows that (3.10.1) implies ordinary c-strong convexity in Euclidean norm. The next proposition connects relative strong convexity with the Bregman divergence introduced in Eq. 3.7.10.

Proposition 3.10.2 (Proposition 1.1 in [LFN18])**.** *A twice differentiable function $\boldsymbol{F} : Q \to \mathbb{R}$ is relatively c-strongly convex if and only if*

$$\boldsymbol{F}(\mathbf{m}) \geq \boldsymbol{F}(\boldsymbol{n}) + \langle \nabla \boldsymbol{F}(\boldsymbol{n}), \mathbf{m} - \boldsymbol{n}\rangle + cD_{-\boldsymbol{h}}(\mathbf{m}, \boldsymbol{n}), \qquad \forall \mathbf{m}, \boldsymbol{n} \in Q\,. \tag{3.10.3}$$

Lemma 3.10.3. *For $\mathbf{m}, \boldsymbol{n} \in (-1,1)^N$,*

$$D_{-\boldsymbol{h}}(\mathbf{m}, \boldsymbol{n}) \geq \frac{\|\mathbf{m} - \boldsymbol{n}\|_2^2}{2}\,, \tag{3.10.4}$$

$$D_{-\boldsymbol{h}}(\mathbf{m}, \boldsymbol{n}) \leq 10N\left(1 + \frac{\|\operatorname{arctanh}(\boldsymbol{n})\|_2}{\sqrt{N}}\right)\,, \tag{3.10.5}$$

$$D_{-\boldsymbol{h}}(\mathbf{m}, \boldsymbol{n}) \leq \|\operatorname{arctanh}(\mathbf{m}) - \operatorname{arctanh}(\boldsymbol{n})\|_2^2\,. \tag{3.10.6}$$

Proof. Observe that $h''(x) = -1/(1-x^2) \leq -1$ for all $x \in (-1,1)$ with equality if and only if $x = 0$.

Therefore

$$\begin{aligned} D_{-\boldsymbol{h}}(\mathbf{m}, \boldsymbol{n}) &= \sum_{i=1}^{N} \int_{m_i}^{N_i} (x - m_i)(-h''(x)) \, \mathrm{d}x \\ &= \sum_{i=1}^{N} \frac{(N_i - m_i)^2}{2} \, . \end{aligned}$$

This proves Eq. (3.10.4).

Next, Eq. (3.10.5) follows from Eq. (3.7.10) and the fact that the binary entropy $h : \mathbb{R} \to \mathbb{R}$ is uniformly bounded.

Finally Eq. (3.10.6) follows from

$$\begin{aligned} D_{-\boldsymbol{h}}(\mathbf{m}, \boldsymbol{n}) &\le \langle \nabla h(\boldsymbol{n}) - \nabla h(\mathbf{m}), \mathbf{m} - \boldsymbol{n} \rangle \\ &= \big\langle \operatorname{arctanh}(\mathbf{m}) - \operatorname{arctanh}(\boldsymbol{n}), \mathbf{m} - \boldsymbol{n} \big\rangle \\ &\le \big\| \operatorname{arctanh}(\mathbf{m}) - \operatorname{arctanh}(\boldsymbol{n}) \big\|_2^2 \, . \end{aligned}$$

Here in the last step we used that $\tanh(\cdot)$ is 1-Lipschitz. □

Lemma 3.10.4. *If $\boldsymbol{F} : Q \to \mathbb{R}$ is relatively c-strongly convex for some convex set $Q \subseteq (-1, 1)^N$, and $\nabla \mathcal{F}(\mathbf{m}_*) = 0$ for $\mathbf{m}_* \in Q$, it follows that*

$$\boldsymbol{F}(\mathbf{m}) - \boldsymbol{F}(\mathbf{m}_*) \ge \frac{c \|\mathbf{m} - \mathbf{m}_*\|_2^2}{2} \, .$$

for all $\mathbf{m} \in Q$.

Proof. Using (3.10.3) and (3.10.4), and observing that $\nabla \mathcal{F}(\mathbf{m}_*) = 0$, we obtain

$$\frac{\mathcal{F}(\mathbf{m}) - \mathcal{F}(\mathbf{m}_*)}{\|\mathbf{m} - \mathbf{m}_*\|_2^2} \ge \frac{\mathcal{F}(\mathbf{m}) - \mathcal{F}(\mathbf{m}_*)}{2 \cdot D_{-\boldsymbol{h}}(\mathbf{m}, \mathbf{m}_*)} \ge \frac{c}{2} \, .$$

□

Lemma 3.10.5. *Suppose $\boldsymbol{F} : Q_* \to \mathbb{R}$ is c-strongly convex in the convex set $Q_* := B(\boldsymbol{m}, \rho) \cap (-1, 1)^N$. If $\boldsymbol{x}_* \in \partial Q_*$, $x_{*,k} = +1$ (respectively, $x_{*,k} = -1$) and $|x_j| < 1$ for all $j \in [n] \setminus \{k\}$, then $\lim_{t \to 0+} \partial_{x_k} \boldsymbol{F}(\boldsymbol{x}_* - t\boldsymbol{e}_k) = +\infty$ (respectively $\lim_{t \to 0+} \partial_{x_k} \boldsymbol{F}(\boldsymbol{x}_* + t\boldsymbol{e}_k) = -\infty$.)*

Proof. Consider the case $x_k = +1$ (as the case $x_k = -1$ follows by symmetry.) Then there exists

$t_0 > 0$ such that $\boldsymbol{x}_* - t\boldsymbol{e}_k \in Q_*$ for all $t \in (0, t_0]$. Let $\boldsymbol{x}(s) := \boldsymbol{x}_* - (t_0 - s)\boldsymbol{e}_k$, $s \in [0, t_0)$. Then

$$\begin{aligned}
\partial_{x_k} F(\boldsymbol{x}(s)) &= \partial_{x_k} F(\boldsymbol{x}(0)) + \int_0^s \partial_{x_k}^2 F(\boldsymbol{x}(u))\,\mathrm{d}u \\
&= \partial_{x_k} F(\boldsymbol{x}(0)) + \int_0^s \langle \boldsymbol{e}_k, \nabla^2 F(\boldsymbol{x}(u))\boldsymbol{e}_k\rangle\,\mathrm{d}u \\
&\geq \partial_{x_k} F(\boldsymbol{x}(0)) + c\int_0^s (1 - x_k(u)^2)^{-1}\,\mathrm{d}u \\
&\geq \partial_{x_k} F(\boldsymbol{x}(0)) + c\int_0^s (1 - (1 - t_0 + u)^2)^{-1}\,\mathrm{d}u,.
\end{aligned}$$

The last integral diverges as $s \uparrow t_0$, thus proving the claim. □

Lemma 3.10.6. *Suppose $\boldsymbol{F} : Q \to \mathbb{R}$ is c-strongly convex for a convex set $Q \subseteq (-1,1)^N$. Moreover suppose that*

$$\|\nabla \mathcal{F}(\mathbf{m})\| \leq c\sqrt{\varepsilon N}$$

for some $\mathbf{m} \in Q$ with

$$B\left(\mathbf{m}, 2\sqrt{\varepsilon N}\right) \cap (-1,1)^N \subseteq Q\,.$$

Then there exists a unique $\mathbf{m}_* \in B\left(\mathbf{m}, 2\sqrt{\varepsilon N}\right) \cap (-1,1)^N$ satisfying $\nabla \boldsymbol{F}(\mathbf{m}_*) = 0$, which is in fact a global minimizer of $\boldsymbol{F}$ on Q. Moreover

$$\boldsymbol{F}(\mathbf{m}) - \boldsymbol{F}(\mathbf{m}_*) \leq 2c\varepsilon N\,. \tag{3.10.7}$$

Proof. Let $Q_{\leq} := \{\boldsymbol{x} \in Q : \boldsymbol{F}(\boldsymbol{x}) \leq \boldsymbol{F}(\mathbf{m})\}$. Then, for any $\boldsymbol{x} \in Q_0$, we have

$$\begin{aligned}
0 &\geq \boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\mathbf{m}) \\
&\geq -c\sqrt{\varepsilon N}\|\boldsymbol{x} - \mathbf{m}\|_2 + cD_{-\boldsymbol{h}}(\boldsymbol{x}; \mathbf{m}) \\
&\geq -c\sqrt{\varepsilon N}\|\boldsymbol{x} - \mathbf{m}\|_2 + \frac{c}{2}\|\boldsymbol{x} - \mathbf{m}\|_2^2\,.
\end{aligned}$$

Hence $Q_{\leq} \subseteq Q_* := B\left(\mathbf{m}, \sqrt{\varepsilon N}\right) \cap (-1,1)^N$, $Q_* \subseteq Q$. By continuity three cases are possible: (i) The minimum of $\boldsymbol{F}$ is achieved in the interior of $Q_{\leq}$; (ii) The minimum is achieved along a sequence $(\boldsymbol{x}_i)_{i\geq 0}$, $\|\boldsymbol{x}_i\|_\infty \to 1$; (iii) the minimum is achieved at $\mathbf{m}_* \neq \mathbf{m}$ such that $\boldsymbol{F}(\mathbf{m}_*) = \boldsymbol{F}(\mathbf{m})$. Case (iii) cannot hold by strong convexity, and case (ii) cannot hold by Lemma 3.10.5.

Uniqueness of $\mathbf{m}_*$ follows by strong convexity, and $\nabla \boldsymbol{F}(\mathbf{m}_*) = 0$ by differentiability. Finally

$$\boldsymbol{F}(\mathbf{m}) - \boldsymbol{F}(\mathbf{m}_*) \leq \|\nabla \boldsymbol{F}(\mathbf{m})\| \cdot \|\mathbf{m} - \mathbf{m}_*\| \leq 2c\varepsilon N\,.$$

□

Lemma 3.10.7. *Suppose $\boldsymbol{F} : Q \to \mathbb{R}$ is relatively c-strongly convex. Let $\mathbf{m}_*$ be a local minimum of $\boldsymbol{F}$ belonging to the interior of Q, and suppose that $B\left(\mathbf{m}_*, 2\sqrt{\varepsilon N}\right) \cap (-1,1)^N \subseteq Q$. Consider for $\boldsymbol{y} \in \mathbb{R}^N$ the function*

$$\boldsymbol{F}_{\boldsymbol{y}}(\mathbf{m}) = \boldsymbol{F}(\mathbf{m}) - \langle \boldsymbol{y}, \mathbf{m} \rangle.$$

Then $\boldsymbol{F}_{\boldsymbol{y}}$ is relatively c-strongly convex on Q for any $\boldsymbol{y} \in \mathbb{R}^N$. If $\|\boldsymbol{y}\| \le (c/2)\sqrt{\varepsilon N}$, then $\boldsymbol{F}_{\boldsymbol{y}}$ has a unique stationary point and minimizer $\mathbf{m}_*(\boldsymbol{y}) \in Q$. Moreover if $\|\boldsymbol{y}\|, \|\hat{\boldsymbol{y}}\| \le \frac{c\sqrt{\varepsilon N}}{2}$ then

$$\|\mathbf{m}_*(\boldsymbol{y}) - \mathbf{m}_*(\hat{\boldsymbol{y}})\| \le \frac{\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|}{c}. \tag{3.10.8}$$

Proof. The relative c-strong convexity of $\boldsymbol{F}_{\boldsymbol{y}}$ is clear as the Hessian of $\boldsymbol{F}_{\boldsymbol{y}}$ does not depend on $\boldsymbol{y}$. For $\|\boldsymbol{y}\| \le (c/2)\sqrt{\varepsilon N}$, because

$$\|\nabla \boldsymbol{F}_{\boldsymbol{y}}(\mathbf{m}_*)\| = \|\boldsymbol{y}\| \le \frac{c\sqrt{\varepsilon N}}{2} \quad \text{and} \quad B\left(\mathbf{m}_*, \sqrt{\varepsilon N}\right) \cap (-1,1)^N \subseteq Q\,,$$

Lemma 3.10.6 implies the existence of a unique minimizer

$$\mathbf{m}_*(\boldsymbol{y}) \in B\left(\mathbf{m}_*, \sqrt{\varepsilon N}\right) \cap (-1,1)^N \subseteq Q\,.$$

If $\|\hat{\boldsymbol{y}}\| \le (c/2)\sqrt{\varepsilon N}$ also holds, $\boldsymbol{F}_{\hat{\boldsymbol{y}}}$ is c-strongly convex on

$$B\left(\mathbf{m}_*(\hat{\boldsymbol{y}}), \sqrt{\varepsilon N}\right) \cap (-1,1)^N \subseteq B\left(\mathbf{m}_*, 2\sqrt{\varepsilon N}\right) \cap (-1,1)^N \subseteq Q.$$

Moreover since $\|\boldsymbol{y} - \hat{\boldsymbol{y}}\| \le c\sqrt{\varepsilon N}$, we obtain

$$\|\nabla \boldsymbol{F}_{\hat{\boldsymbol{y}}}(\mathbf{m}_*(\boldsymbol{y}))\| = \|\boldsymbol{y} - \hat{\boldsymbol{y}}\| = c\sqrt{\varepsilon' N}\,,$$

for $\varepsilon' = \frac{\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2}{c^2 N} \le \varepsilon$. Therefore the conditions of Lemma 3.10.6 are satisfied with $(\boldsymbol{F}_{\hat{\boldsymbol{y}}}, \mathbf{m}_*(\boldsymbol{y}), \varepsilon')$ in place of $(\boldsymbol{F}, \mathbf{m}, \varepsilon)$. Equation (3.10.8) now follows since

$$\|\mathbf{m}_*(\boldsymbol{y}) - \mathbf{m}_*(\hat{\boldsymbol{y}})\| \le \sqrt{\varepsilon' N} = \frac{\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|}{c}\,.$$

□

We now analyze the convergence of Algorithm 3 from a good initialization.

Lemma 3.10.8. *Suppose $\boldsymbol{F}(\cdot) = \mathscr{F}_{\text{TAP}}(\,\cdot\,; \boldsymbol{y}, q_K(\beta, t))$ has a local minimum at $\mathbf{m}_*$ and is relatively c-strongly-convex on $B(\mathbf{m}_*, \sqrt{\varepsilon N}) \cap (-1,1)^N$, and also C-relatively smooth on $(-1,1)^N$. Suppose*

$$\hat{\boldsymbol{m}}^0 \in B\left(\mathbf{m}_*, \sqrt{\varepsilon N}\right) \cap (-1,1)^N \tag{3.10.9}$$

satisfies

$$\boldsymbol{F}(\hat{\boldsymbol{m}}^0) < \boldsymbol{F}(\mathbf{m}_*) + \frac{c\varepsilon N}{8}. \tag{3.10.10}$$

Then there exist constants $\eta_0, C' > 0$ depending only on (C, c, ε) such that the following holds. If Algorithm 3 is initialized at $\hat{\boldsymbol{m}}^0$ with learning rate $\eta = 1/L \in (0, \eta_0)$, then, for every $K \geq 1$

$$\boldsymbol{F}(\hat{\boldsymbol{m}}^K) \leq \boldsymbol{F}(\mathbf{m}_*) + C'N\left(1 + \frac{\|\operatorname{arctanh}(\hat{\boldsymbol{m}}^0)\|_2}{\sqrt{N}}\right)(1 - c\eta)^K, \tag{3.10.11}$$

$$\|\hat{\boldsymbol{m}}^K - \mathbf{m}_*\|_2 \leq C'\sqrt{N}\left(1 + \frac{\|\operatorname{arctanh}(\hat{\boldsymbol{m}}^0)\|_2}{\sqrt{N}}\right)(1 - c\eta)^{K/2}. \tag{3.10.12}$$

Proof. Recall Eq. (3.7.11), which we copy here for the reader's convenience:

$$\hat{\boldsymbol{m}}^{i+1} = \arg\min_{\boldsymbol{x} \in (-1,1)^N} \left\langle \nabla \boldsymbol{F}(\hat{\boldsymbol{m}}^i), \boldsymbol{x} - \hat{\boldsymbol{m}}^i \right\rangle + L \cdot D_{-\boldsymbol{h}}(\boldsymbol{x}, \hat{\boldsymbol{m}}^i). \tag{3.10.13}$$

If $\eta_0 \leq \frac{1}{2C}$ then [LFN18, Lemma 3.1] applied to the linear (hence convex) function $\langle \nabla \boldsymbol{F}(\hat{\boldsymbol{m}}^i), \cdot \rangle$ states that for all $\mathbf{m} \in (-1,1)^N$,

$$\langle \nabla \boldsymbol{F}(\hat{\boldsymbol{m}}^i), \hat{\boldsymbol{m}}^{i+1} \rangle + L D_{-\boldsymbol{h}}(\hat{\boldsymbol{m}}^{i+1}, \hat{\boldsymbol{m}}^i) + L D_{-\boldsymbol{h}}(\mathbf{m}, \hat{\boldsymbol{m}}^{i+1}) \leq \langle \nabla \boldsymbol{F}(\hat{\boldsymbol{m}}^i), \mathbf{m} \rangle + L D_{-\boldsymbol{h}}(\mathbf{m}, \hat{\boldsymbol{m}}^i). \tag{3.10.14}$$

Moreover the global relative smoothness shown in (3.7.8) implies that for $\mathbf{m}, \mathbf{m}' \in (-1,1)^N$,

$$\boldsymbol{F}(\mathbf{m}) \leq \boldsymbol{F}(\mathbf{m}') + \langle \nabla \boldsymbol{F}(\mathbf{m}'), \mathbf{m} - \mathbf{m}' \rangle + C \cdot D_{-\boldsymbol{h}}(\mathbf{m}, \mathbf{m}'). \tag{3.10.15}$$

Combining Eqs. (3.10.14) and (3.10.15) yields

$$\begin{aligned} \boldsymbol{F}(\hat{\boldsymbol{m}}^{i+1}) &\leq \boldsymbol{F}(\hat{\boldsymbol{m}}^i) + \langle \nabla \boldsymbol{F}(\hat{\boldsymbol{m}}^i), \hat{\boldsymbol{m}}^{i+1} - \hat{\boldsymbol{m}}^i \rangle + L D_{-\boldsymbol{h}}(\hat{\boldsymbol{m}}^{i+1}, \hat{\boldsymbol{m}}^i) \\ &\leq \boldsymbol{F}(\hat{\boldsymbol{m}}^i) + \langle \nabla \boldsymbol{F}(\hat{\boldsymbol{m}}^i), \mathbf{m} - \hat{\boldsymbol{m}}^i \rangle + L D_{-\boldsymbol{h}}(\mathbf{m}, \hat{\boldsymbol{m}}^i) - L D_{-\boldsymbol{h}}(\mathbf{m}, \hat{\boldsymbol{m}}^{i+1}). \end{aligned} \tag{3.10.16}$$

Setting $\mathbf{m} = \hat{\boldsymbol{m}}^i$, we find

$$\boldsymbol{F}(\hat{\boldsymbol{m}}^{i+1}) \leq \boldsymbol{F}(\hat{\boldsymbol{m}}^i), \quad \forall\, i \in [K].$$

We next prove by induction that for each $i \geq 1$,

$$\boldsymbol{F}(\hat{\boldsymbol{m}}^i) < \boldsymbol{F}(\mathbf{m}_*) + \frac{c\varepsilon N}{8}, \qquad \|\hat{\boldsymbol{m}}^i - \mathbf{m}_*\| < \sqrt{\varepsilon N}. \tag{3.10.17}$$

The base case $i = 0$ holds by assumption. Suppose (3.10.17) holds for i. It follows that

$$\boldsymbol{F}(\hat{\boldsymbol{m}}^{i+1}) \leq \boldsymbol{F}(\hat{\boldsymbol{m}}^i) \leq \boldsymbol{F}(\mathbf{m}_*) + \frac{c\varepsilon N}{8}.$$

In fact, local c-strong convexity

$$\nabla^2 \boldsymbol{F}(\mathbf{m}) \succeq c\boldsymbol{D}(\mathbf{m}) \succeq c\boldsymbol{I}_N, \quad \mathbf{m} \in B(\mathbf{m}_*, \sqrt{\varepsilon N}) \cap (-1,1)^N$$

implies $\hat{\boldsymbol{m}}^i$ is even closer to $\mathbf{m}_*$ than required by (3.10.17):

$$\|\hat{\boldsymbol{m}}^i - \mathbf{m}_*\|_2 \le \sqrt{\frac{\boldsymbol{F}(\hat{\boldsymbol{m}}^i) - \boldsymbol{F}(\mathbf{m}_*)}{c}} \le \frac{\sqrt{\varepsilon N}}{2}.$$

Next we bound the movement from a single NGD step. Comparing values of (3.10.13) at $\hat{\boldsymbol{m}}^i$ and the minimizer $\hat{\boldsymbol{m}}^{i+1}$ implies

$$\langle \nabla \boldsymbol{F}(\hat{\boldsymbol{m}}^i), \hat{\boldsymbol{m}}^{i+1} - \hat{\boldsymbol{m}}^i \rangle + L D_{-\boldsymbol{h}}(\hat{\boldsymbol{m}}^{i+1}, \hat{\boldsymbol{m}}^i) \le 0. \tag{3.10.18}$$

From definition of Bregman divergence and the fact that (on the high probability event $\|\boldsymbol{A}\|_{\rm op} \le 3$) $\|\nabla \boldsymbol{F} + \nabla \boldsymbol{h}\|_2 \le C\sqrt{N}$ (thanks to the special form of $\boldsymbol{F}(\,\cdot\,) = \mathscr{F}_{\rm TAP}(\,\cdot\,; \boldsymbol{y}, q_K(\beta, t))$,

$$\begin{aligned}|\langle \nabla \boldsymbol{F}(\hat{\boldsymbol{m}}^i), \hat{\boldsymbol{m}}^{i+1} - \hat{\boldsymbol{m}}^i \rangle + D_{-\boldsymbol{h}}(\hat{\boldsymbol{m}}^{i+1}, \hat{\boldsymbol{m}}^i)| &= |\langle \nabla \boldsymbol{F}(\hat{\boldsymbol{m}}^i) + \nabla \boldsymbol{h}(\hat{\boldsymbol{m}}^i), \hat{\boldsymbol{m}}^{i+1} - \hat{\boldsymbol{m}}^i \rangle - \boldsymbol{h}(\hat{\boldsymbol{m}}^{i+1}) + \boldsymbol{h}(\hat{\boldsymbol{m}}^i)| \\ &\le C_1 N \left(1 + \frac{\|\hat{\boldsymbol{m}}^{i+1} - \hat{\boldsymbol{m}}^i\|}{\sqrt{N}}\right).\end{aligned}$$

Moreover assuming $L > 1$, (3.10.4) implies

$$(L-1) D_{-\boldsymbol{h}}(\hat{\boldsymbol{m}}^{i+1}, \hat{\boldsymbol{m}}^i) \ge \frac{L-1}{2} \|\hat{\boldsymbol{m}}^{i+1} - \hat{\boldsymbol{m}}^i\|^2.$$

Substituting the previous two displays into (3.10.18) yields

$$0 \ge \frac{L-1}{2} \|\hat{\boldsymbol{m}}^{i+1} - \hat{\boldsymbol{m}}^i\|^2 - C_2\sqrt{N} \|\hat{\boldsymbol{m}}^{i+1} - \hat{\boldsymbol{m}}^i\|_2 - C_2 n$$

and so

$$\|\hat{\boldsymbol{m}}^{i+1} - \hat{\boldsymbol{m}}^i\|_2 \le \frac{C_3\sqrt{N}}{\sqrt{L-1}}.$$

Taking L large enough, it follows that

$$\|\hat{\boldsymbol{m}}^{i+1} - \mathbf{m}_*\| \le \|\hat{\boldsymbol{m}}^{i+1} - \hat{\boldsymbol{m}}^i\|_2 + \|\hat{\boldsymbol{m}}^i - \mathbf{m}_*\|_2 \le \sqrt{\varepsilon N}.$$

This completes the inductive proof of Eq. (3.10.17), which we now use to analyze convergence of Algorithm 3. Indeed from the first part of (3.10.17), the local relative strong convexity of $\boldsymbol{F}$ implies

$$\boldsymbol{F}(\hat{\boldsymbol{m}}^i) + \langle \nabla \boldsymbol{F}(\hat{\boldsymbol{m}}^i), \mathbf{m}_* - \hat{\boldsymbol{m}}^i \rangle \le \boldsymbol{F}(\mathbf{m}_*) - c D_{-\boldsymbol{h}}(\mathbf{m}_*, \hat{\boldsymbol{m}}^i), \qquad \forall\, i \in [K].$$

Setting $\mathbf{m} = \mathbf{m}_*$ in (3.10.16) and combining yields

$$\boldsymbol{F}(\hat{\boldsymbol{m}}^{i+1}) \le \boldsymbol{F}(\mathbf{m}_*) + (L-c)D_{-\boldsymbol{h}}(\mathbf{m}_*, \hat{\boldsymbol{m}}^i) - LD_{-\boldsymbol{h}}(\mathbf{m}_*, \hat{\boldsymbol{m}}^{i+1}).$$

Multiplying by $\left(\frac{L}{L-c}\right)^{i+1}$ and summing over i gives

$$\sum_{i=0}^{K-1}\left(\frac{L}{L-c}\right)^{i+1}\boldsymbol{F}(\hat{\boldsymbol{m}}^{i+1}) \le \sum_{i=0}^{K-1}\left(\frac{L}{L-c}\right)^{i+1}\boldsymbol{F}(\mathbf{m}_*) + LD_{-\boldsymbol{h}}(\mathbf{m}_*, \hat{\boldsymbol{m}}^0).$$

Since the values $\boldsymbol{F}(\hat{\boldsymbol{m}}^i)$ are decreasing, we find

$$\begin{aligned}\boldsymbol{F}(\hat{\boldsymbol{m}}^K) &\le \boldsymbol{F}(\mathbf{m}_*) + L\left(\sum_{i=0}^{K-1}\left(\frac{L}{L-c}\right)^{i+1}\right)^{-1} D_{-\boldsymbol{h}}(\mathbf{m}_*, \hat{\boldsymbol{m}}^0)\\ &\le \boldsymbol{F}(\mathbf{m}_*) + L\left(1-c\eta\right)^K D_{-\boldsymbol{h}}(\mathbf{m}_*, \hat{\boldsymbol{m}}^0).\end{aligned}$$

Using Eq. (3.10.5) together with the last display proves Eq. (3.10.11).

It was shown above by induction that $\hat{\boldsymbol{m}}^K$ is in a c-strongly convex neighborhood of $\mathbf{m}_*$. Using strong convexity in Euclidean norm yields

$$\|\hat{\boldsymbol{m}}^k - \mathbf{m}_*\| \le \sqrt{\frac{\boldsymbol{F}(\hat{\boldsymbol{m}}^K) - \boldsymbol{F}(\mathbf{m}_*)}{c}}$$

and so (3.10.12) follows as well. □

Lemma 3.10.9. *Assume $\|\boldsymbol{A}\|_{op} \le 3$. For any $\mathbf{m}, \boldsymbol{n} \in (-1,1)^N$, and $\boldsymbol{y}, \hat{\boldsymbol{y}} \in \mathbb{R}^N$, and $q \in [0,1]$:*

$$\|\nabla\mathscr{F}_{\text{TAP}}(\mathbf{m}, \boldsymbol{y}, q) - \nabla\mathscr{F}_{\text{TAP}}(\boldsymbol{n}, \hat{\boldsymbol{y}}, q)\| \le (4\beta^2+4)\|\operatorname{arctanh}(\mathbf{m}) - \operatorname{arctanh}(\boldsymbol{n})\| + \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|. \quad (3.10.19)$$

Proof. The inequality (3.10.19) follows with the smaller constant factor $\beta^2 + 3\beta + 1 \le 4\beta^2 + 4$ using (3.7.4) and the fact that $\tanh(\cdot)$ is 1-Lipschitz.

□

3.10.2 Proof of Lemma 3.7.1

We split the proof into four parts.

Proof of Lemma 3.7.1, Part 1. Fix $c = (1/4) - (\beta/2) > 0$. Lemma 3.7.2 implies that for $K_{\text{AMP}} =$

$K_{\mathsf{AMP}}(\beta, \mathsf{T}, \varepsilon)$ sufficiently large, we have with probability $1 - o_N(1)$

$$\|\nabla \mathscr{F}_{\rm TAP}(\hat{\boldsymbol{m}}^{\mathsf{AMP}}; \boldsymbol{y}, q_*)\| \le \frac{c\sqrt{\varepsilon t N}}{4}\,, \tag{3.10.20}$$
$$\hat{\boldsymbol{m}}^{\mathsf{AMP}} := \mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}(t); K_{\mathsf{AMP}}), \quad q_* := q_*(\beta, t)\,.$$

Therefore, if $\|\boldsymbol{y}(t) - \hat{\boldsymbol{y}}\| \le (c\sqrt{\varepsilon t N})/4$ then

$$\|\nabla \mathscr{F}_{\rm TAP}(\hat{\boldsymbol{m}}^{\mathsf{AMP}}; \hat{\boldsymbol{y}}, q_*)\| \le \|\nabla \mathscr{F}_{\rm TAP}(\hat{\boldsymbol{m}}^{\mathsf{AMP}}; \boldsymbol{y}(t), q_*)\| + \|\boldsymbol{y} - \hat{\boldsymbol{y}}\| \le \frac{c}{2}\sqrt{\varepsilon t}$$

Moreover Lemma 3.7.3 implies that there exist ε_0, $c > 0$ such that for all $\varepsilon \in (0, \varepsilon_0)$,

$$\nabla^2 \mathscr{F}_{\rm TAP}(\mathbf{m}; \hat{\boldsymbol{y}}, q_*) = \nabla^2 \mathscr{F}_{\rm TAP}(\mathbf{m}; \boldsymbol{y}(t), q_*) \succeq c\boldsymbol{D}(\mathbf{m}), \qquad \forall\, \mathbf{m} \in B\left(\hat{\boldsymbol{m}}^{\mathsf{AMP}}, \sqrt{\varepsilon t N}\right) \cap (-1, 1)^N.$$

Using $\varepsilon t/4$ in place of ε in Lemma 3.10.6, it follows that there exists a local minimum

$$\mathbf{m}_*(\boldsymbol{A}, \hat{\boldsymbol{y}}; q_*) \in B\left(\hat{\boldsymbol{m}}^{\mathsf{AMP}}, \frac{\sqrt{\varepsilon t N}}{2}\right) \cap (-1, 1)^N$$

of $\mathscr{F}_{\rm TAP}(\,\cdot\,, \hat{\boldsymbol{y}}; q_*)$ which is also the unique stationary point in $B\left(\hat{\boldsymbol{m}}^{\mathsf{AMP}}, (1/2)\sqrt{\varepsilon t N}\right) \cap (-1, 1)^N$.

We next claim that, for any $K > K_{\mathsf{AMP}}$, with probability $1 - o_N(1)$, this local minimum is also the unique stationary point in $B\left(\mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}(t); k), (1/2)\sqrt{\varepsilon t N}\right) \cap (-1, 1)^N$. Indeed for K_{AMP} sufficiently large (writing for simplicity $\boldsymbol{y} = \boldsymbol{y}(t)$):

$$\begin{aligned}\operatorname*{p-lim}_{N\to\infty} \sup_{k_1, k_2 \in [K_{\mathsf{AMP}}, K]} \|\mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; k_1) - \mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; k_2)\|^2 &= \sup_{k_1, k_2 \in [k_{\rm alg}, K]} \operatorname*{p-lim}_{N\to\infty} \|\mathsf{AMP}_\beta(\boldsymbol{A}, \boldsymbol{y}; k_1) - \mathsf{AMP}_\beta(\boldsymbol{A}, \boldsymbol{y}; k_2)\|^2 \\ &\le N \cdot \sup_{k_1, k_2 \ge K_{\mathsf{AMP}}} |q_{k_1}(\beta, t) - q_{k_2}(\beta, t)|.\end{aligned}$$

From Eq. (3.6.8), by eventually increasing K_{AMP}, we have

$$\sup_{k_1, k_2 \ge K_{\mathsf{AMP}}} |q_{k_1}(\beta, t) - q_{k_2}(\beta, t)| \le \frac{\varepsilon t}{16}.$$

For such K_{AMP}, with probability $1 - o_N(1)$, all $k \in [K_{\mathsf{AMP}}, K]$ satisfy

$$\begin{aligned}\|\mathbf{m}_*(\boldsymbol{A}, \boldsymbol{y}; q_{K_{\mathsf{AMP}}}) - \mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; k)\| &\le \|\mathbf{m}_*(\boldsymbol{A}, \boldsymbol{y}; q_{\mathsf{AMP}}) - \mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; K_{\mathsf{AMP}})\| \\ &\quad + \|\mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; k) - \mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; K_{\mathsf{AMP}})\| \\ &\le \frac{\sqrt{\varepsilon t N}}{2} + \sqrt{\frac{\varepsilon t N}{4}} \\ &\le \frac{3}{4}\sqrt{\varepsilon t N}.\end{aligned}$$

Let

$$S(k,\rho) := B\left(\mathsf{AMP}_\beta(\boldsymbol{A},\boldsymbol{y};k),\rho\right) \cap (-1,1)^N\,, \quad \rho_{N,t} := \sqrt{\varepsilon n t}$$

Recall that $\mathbf{m}_*(\boldsymbol{A},\boldsymbol{y};q_*)$ is the unique stationary point of $\mathscr{F}_{\mathrm{TAP}}(\,\cdot\,;\boldsymbol{y},q_*)$ in $S(K_{\mathsf{AMP}},\rho_{N,t})$. By the above, it is also a stationary point in $S(k,\rho_{N,t})$, for $k \in [K_{\mathsf{AMP}},K]$. Repeating the same argument as before, there is only one stationary point inside $S(k,\rho_{N,t})$, hence this must coincide with $\mathbf{m}_*(\boldsymbol{A},\boldsymbol{y};q_*)$. □

Proof of Lemma 3.7.1, Part 2. Because K_{AMP} is large depending on δ_0, Lemma 3.7.2 implies that with probability $1-o_N(1)$,

$$\|\nabla\mathscr{F}_{\mathrm{TAP}}(\mathsf{AMP}(\boldsymbol{A},\boldsymbol{y};K_{\mathsf{AMP}}),\boldsymbol{y};q_*)\| \le \frac{c\delta_0\sqrt{tN}}{4}.$$

Using $\frac{\delta_0\sqrt{t}}{4}$ in place of ε in Lemma 3.10.6, it follows that the local minimizer $\mathbf{m}_*(\boldsymbol{A},\boldsymbol{y};q_*)$ of $\mathscr{F}_{\mathrm{TAP}}(\,\cdot\,;\boldsymbol{y},q_*)$ satisfies

$$\|\mathsf{AMP}(\boldsymbol{A},\boldsymbol{y};K_{\mathsf{AMP}}) - \mathbf{m}_*(\boldsymbol{A},\boldsymbol{y};q_*)\| \le \frac{\delta_0\sqrt{tN}}{2}.$$

Since K is sufficiently large depending on δ_0, Lemma implies that with probability $1-o_N(1)$,

$$\|\mathbf{m}(\boldsymbol{A},\boldsymbol{y}) - \mathsf{AMP}(\boldsymbol{A},\boldsymbol{y};K_{\mathsf{AMP}})\| \le \frac{\delta_0\sqrt{tN}}{2}.$$

Combining, we obtain that with probability $1-o_N(1)$,

$$\begin{aligned}\|\mathbf{m}(\boldsymbol{A},\boldsymbol{y}) - \mathbf{m}_*(\boldsymbol{A},\boldsymbol{y};q_*)\| &\le \|\mathbf{m}(\boldsymbol{A},\boldsymbol{y}) - \mathsf{AMP}(\boldsymbol{A},\boldsymbol{y};K_{\mathsf{AMP}})\| + \|\mathsf{AMP}(\boldsymbol{A},\boldsymbol{y};K_{\mathsf{AMP}}) - \mathbf{m}_*(\boldsymbol{A},\boldsymbol{y};q_*)\| \\ &\le \delta_0\sqrt{tN}.\end{aligned}$$

□

Proof of Lemma 3.7.1, Part 3. The result is immediate from (3.10.8). □

Proof of Lemma 3.7.1, Part 4. We apply Lemma 3.10.8 with $\boldsymbol{F}(\,\cdot\,) = \mathscr{F}_{\mathrm{TAP}}(\,\cdot\,;\hat{\boldsymbol{y}},q_*)$ and $\mathbf{m}_* = \mathbf{m}_*(\boldsymbol{A},\hat{\boldsymbol{y}};q_*))$ (with $q_* = q_*(\beta,t)$). We need to check that assumptions (3.10.9), (3.10.10) of Lemma 3.10.8 hold for $\hat{\boldsymbol{m}}^0 = \tanh(\boldsymbol{u}^0)$ with $\boldsymbol{u}^0$ satisfying Eq. (3.7.2).

To check assumption (3.10.9), we take K_{AMP} sufficiently large and δ_0 sufficiently small, obtaining

$$\begin{aligned}\|\hat{\boldsymbol{m}}^0 - \mathbf{m}_*(\boldsymbol{A}, \hat{\boldsymbol{y}}; q_*)\| &\le \|\hat{\boldsymbol{m}}^0 - \mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; K_{\mathsf{AMP}})\| + \|\mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; K_{\mathsf{AMP}}) - \mathbf{m}(\boldsymbol{A}, \boldsymbol{y})\| \\ &\quad + \|\mathbf{m}(\boldsymbol{A}, \boldsymbol{y}) - \mathbf{m}_*(\boldsymbol{A}, \boldsymbol{y}; q_*)\| + \|\mathbf{m}_*(\boldsymbol{A}, \boldsymbol{y}; q_*) - \mathbf{m}_*(\boldsymbol{A}, \hat{\boldsymbol{y}}; q_*)\| \\ &\overset{(a)}{\le} \frac{c\sqrt{\varepsilon t N}}{96(\beta^2+1)} + \frac{1}{100}\sqrt{\varepsilon t N} + \delta_0\sqrt{tN} + \frac{\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|}{c} \\ &\le \frac{\sqrt{\varepsilon t N}}{3}\end{aligned}$$

where inequality (a) holds with probability $1 - o_N(1)$. In the last step we used $c \le 1$.

To check Eq. (3.10.10), we use (3.10.19) we find that with probability $1 - o_N(1)$,

$$\begin{aligned}\|\nabla \mathscr{F}_{\mathrm{TAP}}(\hat{\boldsymbol{m}}^0; \hat{\boldsymbol{y}}, q_*)\| &\le \|\nabla \mathscr{F}_{\mathrm{TAP}}(\mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; K_{\mathsf{AMP}}); \boldsymbol{y}, q_*)\| + \|\boldsymbol{y} - \hat{\boldsymbol{y}}\| \\ &\quad + (4\beta^2 + 4)\|\operatorname{arctanh}(\hat{\boldsymbol{m}}^0) - \operatorname{arctanh}(\mathsf{AMP}(\boldsymbol{A}, \hat{\boldsymbol{y}}; K_{\mathsf{AMP}}))\| \\ &\le \|\nabla \mathscr{F}_{\mathrm{TAP}}(\mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; K_{\mathsf{AMP}}); \boldsymbol{y}, q_*)\| + \frac{c\sqrt{\varepsilon t N}}{24} + \frac{c\sqrt{\varepsilon t N}}{4}.\end{aligned}$$

Combining with Eq. (3.10.20), we find that with probability $1 - o_N(1)$,

$$\|\nabla \mathscr{F}_{\mathrm{TAP}}(\hat{\boldsymbol{m}}^0; \hat{\boldsymbol{y}}, q_*)\| \le \frac{c\sqrt{\varepsilon t N}}{6}.$$

Finally, we apply Lemma 3.10.6 with $\frac{\varepsilon t}{9}$ in place of ε, to get

$$\mathscr{F}_{\mathrm{TAP}}(\hat{\boldsymbol{m}}^0; \hat{\boldsymbol{y}}, q_*) \le \mathscr{F}_{\mathrm{TAP}}(\mathbf{m}_*(\boldsymbol{A}, \hat{\boldsymbol{y}}; q_*); \hat{\boldsymbol{y}}, q_*) + \frac{Nc\varepsilon t}{9}.$$

Lemma 3.10.8 now applies for η_0 sufficiently small. Moreover, with probability $1 - o_N(1)$ the initialization $\boldsymbol{x}^0$ satisfies

$$\begin{aligned}\|\operatorname{arctanh}(\hat{\boldsymbol{m}}^0)\| &\le \|\operatorname{arctanh}(\hat{\boldsymbol{m}}^0) - \operatorname{arctanh}(\mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; K_{\mathsf{AMP}}))\| + \|\operatorname{arctanh}(\mathsf{AMP}(\boldsymbol{A}, \boldsymbol{y}; K_{\mathsf{AMP}}))\| \\ &\le \frac{c\sqrt{\varepsilon t N}}{96(\beta^2+1)} + \sqrt{3(\gamma_*(\beta, t) + t)}\sqrt{N} \\ &\le C(\beta, c, \mathsf{T})\sqrt{tN}.\end{aligned}$$

Thus, (3.10.12) implies (3.7.3) for a sufficiently large number K_{NGD} of natural gradient iterations. □

Part II

Optimization of Mean-Field Spin Glasses

Chapter 4

Optimizing Mean-Field Spin Glasses via Approximate Message Passing

4.1 Introduction

Optimizing non-convex functions in high dimensions is well-known to be computationally intractible in general. In this chapter we study the optimization of a natural class of *random* non-convex functions, namely the Hamiltonians of mean-field spin glasses. These functions H_N are defined on either the cube $\Sigma_N = \{-1, 1\}^N$ or the sphere $\mathbb{S}^{N-1}(\sqrt{N})$ of radius $\sqrt{N}$ and have been studied since [SK75] as models for the behavior of disordered magnetic systems.

The distribution of an N-dimensional mean-field spin glass Hamiltonian H_N is described by an exponentially decaying sequence $(c_p)_{p\geq 2}$ of non-negative real numbers as well as an external field probability distribution $\mathcal{L}_h$ on $\mathbb{R}$ with finite second moment. Given these data, one samples $h_1, \ldots, h_N \sim \mathcal{L}_h$ and standard Gaussians $g_{i_1,\ldots,i_p} \sim \mathsf{N}(0,1)$ and then defines $H_N : \mathbb{R}^N \to \mathbb{R}$ by

$$H_N(\boldsymbol{x}) = \sum_i h_i x_i + \widetilde{H}_N(\boldsymbol{x}),$$
$$\widetilde{H}_N(\boldsymbol{x}) = \sum_{p=2}^{\infty} \frac{c_p}{N^{(p-1)/2}} \sum_{i_1,\ldots,i_p=1}^{N} g_{i_1,\ldots,i_p} x_{i_1} \cdots x_{i_p}.$$

The distribution of the non-linear part $\widetilde{H}_N$ is characterized by the mixture function $\xi(z) = \sum_{p\geq 2} c_p^2 z^p$ - there are no issues of convergence for $|z| \leq 1+\eta$ thanks to the exponential decay assumption. We assume throughout that ξ is not the zero function so that we study a genuine spin glass. $\widetilde{H}_N$ is then a centered Gaussian process with covariance

$$\mathbb{E}\big[\widetilde{H}_N(\boldsymbol{x}_1)\widetilde{H}_N(\boldsymbol{x}_2)\big] = N\xi\left(\frac{\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle}{N}\right).$$

Spin glasses were introduced to model the magnetic properties of diluted materials and have been studied in statistical physics and probability since the seminal work [SK75]. In this context, the object of study is the Gibbs measure $\frac{e^{\beta H_N(\boldsymbol{x})}d\mu(\boldsymbol{x})}{Z_{N,\beta}}$ where $\beta > 0$ is the inverse-temperature, $\mu(\boldsymbol{x})$ is a fixed reference measure and $Z_{N,\beta}$ is a random normalizing constant known as the partition function. The most common choice is to take $\mu(\cdot)$ the uniform measure on $\boldsymbol{\Sigma}_N = \{-1,1\}^N$, and another canonical choice is the uniform measure on $\mathbb{S}^{N-1}(\sqrt{N})$. These two choices define *Ising* and *spherical* spin glasses. The quantity of primary interest is the free energy

$$F_N(\beta) = \log \mathbb{E}^{\boldsymbol{x}\sim\mu}[e^{\beta H_N(\boldsymbol{x})}].$$

The in-probability normalized limit $F(\beta) = \text{p-lim}_{N\to\infty} \frac{F_N(\beta)}{N}$ of the free energy at temperature β is famously given by an infinite-dimensional variational problem known as the Parisi formula (or the Cristanti-Sommers formula in the spherical case) as we review in the next section. These free energies are well-concentrated and taking a second limit $\lim_{\beta\to\infty} \frac{F(\beta)}{\beta}$ yields the asymptotic ground state energies

$$GS(\xi, \mathcal{L}_h) = \text{p-lim}_{N\to\infty} \max_{\boldsymbol{x}\in\boldsymbol{\Sigma}_N} \frac{H_N(\boldsymbol{x})}{N},$$
$$GS_{\text{sph}}(\xi, \mathcal{L}_h) = \text{p-lim}_{N\to\infty} \max_{\boldsymbol{x}\in\mathbb{S}^{N-1}(\sqrt{N})} \frac{H_N(\boldsymbol{x})}{N}.$$

From the point of view of optimization, spin glass Hamiltonians serve as natural examples of highly non-convex functions. Indeed, the landscape of H_N can exhibit quite complicated behavior. For instance H_N may have exponentially many near-maxima on $\boldsymbol{\Sigma}_N$ [Cha09, DEZ15, CHL18]. The structure of these near-maxima is highly nontrivial; the Gibbs measures on $\boldsymbol{\Sigma}_N$ are approximate ultrametrics in a certain sense, at least in the so-called generic models [Jag17, CS21]. Moreover spherical spin glasses typically have exponentially many *local* maxima and saddle points, which are natural barriers to gradient descent and similar optimization algorithms [ABA13, ABAČ13, Sub17, AMMN19]. The utility of a rich model of random functions is made clear by a comparison to the theory of high-dimensional non-convex optimization in the worst-case setting. In the black-box model of optimization based on querying function values, gradients, and Hessians, approximately optimizing an unknown non-convex function in high-dimension efficiently is trivially impossible and

substantial effort has gone towards the more modest task of finding a local optimum or stationary point [CDHS17, JGN$^+$17, AAZB$^+$17, CDHS18, CDHS19]. Even for quadratic polynomials in N variables, it is quasi-NP hard to reach within a factor $\log(N)^{\varepsilon}$ of the optimum [ABE$^+$05]. For polynomials of degree $p \geq 3$ on the sphere, [BBH$^+$12] proves that even an approximation ratio $e^{(\log N)^{\varepsilon}}$ is computationally infeasible to obtain.

Despite the worst-case obstructions just outlined, a series of recent works have found great success in approximately maximizing certain spin glass Hamiltonians. By *approximate maximization* we always mean maximization up to a factor $(1+\varepsilon)$, where $\varepsilon > 0$ is an arbitrarily small positive constant; we similarly refer to a point $\boldsymbol{x} \in \boldsymbol{\Sigma}_N$ or $\boldsymbol{x} \in \mathbb{S}^{N-1}(\sqrt{N})$ achieving such a nearly optimal value as an *approximate maximizer* (where the small constant ε is implicit). Subag showed in [Sub21] how to approximately maximize spherical spin glasses by using top eigenvectors of the Hessian $\nabla^2 H_N$. Subsequently [Mon21] developed a message passing algorithm with similar guarantees for the Ising case. These works all operate under an assumption of no overlap gap, a condition which is expected (known in the spherical setting) to hold for some but not all models $(\xi, \mathcal{L}_h)$ - otherwise they achieve an explicit, sub-optimal energy value. Such a no overlap gap assumption is expected to be necessary to find approximate maxima efficiently. Indeed, the works [BAJ18, GJ21, GJW20b] rule out various algorithms for optimizing spin glasses when an overlap gap holds. Variants of the overlap gap property have been shown to rule out $(1+\varepsilon)$-approximation by certain classes of algorithms for random optimization problems on sparse graphs [MMZ05, ACORT11, GS14, RV17b, GS17, CGPR19, Wei22]. Overlap-gaps have also been proposed as evidence of computational hardness for a range of statistical tasks including planted clique, planted dense submatrix, sparse regression, and sparse principal component analysis [GZ17, GL18, GJS19, GZ19, AWZ20]. The overlap gap property is extensively discussed and generalized in the next chapter.

Our main algorithm consists of two stages of message passing. The first stage is inspired by the work [Bol14] which constructs solutions to the TAP equations for the SK model at high temperature. We construct approximate solutions to the generalized TAP equations of [Sub18, CPS22, CPS19], which heuristically amounts to locating the root of the ultrametric tree of approximate maxima. The second stage extends the algorithm of [Mon21], using incremental approximate message passing to descend the ultrametric tree by simulating the SDE corresponding to a candidate solution for the Parisi variational problem.

While the primary goal in this line of work is to construct a single approximate maximizer, Subag beautifully observed in [Sub21, Remark 6] that an extension of his Hessian-based construction for spherical models produces approximate maximizers arranged into a completely arbitary ultrametric space obeying an obvious diameter upper bound. The overlap gap property essentially states that distances between approximate maximizers cannot take certain values, and so this is a sort of constructive converse result. In Section 4.4 we give a branching version of our main algorithm, following a suggestion of [AM20], which constructs an arbitrary ultrametric space of approximate

maximizers in the Ising case (again subject to a diameter upper bound). This is a converse to and major motivation for the results of the next chapter on the branching overlap gap property.

4.1.1 Optimizing Ising Spin Glasses

To state our results we require the Parisi formula for the ground state of a mean field Ising spin glass as given in [AC17b]. Let $\mathscr{U}$ be the function space

$$\mathscr{U} = \left\{ \gamma : [0,1) \to [0,\infty) : \gamma \text{ is non-decreasing}, \int_0^1 \gamma(t)\mathrm{d}t < \infty \right\}.$$

The functions γ are meant to correspond to cumulative distribution functions - for finite β the corresponding Parisi formula requires $\gamma(1) = 1$, but this constraint disappears in renormalizing to obtain a zero-temperature limit. For $\gamma \in \mathscr{U}$ we take $\Phi_\gamma(t,x) : [0,1] \times \mathbb{R} \to \mathbb{R}$ to be the solution of the following Parisi PDE:

$$\partial_t \Phi_\gamma(t,x) + \frac{1}{2}\xi''(t)\left(\partial_{xx}\Phi_\gamma(t,x) + \gamma(t)(\partial_x \Phi_\gamma(t,x))^2\right) = 0, \tag{4.1.1}$$

$$\Phi_\gamma(1,x) = |x|. \tag{4.1.2}$$

Intimately related to the above PDE is the stochastic differential equation

$$\mathrm{d}X_t = \xi''(t)\gamma(t)\partial_x \Phi_\gamma(t, X_t)\mathrm{d}t + \sqrt{\xi''(t)}\mathrm{d}B_t, \quad X_0 \sim \mathcal{L}_h. \tag{4.1.3}$$

which we call the Parisi SDE. The Parisi functional $\mathsf{P} : \mathscr{U} \to \mathbb{R}$ with external field distribution $\mathcal{L}_h$ is given by:

$$\mathsf{P}_{\xi,\mathcal{L}_h}(\gamma) = \mathbb{E}^{h\sim\mathcal{L}_h}[\Phi_\gamma(0,h)] - \frac{1}{2}\int_0^1 t\xi''(t)\gamma(t)\mathrm{d}t. \tag{4.1.4}$$

The Parisi formula for the ground state energy is as follows.

Theorem 14 ([Tal06d, Pan14, AC17b, CHL18])**.**

$$GS(\xi, \mathcal{L}_h) = \inf_{\gamma\in\mathscr{U}} \mathsf{P}_{\xi,\mathcal{L}_h}(\gamma).$$

Moreover the minimum is attained at a unique $\gamma_*^{\mathscr{U}} \in \mathscr{U}$*.*

Throughout this chapter, $\gamma_*^{\mathscr{U}}$ will always refer to the minimizer of Theorem 14. We now turn to algorithms. In [Mon21], Montanari introduced the class of *incremental approximate message*

passing (IAMP) algorithms to optimize the SK model. These are a special form of the well-studied approximate message passing (AMP) algorithms, reviewed in Subsection 4.2.1. The work [AMS21] showed that the maximum asympototic value of H_N achievable by IAMP algorithms is given by the minimizer of P, assuming it exists, over a larger class of non-monotone functions, when $\mathcal{L}_h = \delta_0$ so there is no external field. This larger class is:

$$\mathscr{L} = \left\{ \gamma : [0,1) \to [0,\infty) : \gamma \text{ is right-continuous}, \|\xi'' \cdot \gamma\|_{TV[0,t]} < \infty \forall t \in [0,1), \int_0^1 \xi''(t)\gamma(t)\mathrm{d}t < \infty \right\}. \tag{4.1.5}$$

Here $TV[0,t]$ denotes the total variation norm

$$\|f\|_{TV[0,t]} \equiv \sup_n \sup_{0 \le t_0 < t_1 < \cdots < t_k \le t} \sum_{i=1}^k |f(t_i) - f(t_{i-1})| .$$

The Parisi PDE (4.1.4) and associated SDE extend also to $\mathscr{L}$. We denote by $\gamma_*^{\mathscr{L}} \in \mathscr{L}$ the minimizer of P over $\mathscr{L}$, assuming that it exists. Note that uniqueness always holds by Lemma 4.1.3 below. We define the *support* $\mathrm{supp}(\gamma)$ of $\gamma \in \mathscr{L}$ to be the closure in $[0,1)$ of $S(\gamma) \equiv \{x \in [0,1) : \gamma(x) > 0\}$. Note that this is not the same as the support of the signed measure with CDF γ.

Theorem 15 ([Mon21])**.** *For the Sherrington-Kirkpatrick model with $\mathcal{L}_h = \delta_0$ and $\xi(x) = x^2/2$, suppose $\inf_{\gamma \in \mathscr{U}} \mathsf{P}(\gamma)$ is achieved at a strictly increasing function $\gamma_*^{\mathscr{U}}$. Then for any $\varepsilon > 0$ there exists an efficient AMP algorithm which outputs $\boldsymbol{\sigma} \in \Sigma_N$ satisfying*

$$\frac{H_N(\boldsymbol{\sigma})}{N} \in \left[\mathsf{P}(\gamma_*^{\mathscr{U}}) - \varepsilon, \mathsf{P}(\gamma_*^{\mathscr{U}}) + \varepsilon\right]$$

with probability tending to 1 as $N \to \infty$.

The Parisi formula implies that $\mathsf{P}(\gamma_*^{\mathscr{U}})$ coincides with the ground state energy in the SK model. Thus, under the hypothesis that the minimizer $\gamma_*^{\mathscr{U}}$ is strictly increasing (which is supported by numerical simulations [CR02, OSS07, SO08]), Theorem 15 succeeds in efficient optimization up to $o(1)$ relative error.

We expand upon our use of the word "efficient" in Subsection 4.2.1 – in short, it means that $O_\varepsilon(1)$ evaluations of $\nabla \widetilde{H}_N$ and first or second partial derivatives of Φ_{γ_*} are required. In general, minimizing over the larger space $\mathscr{L}$ instead of $\mathscr{U}$ may decrease the infimum value of P, so that IAMP algorithms fail to approximately maximize H_N. However if $\gamma_*^{\mathscr{U}}$ is strictly increasing, then the infima are equal.

We now present our new results for more general mixed p-spin models in the presence of a non-trivial external field $\boldsymbol{h}$ with coordinate distribution $\mathcal{L}_h \neq \delta_0$. We first point out the external field requires a qualitatively new idea. Indeed the following proposition shows that any nonzero $\boldsymbol{h}$ forces $\gamma_*^{\mathscr{U}}(t) = 0$ in a neighborhood of $t = 0$, hence the strictly increasing assumption above cannot

apply. The proof is exactly the same as [Pan, Lemma A.19] (the same result stated for positive temperature).

Proposition 4.1.1. *We have* $0 \in \mathrm{supp}(\gamma_*^{\mathscr{U}})$ *if and only if* $\mathcal{L}_h = \delta_0$.

Despite this, we will show that approximate maximization is still possible if $\gamma_*^{\mathscr{U}}$ is strictly increasing on $[\underline{q}, 1)$ for $\underline{q} = \inf(\mathrm{supp}(\gamma_*^{\mathscr{U}}))$. If this condition holds, we give a two-phase approximate message passing algorithm which first locates a suitable point $\boldsymbol{m}_{\underline{\ell}}$ with L^2 norm $\|\boldsymbol{m}_{\underline{\ell}}\| \approx \sqrt{\underline{q}N}$, and then proceeds as in the no-external-field case. The relevant condition is precisely defined as follows.

Definition 4.1.2. *For* $\gamma_* \in \mathscr{L}$, *let* $\underline{q} = \inf(\mathrm{supp}(\gamma_*))$. *We say* γ_* *is* $\underline{q}$-***optimizable*** *if, with* X_t *given by* (4.1.3)*:*

$$\mathbb{E}[\partial_x \Phi_{\gamma_*}(t, X_t)^2] = t, \quad t \in [\underline{q}, 1). \tag{4.1.6}$$

We say $\gamma_* \in \mathscr{L}$ *is optimizable if it is* $\underline{q}$*-optimizable for* $\underline{q} = \inf(\mathrm{supp}(\gamma_*))$. *We say that* $(\xi, \mathcal{L}_h)$ *is optimizable, or equivalently that the* ***no overlap gap*** *property holds for* $(\xi, \mathcal{L}_h)$, *if the function* $\gamma_*^{\mathscr{U}}$ *is optimizable.*

Our preliminary numerical simulations suggest that the SK model retains the no overlap gap property with any constant external field $\mathcal{L}_h = \delta_h$. However proving this conjecture rigorously for any value of h seems difficult.

For $\underline{q} \in [0, 1)$, let $\mathscr{L}_{\underline{q}} = \{\gamma \in \mathscr{L} : \inf(\mathrm{supp}(\gamma)) \geq \underline{q}\}$ consist of functions in $\mathscr{L}$ vanishing on $[0, \underline{q})$. The next lemma shows optimizability is equivalent to minimizing P over either $\mathscr{L}$ or $\mathscr{L}_{\underline{q}}$. It is related to results in [AC15, JT16] which show that (4.1.6) holds at points of increase t for $\gamma_*^{\mathscr{U}}$. The proof is given in Appendix B.

Lemma 4.1.3. *For* $\gamma_* \in \mathscr{L}$ *and* $\underline{q} = \inf(\mathrm{supp}(\gamma_*))$, *the following are equivalent:*

1. γ_* *is optimizable.*
2. $\mathsf{P}(\gamma_*) = \inf_{\gamma \in \mathscr{L}} \mathsf{P}(\gamma)$.
3. $\mathsf{P}(\gamma_*) = \inf_{\gamma \in \mathscr{L}_{\underline{q}}} \mathsf{P}(\gamma)$.

Moreover if a minimizer exists in either variational problem just above, then it is unique.

Lemma 4.1.3 implies that any optimizable γ_* is in fact the unique minimizer $\gamma_*^{\mathscr{L}} \in \mathscr{L}$ of the Parisi functional. However throughout much of the chapter we will use γ_* to denote a general optimizable function without making use of this result. We made this choice because while Lemma 4.1.3 is important to make sense of our results, it is not necessary for proving e.g. Theorem 37 below.

Theorem 16. *Suppose $\gamma_* \in \mathscr{L}$ is optimizable. Then for any $\varepsilon > 0$ there exists an efficient AMP algorithm which outputs $\boldsymbol{\sigma} \in \Sigma_N$ such that*

$$\frac{H_N(\boldsymbol{\sigma})}{N} \in [\mathsf{P}(\gamma_*) - \varepsilon, \mathsf{P}(\gamma_*) + \varepsilon]$$

with probability tending to 1 as $N \to \infty$.

Lemma 4.1.4. *If $\gamma_*^{\mathscr{U}}$ strictly increases on $[\underline{q}, 1)$ for $\underline{q} = \inf(\mathrm{supp}(\gamma_*^{\mathscr{U}}))$, then no overlap gap holds, i.e. $\gamma_*^{\mathscr{U}}$ is optimizable.*

Corollary 4.1.5. *Suppose no overlap gap holds. Then for any $\varepsilon > 0$ an efficient AMP algorithm outputs $\boldsymbol{\sigma} \in \Sigma_N$ satisfying*

$$\frac{H_N(\boldsymbol{\sigma})}{N} \in [GS(\xi, \mathcal{L}_h) - \varepsilon, GS(\xi, \mathcal{L}_h) + \varepsilon]$$

with probability tending to 1 as $N \to \infty$.

Remark 4.1.1. Unlike for $\mathscr{U}$ the infimum $\inf_{\gamma \in \mathscr{L}} \mathsf{P}(\gamma)$ need not be achieved, i.e. an optimizable γ_* need not exist. For instance, one has $\xi''(0) = 0$ whenever $c_2 = 0$. On the other hand if γ is optimizable, Corollary B.1.6 and Lemma B.2.5 (with $\underline{q} = 0$) yield

$$\int_0^t \xi''(s)\, \mathbb{E}[\partial_{xx} \Phi_{\gamma_*}(s, X_s)^2] \mathrm{d}s = \mathbb{E}[\partial_x \Phi_{\gamma_*}(t, X_t)^2] \geq t, \quad t \geq 0.$$

In light of Lemma B.1.7 the integrand on the left-hand side is $O(\xi''(s)) = o(1)$ so the above cannot hold for small t. Hence if $c_2 = 0$ there exists no optimizable γ_*. We conjecture that conversely a minimizing $\gamma_*^{\mathscr{L}} \in \mathscr{L}$ exists whenever $c_2 > 0$, but we do not have a proof.

Remark 4.1.2. By the symmetry of $\widetilde{H}_N$, the external field can also be a deterministic vector $\boldsymbol{h} = (h_1, \ldots, h_N)$. As long as the empirical distribution of the values $(h_i)_{i \in [N]}$ is close to $\mathcal{L}_h$ in W_2 distance and the external field is independent of $\widetilde{H}_N$, exactly the same results hold. Indeed, in Theorem 39 we establish state evolution in this generality.

We conclude this subsection with some comments regarding our choices of terminology. Our definition of optimizability is closely related to "full" or "continuous" replica symmetry breaking. For example, the definitions of full RSB used in [Mon21, Sub21] essentially coincide with 0-optimizability. However these terms seem to be slightly ambiguous, as they can also refer to functions $\gamma_*^{\mathscr{U}}$ which are strictly increasing on **any** nontrivial interval instead of being piece-wise constant as in finite replica symmetry breaking. For example, the physics paper [CKP+14] describes "the case where the function $\Delta(x)$ is allowed to have a continuous part: this can be thought as an appropriate limit of the k-RSB construction when $k \to \infty$ and is therefore called 'fullRSB' or '∞-RSB' ". Adding to

the potential confusion, [ACZ17] uses the term "infinite step" RSB to refer to functions $\gamma_*^{\mathscr{U}}$ with infinity many points of increase, possibly at a discrete set. We therefore use "no overlap gap" as an unambiguous term for the condition that $\gamma_*^{\mathscr{U}}$ is optimizable, while keeping in mind that it closely is implied via Lemma 4.1.4 by a strong, specific form of full replica symmetry breaking.

4.1.2 Branching IAMP and Spherical Spin Glasses

Under no overlap gap, one expects that any finite ultrametric space of diameter at most $\sqrt{2(1-\underline{q})}$ (with size independent of N) can be realized by approximate maximizers of H_N. In fact a modification of our $\underline{q}$-IAMP algorithm is capable of explicitly producing such realizations. In Section 4.4 we give a *branching* $\underline{q}$-IAMP algorithm which for any finite ultrametric space X and optimizable γ_* constructs points $(\boldsymbol{\sigma}_x)_{x\in X}$ such that $\frac{H_N(\boldsymbol{\sigma}_x)}{N} \simeq \mathsf{P}(\gamma_*)$ and $\frac{\|\boldsymbol{\sigma}_x-\boldsymbol{\sigma}_y\|_2}{\sqrt{N}} \simeq d_X(x,y)$ for each $x, y \in X$. The idea is to occasionally reset the IAMP part of the algorithm with external randomness. A similar strategy was proposed but not analyzed in [AM20].

Theorem 17. *Let $\gamma_* \in \mathscr{L}$ be optimizable, and fix a finite ultrametric space (X, d_X) with diameter at most $\sqrt{2(1-\underline{q})}$ as well as $\varepsilon > 0$. Then an efficient AMP algorithm constructs points $\{\boldsymbol{\sigma}_x | x \in X\}$ in $\boldsymbol{\Sigma}_N$ satisfying*

$$\frac{H_N(\boldsymbol{\sigma}_x)}{N} \in [\mathsf{P}(\gamma_*) - \varepsilon, \mathsf{P}(\gamma_*) + \varepsilon], \quad x \in X,$$
$$\frac{\|\boldsymbol{\sigma}_x - \boldsymbol{\sigma}_y\|}{\sqrt{N}} \in [d_X(x,y) - \varepsilon, d_X(x,y) + \varepsilon], \quad x, y \in X$$

with probability tending to 1 as $N \to \infty$.

In Section 4.5 we give corresponding results for spherical spin glasses, extending [Sub21] to the case of non-trivial external field. At zero temperature, [CS17, Theorem 1] determines the free energy in spherical spin glasses based on a positive, non-decreasing function $\alpha : [0,1) \to [0,\infty)$ as well as a constant L. (See also [JT17] for related results.) More precisely, they show the asymptotic ground state energy is given by the unique minimizer to the variational problem:

$$GS_{\text{sph}}(\xi, h) = \min_{L,\alpha\in\mathcal{K}} \mathcal{Q}(L,\alpha); \tag{4.1.7}$$
$$\mathcal{K} = \left\{(L,\alpha) \in (0,\infty) \times \mathscr{U} \ : \ L > \int_0^1 \alpha(s)\mathrm{d}s\right\};$$
$$2\mathcal{Q}(L,\alpha) = (\xi'(1) + h^2)L - \int_0^1 \xi''(q)\left(\int_0^q \alpha(s)\mathrm{d}s\right)\mathrm{d}q + \int_0^1 \frac{\mathrm{d}q}{L - \int_0^q \alpha(s)\mathrm{d}s}.$$

The associated definition of no overlap gap is as follows.

Definition 4.1.6. *The spherical mixed p-spin model is said* no overlap gap *if for some* $\underline{q}_{sph} \in [0,1)$*, the unique minimizing* $\alpha \in \mathscr{U}$ *in* (4.1.7) *is strictly increasing on* $[\underline{q}_{sph}, 1)$ *and satisfies* $\alpha(q) = 0$ *for* $q \leq \underline{q}_{sph}$*.*

Unlike the Ising case, we do not formulate a generalized variational principle and only show how to achieve a natural energy value, which coincides with the ground state energy when no overlap gap holds by [CS17, Proposition 2]. We also exactly characterize the spherical models exhibiting no overlap gap, which slightly extends the same result.

Theorem 18. *Suppose* ξ *and* $\mathcal{L}_h$ *satisfy* $\mathbb{E}[h^2] + \xi'(1) < \xi''(1)$*, and let* $\underline{q}_{sph} \in (0,1)$ *be the unique solution to* $\mathbb{E}[h^2] + \xi'(\underline{q}_{sph}) = \underline{q}_{sph}\xi''(\underline{q}_{sph})$*. Then the spherical spin glass with parameters* $\xi, \mathcal{L}_h$ *has no overlap gap if and only if* $\xi''(q)^{-1/2}$ *is concave on* $q \in [\underline{q}_{sph}, 1]$*, in which case* α *is supported on* $[\underline{q}_{sph}, 1]$ *and takes the explicit form*

$$\alpha(s) = \begin{cases} 0, & s \in [0, \underline{q}_{sph}) \\ \frac{\xi'''(s)}{2\xi''(s)^{3/2}}, & s \in [\underline{q}_{sph}, 1]. \end{cases}$$

Moreover the ground-state energy satisfies

$$GS_{sph}(\xi, \mathcal{L}_h) \geq \underline{q}_{sph}\sqrt{\xi''(\underline{q}_{sph})} + \int_{\underline{q}_{sph}}^{1} \sqrt{\xi''(q)}\mathrm{d}q$$

with equality if and only if no overlap gap occurs.

Theorem 19. *Suppose* ξ *and* $h \sim \mathcal{L}_h$ *satisfy* $\mathbb{E}[h^2] + \xi'(1) < \xi''(1)$*, and let* $\underline{q}_{sph} \in (0,1)$ *be the unique solution to* $\mathbb{E}[h^2] + \xi'(\underline{q}_{sph}) = \underline{q}_{sph}\xi''(\underline{q}_{sph})$*. Then there exists an efficient AMP algorithm outputting* $\boldsymbol{\sigma} \in \mathbb{S}^{N-1}(\sqrt{N})$ *such that*

$$\frac{H_N(\boldsymbol{\sigma})}{N} \simeq \underline{q}_{sph}\sqrt{\xi''(\underline{q}_{sph})} + \int_{\underline{q}_{sph}}^{1} \sqrt{\xi''(q)}\mathrm{d}q.$$

If on the other hand $\mathbb{E}[h^2] + \xi'(1) \geq \xi''(1)$*, then there is an efficient AMP algorithm outputting* $\boldsymbol{\sigma} \in \mathbb{S}^{N-1}(\sqrt{N})$ *with*

$$\frac{H_N(\boldsymbol{\sigma})}{N} \simeq \sqrt{\mathbb{E}[h^2] + \xi'(1)}.$$

Remark 4.1.3. If $\mathbb{E}[h^2] + \xi'(1) \geq \xi''(1)$ then the model is replica-symmetric by [CS17, Proposition 1]. When $\mathbb{E}[h^2] + \xi'(1) < \xi''(1)$, the function $f(q) = q\xi''(q) - \xi'(q) - \mathbb{E}[h^2]$ is increasing and satisfies $f(0) < 0 < f(1)$, hence has a unique root $\underline{q}_{\mathrm{sph}} \in (0,1)$.

4.2 Technical Preliminaries

We will use ordinary lower-case letters for scalars $(m, x, \dots,)$ and bold lower-case for vectors $(\boldsymbol{m}, \boldsymbol{x})$. Ordinary upper-case letters are used for the state-evolution limits of AMP as in Proposition 4.2.3 such as $(X_j^\delta, Z_j^\delta, N_j^\delta)$ as well as for continuous-time stochastic processes such as (X_t, Z_t, N_t). We denote limits in probability as $N \to \infty$ by $\text{p-lim}_{N\to\infty}(\cdot)$. We write $x \simeq y$ to indicate that $\text{p-lim}_{N\to\infty}(x - y) = 0$ where x, y are random scalars.

We will use the ordinary inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^N x_i y_i$ as well as the normalized inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_N = \frac{\sum_{i=1}^N x_i y_i}{N}$. Here $\boldsymbol{x} = (x_1, \dots, x_N) \in \mathbb{R}^N$ and similarly for $\boldsymbol{y}$. Associated with these are the norms $\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$ and $\|\boldsymbol{x}\|_N = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_N}$. We will also use the notation $\langle \boldsymbol{x} \rangle_N = \frac{\sum_{i=1}^N x_i}{N}$. Often, for example in (4.2.2), we apply a scalar function f to a vector $\boldsymbol{x} \in \mathbb{R}^N$. This will always mean that f is applied entrywise, i.e. $f(x_1, \dots, x_N) = (f(x_1), \dots, f(x_N))$. Similarly for a function $f : \mathbb{R}^{\ell+1} \to \mathbb{R}$, we define

$$f(\boldsymbol{x}^0, \boldsymbol{x}^1, \dots, \boldsymbol{x}^\ell) = \big(f(x_1^0, x_1^1, \dots, x_1^\ell),\ f(x_2^0, x_2^1, \dots, x_2^\ell),\ \dots\ f(x_N^0, x_N^1, \dots, x_N^\ell)\big) \in \mathbb{R}^N. \tag{4.2.1}$$

The following useful a priori estimate shows that all derivatives of $\frac{H_N}{N}$ have order 1 in the $\|\cdot\|_N$ norm. Note that we do not apply any non-standard normalization in the definitions of gradients, Hessians, etc.

Proposition 4.2.1 ([BASZ20, Corollary 59])**.** *Fix a mixture function ξ, external field distribution $\mathcal{L}_h$, $k \in \mathbb{Z}^+, \eta \in \mathbb{R}^+$, and assume that the coefficients of ξ decay exponentially. Then for suitable $C = C(\xi, \mathcal{L}_h, k, \eta)$,*

$$\mathbb{P}\left[\sup_{\|\boldsymbol{x}\| \le (1+\eta)\sqrt{N}} \|\nabla^k \widetilde{H}_N(\boldsymbol{x})\| \le C N^{1-\frac{k}{2}}\right] \ge 1 - e^{-\Omega(N)}.$$

4.2.1 Review of Approximate Message Passing

Here we review the general class of approximate message passing (AMP) algorithms. AMP algorithms are a flexible class of efficient algorithms based on a random matrix or, in our setting, mixed tensor. To specify an AMP algorithm, we fix a probability distribution p_0 on $\mathbb{R}$ with finite second moment and a sequence $f_0, f_1, \dots$ of Lipschitz functions $f_\ell : \mathbb{R}^{\ell+1} \to \mathbb{R}$, with $f_{-1} = 0$. The functions f_ℓ will often be referred to as *non-linearities*. We begin by taking $\boldsymbol{z}^0 \in \mathbb{R}^N$ to have i.i.d. coordinates

$(z_i^0)_{i\in[N]} \sim p_0$. Then we recursively define $\boldsymbol{z}^1, \boldsymbol{z}^2, \dots$ via

$$\boldsymbol{z}^{\ell+1} = \nabla \widetilde{H}_N(f_\ell(\boldsymbol{z}^0, \dots, \boldsymbol{z}^\ell)) - \sum_{j=1}^{\ell} d_{\ell,j} f_{j-1}(\boldsymbol{z}^0, \dots, \boldsymbol{z}^{j-1}), \tag{4.2.2}$$

$$d_{\ell,j} = \xi''\left(\left\langle f_\ell(\boldsymbol{z}^0, \dots, \boldsymbol{z}^\ell), f_{j-1}(\boldsymbol{z}^0, \dots, \boldsymbol{z}^{j-1})\right\rangle_N\right) \cdot \mathbb{E}\left[\frac{\partial f_\ell}{\partial Z^j}(Z^0, \dots, Z^\ell)\right]. \tag{4.2.3}$$

Here the non-linearity f_ℓ is applied coordinate-wise as in (4.2.1). Moreover $Z^0 \sim p_0$ while $(Z^\ell)_{\ell\geq 1}$ is an independent centered Gaussian process with covariance $Q_{\ell,j} = \mathbb{E}[Z^\ell Z^j]$ defined recursively by

$$Q_{\ell+1,j+1} = \xi'\left(\mathbb{E}\left[f_\ell\left(Z^0, \cdots, Z^\ell\right) f_j\left(Z^0, \cdots, Z^j\right)\right]\right), \quad \ell, j \geq 0. \tag{4.2.4}$$

The key property of AMP, stated below in Proposition 4.2.3, is that for any ℓ the empirical distribution of the N sequences $(\boldsymbol{z}_i^1, \boldsymbol{z}_i^2, \dots, \boldsymbol{z}_i^\ell)_{i\in[N]}$ converges in distribution to the law of the Gaussian process $(Z^1, \dots, Z^\ell)$ as $N \to \infty$. This is called *state evolution*.

Definition 4.2.2. *For non-negative integers n, m the function $\psi : \mathbb{R}^\ell \to \mathbb{R}$ is* pseudo-Lipschitz *if for some constant L and any $x, y \in \mathbb{R}^\ell$,*

$$\|\psi(x) - \psi(y)\| \leq L(1 + \|x\| + \|y\|)\|x - y\|.$$

Proposition 4.2.3. *For any pseudo-Lipschitz $\psi : \mathbb{R}^{\ell+1} \to \mathbb{R}$, the AMP iterates satisfy*

$$\operatorname*{p-lim}_{N\to\infty} \left\langle \psi\left(\boldsymbol{z}^0, \cdots, \boldsymbol{z}^\ell\right)\right\rangle_N = \mathbb{E}\left[\psi\left(Z^0, \cdots, Z^\ell\right)\right]. \tag{4.2.5}$$

The first version of state evolution was given for Gaussian random matrices in [Bol14, BM11b]. Since then it has been extended to more general situations in many works including [JM13, BLM15, BMN19, CL21, Fan22]. As state evolution holds for essentially arbitrary non-linearities f_ℓ, it allows a great deal of flexibility in solving problems involving random matrices or tensors.

We remark that when proving state evolution in Theorem 39, we phrase the result in terms of a random mixed tensor $\boldsymbol{W}$, i.e. a sequence of p-tensors $(\boldsymbol{W}^{(p)} \in (\mathbb{R}^N)^{\otimes p})_{p\geq 2}$. The two descriptions are equivalent because $\boldsymbol{W}$ is constructed so that $\sum_{p\geq 2} c_p \langle \boldsymbol{W}^{(p)}, \boldsymbol{x}^{\otimes p}\rangle = \widetilde{H}_N(\boldsymbol{x})$. While the tensor language is better suited to proving state evolution, for our purposes it is more convenient to express AMP just in terms of $\widetilde{H}_N$ and $\nabla \widetilde{H}_N$.

Let us finally discuss the efficiency of our AMP algorithms. The algorithms we give are described by parameters $\bar{q}$ and $\underline{\ell}$ and require oracle access to the function $\Phi_{\gamma_*}(t, x)$ and its derivatives. We do not address the complexity of computing $\Phi_{\gamma_*}(t, x)$. However as stated in [Mon21] it seems unlikely to present a major obstacle because solving for $\gamma_*^{\mathscr{U}}$ is a convex problem which only must be solved once for each $(\xi, \mathcal{L}_h)$. Moreover [AM20] demonstrates that these algorithms are practical to implement.

In the end, our algorithms output rounded points $\boldsymbol{\sigma}$ with $\boldsymbol{\sigma}_i = \text{sign}(f_{\overline{\ell}}(\boldsymbol{z}_i^0, \ldots, \boldsymbol{z}_i^{\ell}))$ for a large value $\overline{\ell} = \overline{\ell}(\bar{q}, \underline{\ell})$. The outputs satisfy

$$\lim_{\bar{q}\to 1} \lim_{\underline{\ell}\to\infty} \text{p-lim}_{N\to\infty} \frac{H_N(\boldsymbol{\sigma})}{N} = H_*$$

for some asymptotic energy value H_*. To achieve an ε-approximation to the value H_*, the parameters $\bar{q}$ and $\underline{\ell}$ must be sent to 1 and ∞ which requires a diverging number of iterations. In particular let χ denote the complexity of computing $\nabla \widetilde{H}_N$ at a point and let χ_1 denote the complexity of computing a single coordinate of $\nabla \widetilde{H}_N$ at a point. Then the total complexity needed to achieve energy $H_* - \varepsilon$ is $C(\varepsilon)(\chi + N) + N\chi_1$. When ξ is a polynomial this complexity is linear in the size of the input specifying H_N. In the statements of our results, we refer to such algorithms as "efficient AMP algorithms".

4.2.2 Initializing AMP

Here we explain some technical points involved in initializing our AMP algorithms and why they arise. First, we would like to use a random external field h_i which varies from coordinate to coordinate. In the most natural AMP implementation, this requires that the non-linearities f_ℓ correspondingly depend on the coordinate rather than being fixed, which is not allowed in state evolution. Second we would like to use many i.i.d. Gaussian vectors throughout the branching version of the algorithm. However Proposition 4.2.3 allows only a single initial vector $\boldsymbol{z}^0$ as a source of external randomness independent of H_N. One could prove a suitable generalization of Proposition 4.2.3, but we instead build these additional vectors into the initialization of the AMP algorithm as a sort of preprocessing phase. To indicate that our constructions here are preparation for the "real algorithm", we reparametrize so the preparatory iterates have negative index.

We begin by taking $p_0 = \mathcal{L}_h$ to be the distribution of the external field itself, and initialize $(\boldsymbol{z}^{-K})_i = h_i \sim \mathcal{L}_h$ for some constant $K \in \mathbb{Z}^+$. We then set $f_{-K}(\boldsymbol{z}^{-K}) = \frac{\boldsymbol{z}^{-K}}{\sqrt{\mathbb{E}^{h\sim\mathcal{L}_h}[h^2]}}$ and $f_{-k}(\boldsymbol{z}^{-K}, \ldots, \boldsymbol{z}^{-k}) = \boldsymbol{z}^{-k}$ for $2 \le k \le K$. Finally we set $f_{-1}(\boldsymbol{z}^{-K}, \ldots, \boldsymbol{z}^{-1}) = c\boldsymbol{z}^{-1}$ for some constant $c > 0$ which the algorithm is free to choose. (Note that the functions f_{-k} correspond to entry-wise applications of the form in (4.2.1).) State evolution immediately implies the following

Proposition 4.2.4. *In the state evolution $N \to \infty$ limit, $(z_i^{-K}, \ldots, z_i^0)$ converges in distribution to the law of an independent $(K+1)$-tuple $(Z^{-K}, Z^{-K+1}, \ldots, Z^{-1}, Z^0)$ with $Z^{-K} \sim \mathcal{L}_h$, $(Z^{-K+1}, \ldots, Z^{-1}) \sim \mathsf{N}(0, I_{K-1})$ i.i.d. standard Gaussian, and $Z^0 \sim \mathsf{N}(0, c^2)$.*

In fact taking $K = 1$ suffices for the main construction of this chapter. In Section 4.4 we require larger values of K for branching IAMP, where the iterates $(\boldsymbol{z}^{-K+1}, \ldots, \boldsymbol{z}^{-1})$ serve as proxies for i.i.d. Gaussian vectors.

Remark 4.2.1. Because the sum defining the Onsager correction term in (4.2.2) starts at $j = 1$, the effect of the external field h_i on future AMP iterates does not enter into any Onsager correction terms in this chapter.

4.2.3 Properties of the Parisi PDE and SDE

Quite a lot is known about the solution Φ_γ to the Parisi PDE. The next results hold for any $\gamma \in \mathscr{L}$ and are shown in the Appendix. Similar results for $\gamma \in \mathscr{U}$ appear in [AC15, JT16]. The following two propositions are shown (with some rearrangement) in Lemmas B.1.2 and B.1.4.

Proposition 4.2.5. *For any $\gamma \in \mathscr{L}$, the solution $\Phi_\gamma(t,x)$ to the Parisi PDE is continuous on $[0,1] \times \mathbb{R}$, convex in x, and further satisfies the following regularity properties for any $\varepsilon > 0$.*

(a) $\partial_x^j \Phi \in L^\infty\big([0, 1-\varepsilon]; L^2(\mathbb{R}) \cap L^\infty(\mathbb{R})\big)$ for $j \geq 2$.

(b) $\partial_t \Phi \in L^\infty([0,1] \times \mathbb{R})$ and $\partial_t \partial_x^j \Phi \in L^\infty\big([0, 1-\varepsilon]; L^2(\mathbb{R}) \cap L^\infty(\mathbb{R})\big)$ for $j \geq 1$.

Proposition 4.2.6. *For any $\gamma \in \mathscr{L}$, Φ_γ satisfies*

$$|\partial_x \Phi_\gamma(t,x)| \leq 1$$

for all $(t,x) \in [0,1] \times \mathbb{R}$.

The next proposition is shown in Lemma B.1.5.

Proposition 4.2.7. *For any $\gamma \in \mathscr{L}$, the Parisi SDE* (4.1.3) *has unique strong solution $(X_t)_{t\in[0,1]}$ which is a.s. continuous and satisfies*

$$\partial_x \Phi_\gamma(t, X_t) = \int_0^t \sqrt{\xi''(s)}\, \partial_{xx} \Phi_\gamma(s, X_s)\, \mathrm{d}B_s\,. \tag{4.2.6}$$

Finally we give two additional properties for optimizable γ_*, which are proved in Chapter B.

Lemma 4.2.8. *If $\gamma_* \in \mathscr{L}$ is $\underline{q}$-optimizable then it satisfies:*

$$\mathbb{E}[\partial_{xx} \Phi_{\gamma_*}(t, X_t)^2] = \frac{1}{\xi''(t)}, \quad t \geq \underline{q}, \tag{4.2.7}$$

$$\mathbb{E}[\partial_{xx} \Phi_{\gamma_*}(t, X_t)] = \int_t^1 \gamma_*(s) \mathrm{d}s, \quad t \in [0,1]. \tag{4.2.8}$$

4.3 The Main Algorithm

In this section we explain our main AMP algorithm and prove Theorem 37. Throughout we take $\gamma_* \in \mathscr{L}$ to be $\underline{q}$-optimizable for $\underline{q} = \inf(\mathrm{supp}(\gamma_*)) \in [0,1)$.

4.3.1 Phase 1: Finding the Root

Here we give the first phase of the algorithm, which proceeds for a large constant number $\underline{\ell}$ of iterations after initialization and approximately converges to a fixed point. The AMP iterates during this first phase are denoted by $(\boldsymbol{w}^k)_{-K \le k \le \underline{\ell}}$. We rely on the function

$$f(x) = \partial_x \Phi_{\gamma_*}(\underline{q}, x)$$

and use non-linearities

$$f_k(\boldsymbol{h}, \boldsymbol{w}^{-K+1}, \dots, \boldsymbol{w}^0, \boldsymbol{w}^1, \dots, \boldsymbol{w}^k) = f(\boldsymbol{h} + \boldsymbol{w}^k)$$

for all $k \ge 1$. (As a reminder, if f is a scalar function, $f(\boldsymbol{x}^k)$ is evaluated entrywise as explained in (4.2.1).) Proposition 4.2.5 implies that each f_k is Lipschitz, so that state evolution applies to the AMP iterates. In the initialization phase we take $c = \sqrt{\xi'(\underline{q})}$ as described in Subsection 4.2.2, so that the coordinates w_i^0 are asymptotically distributed as centered Gaussians with variance $\xi'(\underline{q})$ in the $N \to \infty$ limit. Moreover we set $\boldsymbol{m}^k = f(\boldsymbol{x}^k)$ where $\boldsymbol{x}^k = \boldsymbol{w}^k + \boldsymbol{h}$. This yields the following iteration.

$$\begin{aligned} \boldsymbol{w}^{k+1} &= \nabla \widetilde{H}_N(f(\boldsymbol{x}^k)) - f(\boldsymbol{x}^{k-1}) \xi''\left(\langle f(\boldsymbol{x}^k), f(\boldsymbol{x}^{k-1})\rangle_N\right) \langle \nabla f(\boldsymbol{x}^k)\rangle_N \\ &= \nabla \widetilde{H}_N(\boldsymbol{m}^k) - \boldsymbol{m}^{k-1} \xi''\left(\langle \boldsymbol{m}^k, \boldsymbol{m}^{k-1}\rangle_N\right) \langle \partial_{xx} \Phi_{\gamma_*}(\underline{q}, \boldsymbol{x}^k)\rangle_N, \\ \boldsymbol{x}^{k+1} &= \boldsymbol{w}^{k+1} + \boldsymbol{h} \\ \boldsymbol{m}^k &= f(\boldsymbol{x}^k) = f_k(\boldsymbol{w}^k). \end{aligned} \tag{4.3.1}$$

Lemma 4.3.1. *For f as defined above, $h \sim \mathcal{L}_h$ and $Z \sim \mathsf{N}(0,1)$ an independent standard Gaussian,*

$$\mathbb{E}^{h,Z}\left[f\left(h + Z\sqrt{\xi'(\underline{q})}\right)^2\right] = \underline{q} \tag{4.3.2}$$

$$\mathbb{E}^{h,Z}\left[f'\left(h + Z\sqrt{\xi'(\underline{q})}\right)^2\right] = \frac{1}{\xi''(\underline{q})}. \tag{4.3.3}$$

Proof. The identities follow by taking $t = \underline{q}$ in the definition of optimizability as well as Lemma 4.2.8. Here we use the fact that $X_t = X_0 + Z\sqrt{\xi'(t)}$ is a time-changed Brownian motion started from X_0 for $t \le \underline{q}$. □

Next with $(Z, Z', Z'') \sim \mathsf{N}(0, I_3)$ independent of $h \sim \mathcal{L}_h$, define for $t \le \xi'(\underline{q})$ the function

$$\phi(t) = \mathbb{E}^{h,Z,Z',Z''}\left[f\left(h + Z\sqrt{t} + Z'\sqrt{\xi'(\underline{q}) - t}\right) f\left(h + Z\sqrt{t} + Z''\sqrt{\xi'(\underline{q}) - t}\right)\right]. \tag{4.3.4}$$

Define also $\psi(t) = \xi'(\phi(t))$. It follows from (4.3.2) that

$$\phi(\xi'(\underline{q})) = \underline{q}. \tag{4.3.5}$$

Lemma 4.3.2. *The function ψ is strictly increasing and strictly convex on $[0, \xi'(\underline{q})]$. Moreover*

$$\psi(\xi'(\underline{q})) = \xi'(\underline{q}), \quad \psi'(\xi'(\underline{q})) = 1.$$

Finally $\psi(t) > t$ for all $t < \xi'(\underline{q})$.

Proof. Using Gaussian integration by parts as in [Bol14, Lemma 2.2], we find

$$\begin{aligned}
\phi'(t) &= \overset{h,Z,Z',Z''}{\mathbb{E}}\left[f'\left(\sqrt{t}Z + \sqrt{\xi'(\underline{q}) - t}Z'\right) f'\left(\sqrt{t}Z + \sqrt{\xi'(\underline{q}) - t}Z''\right)\right] \\
&= \overset{h,Z}{\mathbb{E}}\left[\overset{Z'}{\mathbb{E}}\left[f'\left(\sqrt{t}Z + \sqrt{\xi'(\underline{q}) - t}Z'\right)\right]^2\right], \\
\phi''(t) &= \mathbb{E}\left[f''\left(\sqrt{t}Z + \sqrt{\xi'(\underline{q}) - t}Z'\right) f''\left(\sqrt{t}Z + \sqrt{\xi'(\underline{q}) - t}Z''\right)\right] \\
&= \overset{h,Z}{\mathbb{E}}\left[\overset{Z'}{\mathbb{E}}\left[f''\left(\sqrt{t}Z + \sqrt{\xi'(\underline{q}) - t}Z'\right)\right]^2\right].
\end{aligned}$$

These expressions are each strictly positive, as the optimizability of γ_* implies that f', f'' are not identically zero. Therefore ϕ is increasing and convex. Since ξ' is also increasing and convex (being a power series with non-negative coefficients) we conclude the same about their composition ψ. The values $\psi(\xi'(\underline{q})) = \xi'(\underline{q})$ and $\psi'(\xi'(\underline{q})) = 1$ follow from Lemma 4.3.1 and the chain rule. Finally the last claim follows by strict convexity of ψ and $\psi'(\xi'(\underline{q})) = 1$. □

Next, let $h, W^{-1}, (W^j, X^j, M^j)_{j \ge 0}$ be the state evolution limit of the coordinates of

$$(\boldsymbol{h}, \boldsymbol{w}^{-1}, \boldsymbol{w}^0, \boldsymbol{x}^0, \boldsymbol{m}^0, \dots, \boldsymbol{w}^k, \boldsymbol{x}^k, \boldsymbol{m}^k)$$

as $N \to \infty$. Hence each W^j is a centered Gaussian and $X^j = W^j + h$, $M^{j+1} = f(X^j)$ hold for $j \ge 0$. Define the sequence $(a_0, a_1, \dots)$ recursively by $a_0 = 0$ and $a_{k+1} = \psi(a_k)$.

Lemma 4.3.3. *For all non-negative integers $0 \le j < k$, the following equalities hold:*

$$\mathbb{E}[(W^j)^2] = \xi'(\underline{q}) \tag{4.3.6}$$

$$\mathbb{E}[W^j W^k] = a_j \tag{4.3.7}$$

$$\mathbb{E}[(M^j)^2] = \underline{q} \tag{4.3.8}$$

$$\mathbb{E}[M^j M^k] = \phi(a_j). \tag{4.3.9}$$

Moreover $(W^j)_{j\ge 0}$ is independent of h.

Proof. We proceed by induction on j, first showing (4.3.6) and (4.3.8) together. As a base case, (4.3.6) holds for $j = 0$ by initialization. For the inductive step, assume first that (4.3.6) holds for j. Then state evolution and (4.3.5) yield

$$\mathbb{E}\left[(M^j)^2\right] = \phi\left(\xi'(\underline{q})\right) = \underline{q}$$

so that (4.3.6) implies (4.3.8) for each $j \ge 0$. On the other hand, state evolution directly implies that if (4.3.8) holds for j then (4.3.6) holds for $j + 1$. This establishes (4.3.6) and (4.3.8) for all $j \ge 0$.

We similarly show (4.3.7) and (4.3.9) together by induction, beginning with (4.3.7) when $j = 0$. By the initialization of Subsection 4.2.2 it follows that the random variables h, W^{-1}, W^0 are jointly independent. State evolution implies that W^{k-1} is independent of W^{-1} for any $k \ge 0$. Then state evolution yields for any $k \ge 1$:

$$\begin{aligned}
\mathbb{E}[W^0 W^k] &= \xi'(\mathbb{E}[M^{-1} M^{k-1}]) \\
&= \xi'\left(\mathbb{E}\left[W^{-1} f(W^{k-1}\right]\right) \\
&= \xi'(0) \\
&= 0.
\end{aligned}$$

Just as above, it follows from state evolution that (4.3.7) for (j, k) implies (4.3.9) for (j, k) which in turn implies (4.3.7) for $(j+1, k+1)$. Hence induction on j proves (4.3.7) and (4.3.9) for all (j, k). Finally the last independence assertion is immediate from state evolution just because h is the first step in the AMP iteration. □

Lemma 4.3.4.

$$\lim_{k\to\infty} a_k = \xi'(\underline{q}).$$

Proof. Since ψ is strictly increasing and maps $[0, \xi'(\underline{q})] \to [0, \xi'(\underline{q})]$, it follows that $(a_k)_{k \geq 0}$ is a strictly increasing sequence with limiting value in $[0, \xi'(\underline{q})]$. Let $a_\infty = \lim_{k \to \infty} a_k$ be this limit. Then continuity implies $\psi(a_\infty) = a_\infty$ which by the last part of Lemma 4.3.2 implies $a_\infty = \xi'(\underline{q})$. This concludes the proof. □

We now compute the limiting energy

$$\lim_{k \to \infty} \operatorname*{p-lim}_{N \to \infty} \frac{H_N(\boldsymbol{m}_k)}{N}$$

from the first phase. Since the first phase is similar to many "standard" AMP algorithms, this step is comparable to the computation of their final objective value, for example [DAM17, Lemma 6.3]. This computation is straightforward when $\widetilde{H}_N$ is a homogeneous polynomial of degree p, because one can just rearrange the equation for an AMP iteration to solve for

$$\widetilde{H}_N(\boldsymbol{m}^k) = p^{-1} \langle \boldsymbol{m}^k, \nabla \widetilde{H}_N(\boldsymbol{m}^k) \rangle.$$

However it requires more work in our setting because $\nabla \widetilde{H}_N$ acts differently on terms of different degrees. We circumvent this mismatch by applying state evolution to a t-dependent auxiliary AMP step and integrating in t.

Lemma 4.3.5. *With X_t the Parisi SDE* (4.1.3)*,*

$$\begin{aligned} \lim_{k \to \infty} \operatorname*{p-lim}_{N \to \infty} \frac{H_N(\boldsymbol{m}^k)}{N} &= \xi'(\underline{q}) \cdot \mathbb{E}\left[\partial_{xx} \Phi_{\gamma_*}\left(\underline{q}, h + Z\sqrt{\xi'(\underline{q})}\right)\right] + \mathbb{E}\left[h \cdot \partial_x \Phi_{\gamma_*}\left(\underline{q}, h + Z\sqrt{\xi'(\underline{q})}\right)\right] \\ &= \xi'(\underline{q}) \cdot \mathbb{E}\left[\partial_{xx} \Phi_{\gamma_*}\left(\underline{q}, X_{\underline{q}}\right)\right] + \mathbb{E}\left[h \cdot \partial_x \Phi_{\gamma_*}\left(\underline{q}, X_{\underline{q}}\right)\right]. \end{aligned}$$

Proof. The equivalence of the latter two expressions follows from the fact that $X_{\underline{q}} \sim X_0 + \mathsf{N}(0, \xi'(\underline{q}))$ so we focus on the first equality. Observe that

$$\frac{H_N(\boldsymbol{m}^k)}{N} = \langle \boldsymbol{h}, \boldsymbol{m}^k \rangle_N + \int_0^1 \langle \boldsymbol{m}^k, \nabla \widetilde{H}_N(t\boldsymbol{m}^k) \rangle_N \mathrm{d}t \tag{4.3.10}$$

holds for any vector $\boldsymbol{m}^k$ by considering each monomial term of H_N. Our main task now reduces to computing the in-probability limit of the integrand as a function of t. Proposition 4.2.1 ensures that $t \to \langle \boldsymbol{m}^k, \nabla \widetilde{H}_N(t\boldsymbol{m}^k) \rangle_N$ is Lipschitz assuming $\|\boldsymbol{m}^k\|_N \leq 1 + o(1)$. This holds with high probability for each k as $N \to \infty$ by state evolution and Proposition 4.2.6, so we may freely interchange the limit in probability with the integral.

To compute the integrand $\langle \boldsymbol{m}^k, \nabla \widetilde{H}_N(t\boldsymbol{m}^k) \rangle_N$ we analyze a modified AMP which agrees with the AMP we have considered so far up to step k, whereupon we replace the non-linearity $f_k(\boldsymbol{h}, \boldsymbol{x}^k) =$

$f(\boldsymbol{x}^k + \boldsymbol{h})$ by

$$\tilde{f}_k(\boldsymbol{h}, \boldsymbol{x}^k) \equiv t \cdot f_k(\boldsymbol{x}^k)$$

for arbitrary $t \in (0,1)$. We obtain the new iterate

$$\boldsymbol{y}^{k+1}(t) \equiv \nabla \widetilde{H}_N(t\boldsymbol{m}^k) - t\boldsymbol{m}^{k-1}\xi''(t\langle \boldsymbol{m}^k, \boldsymbol{m}^{k-1}\rangle_N)\langle f'(\boldsymbol{x}^k)\rangle_N.$$

Rearranging yields

$$\begin{aligned}\langle \boldsymbol{m}^k, \nabla \widetilde{H}_N(t\boldsymbol{m}^k)\rangle_N &= \langle \boldsymbol{m}^k, \boldsymbol{y}^{k+1}(t)\rangle_N + t\langle \boldsymbol{m}^k, \boldsymbol{m}^{k-1}\rangle_N \xi''(t\langle \boldsymbol{m}^k, \boldsymbol{m}^{k-1}\rangle_N)\langle '(\boldsymbol{x}^{\mathbf{k}})\rangle_{\mathbf{N}} \\ &\simeq \langle \boldsymbol{m}^k, \boldsymbol{y}^{k+1}(t)\rangle_N + ta_{k-1}\xi''(t\phi(a_{k-1}))\langle '(\boldsymbol{x}^{\mathbf{k}})\rangle_{\mathbf{N}}.\end{aligned}$$

We evaluate the $N \to \infty$ limit in probability of the first term, via the state evolution limits $W^k, X^k, Y^{k+1}(t)$. State evolution directly implies

$$\mathbb{E}[W^k Y^{k+1}(t)] = \xi'(t \cdot \mathbb{E}[M^{k-1}M^k]) = \xi'(t\phi(a_{k-1})).$$

Since h is independent of (W^k, Y^{k+1}) we use Gaussian integration by parts to derive

$$\begin{aligned}\mathbb{E}[f(X^k)Y^{k+1}(t)] &= \mathbb{E}[f(h + W^k)Y^{k+1}(t)] \\ &= \mathbb{E}[f'(h + W^k)] \cdot \mathbb{E}[W^k Y^{k+1}(t)] \\ &= \mathbb{E}\left[f'\left(h + Z\sqrt{\xi'(\underline{q})}\right)\right] \cdot \xi'(t\phi(a_{k-1})).\end{aligned}$$

Integrating with respect to t yields

$$\int_0^1 \langle \boldsymbol{m}^k, \nabla \widetilde{H}_N(t\boldsymbol{m}^k)\rangle_N \mathrm{d}t \simeq \mathbb{E}\left[f'\left(h + Z\sqrt{\xi'(\underline{q})}\right)\right] \cdot \int_0^1 \xi'(t\phi(a_{k-1})) + t\phi(a_{k-1})\xi''(t\phi(a_{k-1})) \tag{4.3.11}$$

$$= \mathbb{E}\left[\partial_{xx}\Phi_{\gamma_*}\left(\underline{q}, h + Z\sqrt{\xi'(\underline{q})}\right)\right] \cdot [t\xi'(t\phi(a_{k-1}))]\,|_{t=0}^{t=1} \tag{4.3.12}$$

Finally the first term in (4.3.10) gives energy contribution

$$\begin{aligned}\langle \boldsymbol{h}, \boldsymbol{m}^k\rangle_N &\simeq \mathbb{E}[h \cdot f(Z\sqrt{\xi'(\underline{q})})] \\ &= \mathbb{E}\left[h \cdot \partial_x \Phi_{\gamma_*}\left(\underline{q}, h + Z\sqrt{\xi'(\underline{q})}\right)\right].\end{aligned}$$

Since $\lim_{k\to\infty} a_{k-1} = \xi'(\underline{q})$ and $\psi(\xi'(\underline{q})) = \xi'(\underline{q})$ combining concludes the proof. □

4.3.2 Phase 2: Incremental AMP

We now switch to IAMP, which has a more complicated definition. We will begin from the iterates $\boldsymbol{x}^{\underline{\ell}}, \boldsymbol{m}^{\underline{\ell}}$ from phase 1 for a large $\underline{\ell} \in \mathbb{Z}^+$. We relegate several proofs to Section 4.6. First define the functions

$$u(t,x) = \partial_{xx}\Phi_{\gamma_*}(t,x), \quad v(t,x) = \xi''(t)\gamma_*(t)\partial_x\Phi_{\gamma_*}(t,x).$$

Set $\varepsilon_0 = \frac{\underline{q}}{\phi(a_{\underline{\ell}-1})} - 1$ and $\delta = \xi'(\underline{q}(1+\varepsilon_0)^2) - \xi'(\underline{q})$; observe that $\varepsilon_0, \delta \to 0$ as $\underline{\ell} \to \infty$.[1] Define the sequence $(q_\ell^\delta)_{\ell \geq \underline{\ell}}$ by $q_\ell^\delta = \underline{q} + (\ell - \underline{\ell})\delta$. Fix $\bar{q} \in (\underline{q}, 1)$; the value $\bar{q}$ will be taken close to 1 after sending $\underline{\ell} \to \infty$. In particular we will assume $\delta < 1 - \bar{q}$ holds and set $\bar{\ell} = \min\{\ell \in \mathbb{Z}^+ : q_\ell^\delta \geq \bar{q}\}$. Also define

$$\boldsymbol{n}^{\underline{\ell}} \equiv (1+\varepsilon_0)\boldsymbol{m}^{\underline{\ell}}.$$

Set $\boldsymbol{z}^{\underline{\ell}} = \boldsymbol{w}^{\underline{\ell}}$. So far, we have defined $(\boldsymbol{x}^{\underline{\ell}}, \boldsymbol{z}^{\underline{\ell}}, \boldsymbol{n}^{\underline{\ell}})$. We turn to inductively defining the triples $(\boldsymbol{x}^\ell, \boldsymbol{z}^\ell, \boldsymbol{n}^\ell)$ for $\underline{\ell} \leq \ell \leq \bar{\ell}$. First, the values $(\boldsymbol{z}^\ell)_{\ell \geq \underline{\ell}}$ are defined as AMP iterates via

$$\begin{aligned} \boldsymbol{z}^{\ell+1} &= \nabla \widetilde{H}_N(f_\ell(\boldsymbol{z}^0, \cdots, \boldsymbol{z}^\ell)) - \sum_{j=0}^{\ell} d_{\ell,j} f_{j-1}(\boldsymbol{z}^0, \cdots, \boldsymbol{z}^{j-1}), \\ d_{\ell,j} &= \xi''\left(\mathbb{E}\left[f_\ell(Z^0, \ldots, Z^\ell) f_{j-1}(Z^0, \ldots, Z^{j-1}]\right) \cdot \mathbb{E}\left[\frac{\partial f_\ell}{\partial z^j}(Z^0, \cdots, Z^\ell)\right]\right. . \end{aligned} \tag{4.3.13}$$

(The non-linearities f_ℓ will be specified below). The sequence $(\boldsymbol{x}^{\ell+1})_{\ell \geq \underline{\ell}}$ is defined by

$$\begin{aligned} \boldsymbol{x}^{\ell+1} &\equiv \boldsymbol{x}^{\underline{\ell}} + \sum_{j=\underline{\ell}}^{\ell} v\left(q_j^\delta, \boldsymbol{x}^j\right)\delta + \sum_{j=\underline{\ell}}^{\ell}\left(\boldsymbol{z}^{j+1} - \boldsymbol{z}^j\right) \\ &= \boldsymbol{x}^\ell + v\left(q_\ell^\delta, \boldsymbol{x}^\ell\right)\delta + \left(\boldsymbol{z}^{\ell+1} - \boldsymbol{z}^\ell\right), \quad \underline{\ell} \leq \ell \leq \bar{\ell} - 1. \end{aligned}$$

As usual, $v(q_j^\delta, \cdot)$ is applied component-wise so that $v(q_j^\delta, \boldsymbol{x}^j)_i = v(q_j^\delta, x_i^j)$. Next define the scalar function

$$u_\ell^\delta(x) = \frac{\delta u(q_\ell^\delta, x)}{\left(\xi'(q_\ell^\delta) - \xi'(q_{\ell-1}^\delta)\right)\mathbb{E}\left[u(q_\ell^\delta; X_\ell^\delta)^2\right]}$$

and consider for $\ell \geq \underline{\ell}$ the recursive definition

$$\begin{aligned} \boldsymbol{n}^{\ell+1} &\equiv \boldsymbol{n}^{\underline{\ell}} + \sum_{j=\underline{\ell}}^{\ell} u_j^\delta\left(\boldsymbol{x}^j\right)\left(\boldsymbol{z}^{j+1} - \boldsymbol{z}^j\right) \\ &= \boldsymbol{n}^\ell + u_\ell^\delta\left(\boldsymbol{x}^\ell\right)\left(\boldsymbol{z}^{\ell+1} - \boldsymbol{z}^\ell\right). \end{aligned} \tag{4.3.14}$$

[1] When $\underline{q} = 0$, ε_0 is not defined. In this case we simply take $\delta > 0$ small and begin IAMP at $\boldsymbol{n}^{\underline{\ell}} = (\sqrt{\delta}, \sqrt{\delta}, \ldots, \sqrt{\delta})$.

We define the non-linearity $f_\ell : \mathbb{R}^{\ell+1} \to \mathbb{R}$ to recursively satisfy

$$f_\ell(\boldsymbol{z}^0, \dots, \boldsymbol{z}^\ell) = \boldsymbol{n}^\ell, \quad \ell > \underline{\ell}.$$

It is not difficult to verify that the equations above form a "closed loop" uniquely determining the sequence $(\boldsymbol{x}^\ell, \boldsymbol{z}^\ell, \boldsymbol{n}^\ell)_{\ell \geq \underline{\ell}}$. Since (x_i^ℓ, n_i^ℓ) is determined by the sequence $(z_i^{\underline{\ell}}, \dots, z_i^\ell)$ we may define the state evolution limiting random variables $(X_\ell^\delta, N_\ell^\delta, Z_\ell^\delta)_{\ell \geq \underline{\ell}}$. We emphasize that the IAMP just defined is part of the same $\underline{q}$-AMP algorithm as the first phase defined in the previous subsection. However the variable naming has changed so that the main iterates are $\boldsymbol{z}^\ell$ for $\ell \geq \underline{\ell}$ rather than $\boldsymbol{w}^\ell$ for $\ell \leq \underline{\ell}$. In particular there is no problem in applying state evolution even though the two AMP phases take different forms.

To complete the algorithm, we output the coordinate-wise sign $\boldsymbol{\sigma} = \mathrm{sign}(\boldsymbol{n}^{\bar{\ell}})$ where

$$\mathrm{sign}(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x \leq 0. \end{cases}$$

The key to analyzing the AMP algorithm above is an SDE description in the $\delta \to 0$ limit. Define the filtration

$$\mathcal{F}_\ell^\delta = \sigma\left((Z_k^\delta, N_k^\delta)_{0 \leq k \leq \ell}\right) \tag{4.3.15}$$

for the state evolution limiting process.

Lemma 4.3.6. *The sequences $(Z_{\underline{\ell}}^\delta, Z_{\underline{\ell}+1}^\delta, \dots)$ and $(N_{\underline{\ell}}^\delta, N_{\underline{\ell}+1}^\delta, \dots)$ satisfy for each $\ell \geq \underline{\ell}$:*

$$\begin{aligned}
\mathbb{E}[(Z_{\ell+1}^\delta - Z_\ell^\delta) Z_j^\delta] &= 0, \quad \textit{for all } \underline{\ell} + 1 \leq j \leq \ell \\
\mathbb{E}[(Z_{\ell+1}^\delta - Z_\ell^\delta)^2 | \mathcal{F}_\ell^\delta] &= \xi'(q_{\ell+1}^\delta) - \xi'(q_\ell^\delta) \\
\mathbb{E}[(Z_\ell^\delta)^2] &= \xi'(q_\ell^\delta)] \\
\mathbb{E}[(N_{\ell+1}^\delta - N_\ell^\delta) | \mathcal{F}_\ell^\delta] &= 0 \\
\mathbb{E}[(N_{\ell+1}^\delta - N_\ell^\delta)^2] &= \delta \\
\mathbb{E}[(N_\ell^\delta)^2] &= q_{\ell+1}^\delta.
\end{aligned}$$

From Lemma 4.3.6 and the fact that $(Z_{\underline{\ell}}^\delta, Z_{\underline{\ell}+1}^\delta, \dots)$ form a Gaussian process, it follows that there is a coupling with a standard Brownian motion $(B_t)_{t \in [0,1]}$ such that $\int_0^{q_\ell^\delta} \sqrt{\xi''(t)} \mathrm{d}B_t = Z_\ell^\delta$ for each ℓ. Denote by $(\mathcal{F}_t)_{t \in [0,1]}$ the associated natural filtration. Recall that X_t is defined as the solution to the SDE

$$dX_t = \gamma_*(t) \partial_x \Phi_{\gamma_*}(t, X_t) \mathrm{d}t + \sqrt{\xi''(t)} \mathrm{d}B_t$$

with initialization $X_0 \sim \mathcal{L}_h$. Recalling Proposition 4.2.7, define processes $(N_t, Z_t)_{t\in[0,1]}$ by

$$\begin{aligned} N_t &\equiv \partial_x \Phi_{\gamma_*}(t, X_t) \\ &= \partial_x \Phi_{\gamma_*}(\underline{q}, X_{\underline{q}}) + \int_{\underline{q}}^t \sqrt{\xi''(s)} u(s, X_s) dB_s, \\ Z_t &\equiv \int_0^t \sqrt{\xi''(s)} dB_s. \end{aligned}$$

The next lemma states that these continuous-time processes are the $\delta \to 0$ limit of $(X_\ell^\delta, N_\ell^\delta, Z_\ell^\delta)_{\ell \geq \underline{\ell}}$.

Lemma 4.3.7. *Fix $\bar{q} \in (\underline{q}, 1)$. There exists a coupling between the families of triples $\{(Z_\ell^\delta, X_\ell^\delta, N_\ell^\delta)\}_{\ell\geq 0}$ and $\{(Z_t, X_t, N_t)\}_{t\geq 0}$ such that the following holds. For some $\delta_0 > 0$ and constant $C > 0$, for every $\delta \leq \delta_0$ and $\ell \geq \underline{\ell}$ with $q_\ell \leq \bar{q}$ we have*

$$\begin{aligned} \max_{\underline{\ell} \leq j \leq \ell} \mathbb{E}\left[\left(X_j^\delta - X_{q_j}\right)^2\right] &\leq C\delta, \\ \max_{\underline{\ell} \leq j \leq \ell} \mathbb{E}\left[\left(N_j^\delta - N_{q_j}\right)^2\right] &\leq C\delta. \end{aligned}$$

Lemmas 4.3.6 and 4.3.7 are proved in Section 4.6.

4.3.3 Computing the Final Energy

In this subsection we establish Theorem 37 by showing $\lim_{\bar{q}\to 1} \lim_{\underline{\ell}\to\infty} \text{p-lim}_{N\to\infty} \frac{H_N(\boldsymbol{\sigma})}{N} = \mathsf{P}(\gamma_*)$. First we show that the replacements $\boldsymbol{m}^{\underline{\ell}} \to \boldsymbol{n}^{\underline{\ell}}$ and $\boldsymbol{n}^{\overline{\ell}} \to \boldsymbol{\sigma}$ have negligible effect on the asymptotic value attained.

Lemma 4.3.8.

$$\lim_{\bar{q}\to 1} \lim_{\underline{\ell}\to\infty} \text{p-lim}_{N\to\infty} \left| \frac{H_N(\boldsymbol{\sigma}) - H_N\left(\boldsymbol{n}^{\overline{\ell}}\right)}{N} \right| = 0, \tag{4.3.16}$$

$$\lim_{\underline{\ell}\to\infty} \text{p-lim}_{N\to\infty} \left| \frac{H_N(\boldsymbol{m}^{\underline{\ell}}) - H_N(\boldsymbol{n}^{\underline{\ell}})}{N} \right| = 0. \tag{4.3.17}$$

Proof. Proposition 4.2.6 implies that $N_t \in [-1, 1]$ almost surely, while optimizability of γ_* implies that $\mathbb{E}[(N_t)^2] = t$ for $t \in [\underline{q}, \bar{q}]$. It follows that

$$\begin{aligned} \lim_{\bar{q}\to 1} \lim_{\underline{\ell}\to\infty} \text{p-lim}_{N\to\infty} \|\boldsymbol{n}^{\overline{\ell}} - \text{sign}(\boldsymbol{n}^{\overline{\ell}})\|_N &= \lim_{\bar{q}\to 1} \sqrt{\mathbb{E}\left[\left(N_{\bar{q}} - \text{sign}(N_{\bar{q}})\right)^2\right]} \\ &= 0. \end{aligned}$$

The limit (4.3.16) follows from Proposition 4.2.1 with $k = 1$. (4.3.17) follows similarly as

$$\boldsymbol{n}^{\underline{\ell}} - \boldsymbol{m}^{\underline{\ell}} = \varepsilon_0 \boldsymbol{m}^{\underline{\ell}}$$

and $\varepsilon_0 \to 0$ as $\underline{\ell} \to \infty$ while $\text{p-lim}_{N\to\infty} \|\boldsymbol{m}^{\underline{\ell}}\|_N \le 1$ thanks to Proposition 4.2.6. □

In the next lemma, proved in Section 4.6, we compute the energy gain during IAMP.

Lemma 4.3.9.

$$\lim_{\bar{q}\to 1} \lim_{\underline{\ell}\to\infty} \text{p-lim}_{N\to\infty} \frac{H_N\left(\boldsymbol{n}^{\bar{\ell}}\right) - H_N\left(\boldsymbol{n}^{\underline{\ell}}\right)}{N} = \int_{\underline{q}}^1 \xi''(t)\mathbb{E}\left[u\left(t, X_t\right)\right] \mathrm{d}t. \tag{4.3.18}$$

We now put everything together. Recall from Lemma 4.3.5 that

$$\lim_{\underline{\ell}\to\infty} \text{p-lim}_{N\to\infty} \frac{H_N(\boldsymbol{m}^{\underline{\ell}})}{N} = \xi'(\underline{q}) \cdot \mathbb{E}\left[\partial_{xx}\Phi_{\gamma_*}\left(\underline{q}, X_{\underline{q}}\right)\right] + \mathbb{E}\left[h \cdot \partial_x \Phi_{\gamma_*}\left(\underline{q}, X_{\underline{q}}\right)\right].$$

Proposition 4.2.7 implies that the process $\partial_x \Phi_{\gamma_*}(t, X_t)$ is a martingale, while Lemma 4.2.8 states that $\mathbb{E}[u(t, X_t)] = \mathbb{E}[\partial_{xx}\Phi_{\gamma_*}(t, X_t)] = \int_t^1 \gamma_*(s)\mathrm{d}s$. Substituting, we find

$$\lim_{\underline{\ell}\to\infty} \text{p-lim}_{N\to\infty} \frac{H_N(\boldsymbol{m}^{\underline{\ell}})}{N} = \xi'(\underline{q}) \int_{\underline{q}}^1 \gamma_*(s)\mathrm{d}s + \mathbb{E}[h\partial_x \Phi_{\gamma_*}(0, h)].$$

Using again that $\mathbb{E}[u(t, X_t)] = \int_t^1 \gamma_*(s)\mathrm{d}s$, the right-hand side of (4.3.18) is

$$\begin{aligned}
\lim_{\bar{q}\to 1} \lim_{\underline{\ell}\to\infty} \text{p-lim}_{N\to\infty} \frac{H_N\left(\boldsymbol{n}^{\bar{\ell}}\right) - H_N\left(\boldsymbol{n}^{\underline{\ell}}\right)}{N} &= \int_{\underline{q}}^1 \xi''(t) \int_t^1 \gamma_*(s)\mathrm{d}s\mathrm{d}t \\
&= \int_{\underline{q}}^1 \int_{\underline{q}}^s \xi''(t)\gamma_*(s)\mathrm{d}t\mathrm{d}s \\
&= \int_{\underline{q}}^1 (\xi'(s) - \xi'(\underline{q}))\gamma_*(s)\mathrm{d}s \\
&= \int_0^1 \xi'(s)\gamma_*(s)\mathrm{d}s - \xi'(\underline{q}) \int_{\underline{q}}^1 \gamma_*(s)\mathrm{d}s.
\end{aligned}$$

Combining with Lemma 4.3.8 yields

$$\begin{aligned}\lim_{\bar{q}\to 1}\lim_{\underline{\ell}\to\infty}\operatorname*{p-lim}_{N\to\infty}\frac{H_N(\boldsymbol{\sigma})}{N} &= \lim_{\bar{q}\to 1}\lim_{\underline{\ell}\to\infty}\operatorname*{p-lim}_{N\to\infty}\frac{1}{N}\cdot\Bigg(H_N\left(\operatorname{sign}\left(\boldsymbol{n}^{\bar{\ell}}\right)\right)-H_N\left(\boldsymbol{n}^{\bar{\ell}}\right)\\ &\qquad + H_N\left(\boldsymbol{n}^{\bar{\ell}}\right)-H_N\left(\boldsymbol{n}^{\underline{\ell}}\right)+H_N\left(\boldsymbol{n}^{\underline{\ell}}\right)-H_N\left(\boldsymbol{m}^{\underline{\ell}}\right)+H_N\left(\boldsymbol{m}^{\underline{\ell}}\right)\Bigg)\\ &= \mathbb{E}^{h\sim\mathcal{L}_h}[h\cdot\partial_x\Phi_{\gamma_*}(0,h)]+\int_0^1\xi'(s)\gamma_*(s)\mathrm{d}s. \end{aligned}\tag{4.3.19}$$

Having computed the limiting energy achieved by our $\underline{q}$-AMP algorithm, it remains to verify that the value in Equation (4.3.19) is equal to $\mathsf{P}_{\xi,h}(\gamma_*)$. Define

$$\Psi_{\gamma_*}(t,x)=\Phi_{\gamma_*}(t,x)-x\partial_x\Phi_{\gamma_*}(t,x)$$

for $(t,x)\in[0,1]\times\mathbb{R}$.

Lemma 4.3.10. *For $h\sim\mathcal{L}_h$, $\bar{q}\geq\underline{q}$, and X_t as in (4.1.3),*

$$\begin{aligned}\mathbb{E}[\Phi_{\gamma_*}(0,h)] &= \mathbb{E}[h\cdot\partial_x\Phi_{\gamma_*}(0,h)]+\mathbb{E}\left[\Psi_{\gamma_*}(\bar{q},X_{\bar{q}})\right]\\ &\quad+\frac{1}{2}\int_0^{\bar{q}}\xi''(t)\gamma_*(t)\,\mathbb{E}\left[\partial_x\Phi_{\gamma_*}(t,X_t)^2\right]\,\mathrm{d}t+\int_0^{\bar{q}}\xi''(t)\,\mathbb{E}\left[\partial_{xx}\Phi_{\gamma_*}(t,X_t)\right]\,\mathrm{d}t.\end{aligned}$$

Proof. We write

$$\begin{aligned}\mathbb{E}[\Psi_{\gamma_*}(\bar{q},X_{\bar{q}})-\Psi_{\gamma_*}(0,X_0)] &= \int_0^{\bar{q}}\frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}\left[\Psi_{\gamma_*}(s,X_s)\right]|_{s=t}\mathrm{d}t\\ &= \int_0^{\bar{q}}\frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}\left[\Phi_{\gamma_*}(s,X_s)-X_s\partial_x\Phi_{\gamma_*}(s,X_s)\right]|_{s=t}\mathrm{d}t\\ &= -\frac{1}{2}\int_0^{\bar{q}}\xi''(t)\gamma_*(t)\,\mathbb{E}\left[\partial_x\Phi_{\gamma_*}(t,X_t)^2\right]\mathrm{d}t\\ &\quad-\int_0^{\bar{q}}\xi''(t)\,\mathbb{E}\left[\partial_{xx}\Phi_{\gamma_*}(t,X_t)\right]\mathrm{d}t.\end{aligned}$$

Rearranging shows:

$$\begin{aligned}\mathbb{E}[\Phi_{\gamma_*}(0,X_0)] &= \mathbb{E}[X_0\partial_x\Phi_{\gamma_*}(0,X_0)]+\mathbb{E}[\Psi_{\gamma_*}(\bar{q},X_{\bar{q}})]\\ &\quad+\frac{1}{2}\int_0^{\bar{q}}\xi''(t)\gamma_*(t)\,\mathbb{E}\left[\partial_x\Phi_{\gamma_*}(t,X_t)^2\right]\mathrm{d}t+\int_0^{\bar{q}}\xi''(t)\,\mathbb{E}\left[\partial_{xx}\Phi_{\gamma_*}(t,X_t)\right]\mathrm{d}t.\end{aligned}$$

As $X_0=h$ this concludes the proof. □

Proof of Theorem 37. Note that $\gamma_*(t) > 0$ implies $t \geq \underline{q}$ and hence by optimizability

$$\mathbb{E}[(\partial_x \Phi_{\gamma_*}(t, X_t))^2] = t.$$

Meanwhile for any $t \in [0, 1]$,

$$\mathbb{E}[\partial_{xx} \Phi_{\gamma_*}(t, X_t)] = \int_t^1 \gamma_*(t) \mathrm{d}t.$$

Therefore

$$\Phi_{\gamma_*}(0, h) = h\partial_x \Phi_{\gamma_*}(0, h) + \mathbb{E}\left[\Psi_{\gamma_*}(\bar{q}, X_{\bar{q}})\right] + \frac{1}{2}\int_0^{\bar{q}} \xi''(t)\gamma_*(t)t \,\mathrm{d}t + \int_0^{\bar{q}} \xi''(t) \int_t^1 \gamma_*(s)\mathrm{d}s \,\mathrm{d}t\,.$$

Recalling (4.1.4), we find

$$\begin{aligned}\mathsf{P}(\gamma_*) &= \mathbb{E}[\Phi_{\gamma_*}(0, h)] - \frac{1}{2}\int_0^1 \xi''(t)\gamma_*(t)t\mathrm{d}t \\ &= h\partial_x \Phi_{\gamma_*}(0, h) + \mathbb{E}\left[\Psi_{\gamma_*}(\bar{q}, X_{\bar{q}})\right] + \int_0^{\bar{q}} \xi''(t) \int_t^1 \gamma_*(s)\mathrm{d}s \,\mathrm{d}t \; + o_{\bar{q}\to 1}(1).\end{aligned}$$

It is not hard to show that $\lim_{\bar{q}\to 1} \Psi_{\gamma_*}(\bar{q}, x) = 0$ holds uniformly in x. Moreover

$$\begin{aligned}\lim_{\bar{q}\to 1} \int_0^{\bar{q}} \xi''(t) \int_t^1 \gamma_*(s)\mathrm{d}s \,\mathrm{d}t &= \int_0^1 \xi''(t) \int_t^1 \gamma_*(s)\mathrm{d}s \,\mathrm{d}t \\ &= \int_0^1 \int_0^s \xi''(t)\gamma_*(s)\mathrm{d}t \,\mathrm{d}s \\ &= \int_0^1 \xi'(s)\gamma_*(s)\mathrm{d}s.\end{aligned}$$

Combining the above and comparing with (4.3.19) yields

$$\begin{aligned}\mathsf{P}(\gamma_*) &= \mathbb{E}\left[h\partial_x \Phi_{\gamma_*}(0, h)\right] + \int_0^1 \xi'(s)\gamma_*(s)\mathrm{d}s \\ &= \lim_{\bar{q}\to 1} \lim_{\underline{\ell}\to\infty} \operatorname*{p\text{-}lim}_{N\to\infty} \frac{H_N(\boldsymbol{\sigma})}{N}.\end{aligned}$$

This completes the proof of Theorem 37. □

4.4 Constructing Many Approximate Maximizers

Here we explain the modifications needed for branching IAMP and Theorem 17. The proofs are a slight extension of those for the main algorithm, and in fact we give many proofs for IAMP directly in this more general setting in Section 4.6. Let us fix values $Q = (q_1, \ldots, q_m)$ with $\underline{q} \leq q_1 < \cdots <$

$q_m < 1$ and an index $B \in [m]$. To construct a pair of approximate maximizers with overlap q_B we first construct $\boldsymbol{n}^{\underline{\ell}}$ exactly as in Subsection 4.3.1. For each $i < B$, set $\mathbf{g}^{(q_i,1)} = \mathbf{g}^{(q_i,2)} = \boldsymbol{z}^{-k_{i,1}} = \boldsymbol{z}^{-k_{i,2}} \in \mathbb{R}^N$ for some $k_{i,1} = k_{i,2} \le K$ as in Subsection 4.2.2. For each $B \le i \le m$, set $\mathbf{g}^{(q_i,1)} = \boldsymbol{z}^{-k_{i,1}}$ and $\mathbf{g}^{(q_i,2)} = \boldsymbol{z}^{-k_{i,2}}$ where $k_{i,1} \neq k_{i,2}$. Because the vectors $\mathbf{g}^{(q_i,a)}$ are constructed using AMP, we require some additional conditions. We insist that $k_{i,a'} - \ell^{\delta}_{q_i} > k_{j,a} - \ell^{\delta}_{q_j} > 0$ for any $i > j$ and $a, a' \in \{1,2\}$, which is satisfied by choosing the values $k_{i,a}$ in increasing order of i. Finally we insist that $\max_{i,a}(k_{i,a}) + \overline{\ell} + 1 < K$, where $h = \boldsymbol{z}^{-K}$ was the AMP initialization, which is satisfied by choosing K large at the end.

Having fixed this setup, we proceed by defining $\boldsymbol{m}^{k,1} = \boldsymbol{m}^{k,2} = \boldsymbol{m}^k$ exactly as in the original first phase. Next we generate two sequences of IAMP iterates using (4.3.14) except at times corresponding to $q_i \in Q$, altogether generating $\boldsymbol{n}^{\ell,a}$ for $\ell > \underline{\ell}$ and $a \in \{1,2\}$ via:

$$\boldsymbol{n}^{\ell,a} = \begin{cases} \boldsymbol{n}^{\ell-1,a} + \sqrt{\delta}\mathbf{g}^{(q_i,a)}, & \ell = \ell^{\delta}_{q_i} \equiv \underline{\ell} + \lceil (q_i - \underline{q})\delta \rceil + 1 \ \text{ for some } i \in [m] \\ \boldsymbol{n}^{\ell-1,a} + u^{\delta}_{\ell-1}\left(\boldsymbol{x}^{\ell-1,a}\right)\left(\boldsymbol{z}^{\ell} - \boldsymbol{z}^{\ell-1,a}\right), & \text{else.} \end{cases} \tag{4.4.1}$$

The definitions of $\boldsymbol{x}^{\ell,a}, \boldsymbol{z}^{\ell,a}$ are the same as before. The following result follows immediately from Lemmas 4.6.5, 4.6.6 and readily implies Theorem 17.

Lemma 4.4.1. *For optimizable γ_*,*

$$\lim_{\overline{q}\to 1}\lim_{\underline{\ell}\to\infty}\operatorname*{p-lim}_{N\to\infty}\frac{H_N(\boldsymbol{n}^{\overline{\ell},a}_{\delta})}{N} = \mathsf{P}(\gamma_*), \quad a \in \{1,2\}$$
$$\lim_{\overline{q}\to 1}\lim_{\underline{\ell}\to\infty}\operatorname*{p-lim}_{N\to\infty}\frac{\langle \boldsymbol{n}^{\overline{\ell},1}_{\delta}, \boldsymbol{n}^{\overline{\ell},2}_{\delta}\rangle}{N} = q_B.$$

Theorem 17. *Let $\gamma_* \in \mathscr{L}$ be optimizable, and fix a finite ultrametric space (X, d_X) with diameter at most $\sqrt{2(1-\underline{q})}$ as well as $\varepsilon > 0$. Then an efficient AMP algorithm constructs points $\{\boldsymbol{\sigma}_x | x \in X\}$ in $\boldsymbol{\Sigma}_N$ satisfying*

$$\frac{H_N(\boldsymbol{\sigma}_x)}{N} \in [\mathsf{P}(\gamma_*) - \varepsilon, \mathsf{P}(\gamma_*) + \varepsilon], \quad x \in X,$$
$$\frac{\|\boldsymbol{\sigma}_x - \boldsymbol{\sigma}_y\|}{\sqrt{N}} \in [d_X(x,y) - \varepsilon, d_X(x,y) + \varepsilon], \quad x, y \in X$$

with probability tending to 1 as $N \to \infty$.

Proof. Recall that any finite ultrametric space X with all pairwise distances in the set $\{\sqrt{2(1-q_i)}\}_{i\in[m]}$ can be identified with a rooted tree $\mathcal{T}$ whose leaves $\partial\mathcal{T}$ are in bijection with X, and so that

$d_X(x_i, x_j) = \sqrt{2(1-q_k)}$ is equivalent to leaves i, j having least common ancestor at depth k. Given $\mathcal{T}$, we may assign to each non-root vertex $u \in \mathcal{T}$ a distinct initialization iterate $\mathbf{g}^{(u)} = \boldsymbol{z}^{-k_u}$, where $k_u < k_{u'}$ if $\mathrm{depth}(u) < \mathrm{depth}(u')$ and again $\max_u(k_u) + \overline{\ell} + 1 < K$. Then for each path $\mathrm{root} = v_0, v_1, \ldots, v_m = x \in \partial\mathcal{T} = X$, we compute the iteration (4.4.1) using $\mathbf{g}^{(k_{v_1})}, \ldots, \mathbf{g}^{(k_{v_m})}$. Applying Lemma 4.4.1 over all pairs of leaves $(x, y) \in X \times X$ implies that the AMP iterates $\boldsymbol{n}_\delta^{\overline{\ell},x}$ satisfy $\frac{H_N(\boldsymbol{n}_\delta^{\overline{\ell},x})}{N} \simeq \mathsf{P}(\gamma_*)$ and $\langle \boldsymbol{n}_\delta^{\overline{\ell},x}, \boldsymbol{n}_\delta^{\overline{\ell},y} \rangle_N \simeq q_j$ if $d_X(x, y) = \sqrt{2(1-q_j)}$. The conclusion follows by rounding $\boldsymbol{n}_\delta^{\overline{\ell},x} \to \boldsymbol{\sigma}_x \in \boldsymbol{\Sigma}_N$ for each $x \in X$ as in the main algorithm. □

We remark that our construction differs from the one proposed in [AM20] only because we construct the vectors $\mathbf{g}^{(u)}$ using AMP rather than taking them to be literally independent Gaussian vectors. While the latter construction almost certainly works as well, the analysis seems to require a more general version of state evolution.

4.5 Spherical Models

We now consider the case of spherical spin glasses with external field. The law of the Hamiltonian H_N is specified according to the same formula as before depending on $(\xi, \mathcal{L})$, however the state space is the sphere $\mathbb{S}^{N-1}(\sqrt{N})$ instead of the hypercube. The free energy in this case is given by a similar Parisi-type formula, however it turns out to dramatically simplify under no overlap gap so we do not give the general formula. At positive temperature the spherical free energy was computed non-rigorously in [CS92] and rigorously in [Tal06b, Che13b], but we rely on [CS17] which directly treats the zero-temperature setting.

Remark 4.5.1. Due to rotational invariance, for spherical models all that matters about $\mathcal{L}_h$ is the squared L^2 norm $\mathbb{E}^{h\sim\mathcal{L}_h}[h^2]$. In particular unlike the Ising case there is no loss of generality in assuming h is constant. We continue to work with coordinates h_i sampled i.i.d. from $\mathcal{L}_h$ and implicitly use this observation when interpreting the results of [CS17].

Our treatment of spherical models is less detailed and we simply show how to obtain the energy value in Theorem 18 which is the ground state in models with $\mathrm{supp}(\gamma_*) = [q, 1)$. In the case that $\mathbb{E}[h^2] + \xi'(1) < \xi''(1)$, we let $\underline{q}_{\mathrm{sph}} \in [0, 1]$ be the unique solution to

$$\underline{q}_{\mathrm{sph}}\xi''(\underline{q}_{\mathrm{sph}}) = \mathbb{E}[h^2] + \xi'(\underline{q}_{\mathrm{sph}}).$$

When $\mathbb{E}[h^2] + \xi'(1) \geq \xi''(1)$, we simply set $\underline{q}_{\mathrm{sph}} = 1$.

Note that when $h = 0$ almost surely it follows that $\underline{q}_{\mathrm{sph}} = 0$, which is the setting of [Sub21]. Generate initial iterates

$$(\boldsymbol{w}_{\mathrm{sph}}^{-K}, \ldots, \boldsymbol{w}_{\mathrm{sph}}^{0})$$

as in Subsection 4.2.2. For non-zero h we take $c = \sqrt{\mathbb{E}[h^2] + \xi'(\underline{q}_{\text{sph}})}$ so that

$$\|\boldsymbol{w}^0_{\text{sph}}\|_N \simeq \sqrt{\mathbb{E}[h^2] + \xi'(\underline{q}_{\text{sph}})}.$$

Generate further iterates via the following AMP iteration.

$$\begin{aligned} \boldsymbol{w}^{k+1}_{\text{sph}} &= \nabla \widetilde{H}_N(\boldsymbol{m}^k_{\text{sph}}) - \boldsymbol{m}^{k-1}_{\text{sph}} \xi''\left(\langle \boldsymbol{m}^k_{\text{sph}}, \boldsymbol{m}^{k-1}_{\text{sph}} \rangle\right) \sqrt{\frac{\underline{q}_{\text{sph}}}{\mathbb{E}[h^2] + \xi'(\underline{q}_{\text{sph}})}} \\ \boldsymbol{x}^k_{\text{sph}} &= \boldsymbol{w}^k_{\text{sph}} + \boldsymbol{h} \\ \boldsymbol{m}^k_{\text{sph}} &= \boldsymbol{x}^k_{\text{sph}} \cdot \sqrt{\frac{\underline{q}_{\text{sph}}}{\mathbb{E}[h^2] + \xi'(\underline{q}_{\text{sph}})}}. \end{aligned} \tag{4.5.1}$$

The next lemma is the spherical analog of Lemmas 4.3.3, 4.3.4, 4.3.5 - the proof is similar to the Ising case and is given in the next subsection.

Lemma 4.5.1. *Using the AMP of* (4.5.1)*, the asymptotic overlaps and energies satisfy*

$$\begin{aligned} \lim_{k,\ell\to\infty} \text{p-lim}_{N\to\infty} \frac{\langle \boldsymbol{w}^k_{sph}, \boldsymbol{w}^\ell_{sph} \rangle}{N} &= \xi'(\underline{q}_{sph}), \\ \lim_{k,\ell\to\infty} \text{p-lim}_{N\to\infty} \frac{\langle \boldsymbol{x}^k_{sph}, \boldsymbol{x}^\ell_{sph} \rangle}{N} &= \mathbb{E}[h^2] + \xi'(\underline{q}_{sph}), \\ \lim_{k,\ell\to\infty} \text{p-lim}_{N\to\infty} \frac{\langle \boldsymbol{m}^k_{sph}, \boldsymbol{m}^\ell_{sph} \rangle}{N} &= \underline{q}_{sph}, \\ \lim_{k,\ell\to\infty} \text{p-lim}_{N\to\infty} \frac{H_N(\boldsymbol{m}^k_{sph})}{N} &= \sqrt{\underline{q}_{sph}(\mathbb{E}[h^2] + \xi'(\underline{q}_{sph}))}. \end{aligned} \tag{4.5.2}$$

Proof of Theorem 19. The latter two parts of Lemma 4.5.1 directly imply Theorem 19 in the case that $\mathbb{E}[h^2] + \xi'(1) \geq \xi''(1)$ (recall $\underline{q}_{\text{sph}} = 1$ in this case). Indeed, it suffices to take

$$\boldsymbol{\sigma}_{\text{sph}} = \frac{\boldsymbol{m}^{\overline{\ell}}_{\text{sph}}}{\|\boldsymbol{m}^{\overline{\ell}}_{\text{sph}}\|_N} \in \mathbb{S}^{N-1}(\sqrt{N}) \tag{4.5.3}$$

for a sufficiently large constant $\overline{\ell}$. When $\mathbb{E}[h^2] + \xi'(1) < \xi''(1)$, we can conclude by mimicking the IAMP phase using the simple non-linearities $u(t,x) = u(t) = \xi''(t)^{-1/2}$ and $v(t,x) = 0$ - see also [AMS21, Remark 2.2]. Lemma 4.3.9 then shows the energy gain from IAMP is

$$\int_{\underline{q}}^1 \xi''(t) u(t) \mathrm{d}t = \int_{\underline{q}}^1 \xi''(t)^{1/2}.$$

As in the Ising case, we may start IAMP from $\boldsymbol{m} = \boldsymbol{m}^k$ for a large constant k. Combining with (4.5.2)

and defining $\boldsymbol{\sigma}_{\rm sph}$ via (4.5.3) with $\boldsymbol{m}^{\overline{\ell}}$ an IAMP iterate, we obtain

$$\operatorname*{p-lim}_{N\to\infty} \frac{H_N(\boldsymbol{\sigma}_{\rm sph})}{N} \geq \underline{q}_{\rm sph}\xi''(\underline{q}_{\rm sph})^{1/2} + \int_{\underline{q}_{\rm sph}}^1 \xi''(\underline{q}_{\rm sph})^{1/2}\mathrm{d}q.$$

Alternatively to IAMP, in the spherical setting it is possible to use the approach of [Sub21]. Indeed [Sub21, Theorem 4] immediately extends to an algorithm taking in an arbitrary point $\boldsymbol{m}$ with $\|\boldsymbol{m}\|_N \leq 1$ and outputting a point $\boldsymbol{m}_* \in \mathbb{S}^{N-1}(\sqrt{N})$ (which may depend on H_N) satisfying

$$\frac{H_N(\boldsymbol{m}_*) - H_N(\boldsymbol{m})}{N} \geq \int_{\|\boldsymbol{m}\|_N^2}^1 \sqrt{\xi''(q)}\mathrm{d}q - \varepsilon$$

with probability $1 - o_{N\to\infty}(1)$ for any desired $\varepsilon > 0$. Either approach completes the proof of Theorem 19. □

4.5.1 Proof of Lemma 4.5.1

For $t \in [0, \xi'(\underline{q}_{\rm sph})]$, take $h \sim \mathcal{L}_h$ and $(Z, Z', Z'') \sim \mathsf{N}(0, I_3)$ and define the function

$$\begin{aligned}\phi_{\rm sph}(t) &= \frac{\underline{q}_{\rm sph}}{\mathbb{E}[h^2] + \xi'(\underline{q}_{\rm sph})} \cdot \mathbb{E}^{h,Z,Z',Z''}\left[\left(h + Z\sqrt{t} + Z'\sqrt{\xi'(\underline{q}_{\rm sph}) - t}\right)\left(h + Z\sqrt{t} + Z''\sqrt{\xi'(\underline{q}_{\rm sph}) - t}\right)\right] \\ &= \frac{\underline{q}_{\rm sph}(\mathbb{E}[h^2] + t)}{\mathbb{E}[h^2] + \xi'(\underline{q}_{\rm sph})}.\end{aligned}$$

so that $\phi_{\rm sph}(\xi'(\underline{q}_{\rm sph})) = \underline{q}_{\rm sph}$. Define $\psi_{\rm sph}(t) = \xi'(\phi_{\rm sph}(t))$.

Lemma 4.5.2. *ψ_{sph} is strictly increasing and convex on $[0, \xi'(\underline{q}_{sph})]$ and*

$$\psi_{sph}(\xi'(\underline{q}_{sph})) = \xi'(\underline{q}_{sph}), \tag{4.5.4}$$

$$\psi'_{sph}(\xi'(\underline{q}_{sph})) = 1, \tag{4.5.5}$$

$$\psi_{sph}(t) > t, \qquad \forall t < \xi'(\underline{q}_{sph}). \tag{4.5.6}$$

Proof. Since ξ' is strictly increasing and convex and $\phi_{\rm sph}$ is affine and increasing, it follows that $\psi_{\rm sph}$ is strictly increasing and convex. (4.5.4) is equivalent to the equation $\underline{q}_{\rm sph}\xi''(\underline{q}_{\rm sph}) = \mathbb{E}[h^2] + \xi'(\underline{q}_{\rm sph})$ defining $\underline{q}_{\rm sph}$. To show (4.5.5) we use the chain rule to write

$$\psi'_{\rm sph}(\xi'(\underline{q}_{\rm sph})) = \xi''(\phi_{\rm sph}(\xi'(\underline{q}_{\rm sph}))) \cdot \phi'_{\rm sph}(\xi'(\underline{q}_{\rm sph})) = \xi''(\underline{q}_{\rm sph}) \cdot (\xi''(\underline{q}_{\rm sph}))^{-1} = 1.$$

Equations (4.5.4) and (4.5.5) and the convexity of $\psi_{\rm sph}$ just shown imply (4.5.6) □

Let $h, W_{\rm sph}^{-1}, (W_{\rm sph}^j, X_{\rm sph}^j, M_{\rm sph}^j)_{j=0}^k$ be the state evolution limit of the coordinates of

$$\big(\boldsymbol{h}, \boldsymbol{w}_{\rm sph}^{-1}, \boldsymbol{w}_{\rm sph}^0, \boldsymbol{x}_{\rm sph}^0, \boldsymbol{m}_{\rm sph}^0, \dots, \boldsymbol{w}_{\rm sph}^k, \boldsymbol{x}_{\rm sph}^k, \boldsymbol{m}_{\rm sph}^k\big)$$

as $N \to \infty$. Define the sequence $(b_0, b_1, \dots)$ recursively by $b_0 = 0$ and $b_{k+1} = \psi_{\rm sph}(b_k)$.

Lemma 4.5.3. *For all non-negative integers $0 \le j < k$ the following equalities hold:*

$$\begin{aligned}
\mathbb{E}[(W_{sph}^j)^2] &= \xi'(\underline{q}_{sph}) \\
\mathbb{E}[W_{sph}^j W_{sph}^k] &= b_j \\
\mathbb{E}[(M_{sph}^j)^2] &= \underline{q}_{sph} \\
\mathbb{E}[M_{sph}^j M_{sph}^k] &= \phi_{sph}(b_j).
\end{aligned}$$

Proof. Follows from state evolution and induction exactly as in Lemma 4.3.3. □

Lemma 4.5.4.

$$\begin{aligned}
\lim_{k\to\infty} b_k &= \xi'(\underline{q}_{sph}), \\
\lim_{k\to\infty} \phi_{sph}(b_k) &= \underline{q}_{sph}.
\end{aligned}$$

Proof. As in the proof of Lemma 4.3.4, the sequence $b_1, b_2, \dots$, must converge up to a limit, and this limit must be a fixed point for $\psi_{\rm sph}$, implying the first claim. The second claim follows by continuity of $\phi_{\rm sph}$. □

Lemma 4.5.5.

$$\lim_{k\to\infty} \operatorname*{p-lim}_{N\to\infty} \frac{H_N(\boldsymbol{m}_{sph}^k)}{N} = \sqrt{\underline{q}_{sph}(\mathbb{E}[h^2] + \xi'(\underline{q}_{sph}))}.$$

Proof. We use again the identity

$$\frac{H_N(\boldsymbol{m}_{\rm sph}^k)}{N} = \big\langle \boldsymbol{h}, \boldsymbol{m}_{\rm sph}^k \big\rangle_N + \int_0^1 \big\langle \boldsymbol{m}_{\rm sph}^k, \nabla \widetilde{H}_N(t\boldsymbol{m}_{\rm sph}^k) \big\rangle_N \mathrm{d}t$$

and interchange the limit in probability with the integral. To compute the main term

$$\operatorname*{p-lim}_{N\to\infty} \langle \boldsymbol{m}_{\rm sph}^k, \nabla \widetilde{H}_N((t\boldsymbol{m}_{\rm sph}^k) \rangle$$

we introduce an auxiliary AMP step

$$\boldsymbol{y}_{\rm sph}^{k+1} = \nabla \widetilde{H}_N(t\boldsymbol{m}_{\rm sph}^k) - t\boldsymbol{m}_{\rm sph}^{k-1} \xi''(t\langle \boldsymbol{m}_{\rm sph}^k, \boldsymbol{m}_{\rm sph}^{k-1} \rangle) \sqrt{\frac{\underline{q}_{\rm sph}}{\mathbb{E}[h^2] + \xi'(\underline{q}_{\rm sph})}}\,.$$

Rearranging yields

$$\langle \boldsymbol{m}_{\text{sph}}^k, \nabla \widetilde{H}_N(t\boldsymbol{m}_{\text{sph}}^k)\rangle_N = \langle \boldsymbol{m}_{\text{sph}}^k, \boldsymbol{y}_{\text{sph}}^{k+1}\rangle_N + t\langle \boldsymbol{m}_{\text{sph}}^k, \boldsymbol{m}_{\text{sph}}^{k-1}\rangle_N \xi''(t\langle \boldsymbol{m}_{\text{sph}}^k, \boldsymbol{m}_{\text{sph}}^{k-1}\rangle_N)\sqrt{\frac{\underline{q}_{\text{sph}}}{\mathbb{E}[h^2]+\xi'(\underline{q}_{\text{sph}})}}$$
$$\simeq \langle \boldsymbol{m}_{\text{sph}}^k, \boldsymbol{y}_{\text{sph}}^{k+1}\rangle_N + tb_{k-1}\xi''(t\phi(b_{k-1}))\sqrt{\frac{\underline{q}_{\text{sph}}}{\mathbb{E}[h^2]+\xi'(\underline{q}_{\text{sph}})}}.$$

For the first term, Gaussian integration by parts with

$$g(x) = (x+h)\cdot\sqrt{\frac{\underline{q}_{\text{sph}}}{\mathbb{E}[h^2]+\xi'(\underline{q}_{\text{sph}})}}$$

yields

$$\mathbb{E}[g(X_{\text{sph}}^k)Y^{k+1}] = \mathbb{E}[g'(X_{\text{sph}}^k)]\cdot\mathbb{E}[X_{\text{sph}}^k Y_{\text{sph}}^{k+1}] = \xi'(t\phi_{\text{sph}}(b_{k-1}))\sqrt{\frac{\underline{q}_{\text{sph}}}{\mathbb{E}[h^2]+\xi'(\underline{q}_{\text{sph}})}}.$$

Integrating with respect to t, we find

$$\int_0^1 \langle \boldsymbol{m}_{\text{sph}}^k, \nabla \widetilde{H}_N(t\boldsymbol{m}_{\text{sph}}^k)\rangle_N \mathrm{d}t \simeq \mathbb{E}\left[g'\left(Z\sqrt{\xi'(\underline{q}_{\text{sph}})}\right)\right]\cdot\int_0^1 \xi'(t\phi_{\text{sph}}(b_{k-1})) + t\phi_{\text{sph}}(b_{k-1})\xi''(t\phi_{\text{sph}}(b_{k-1}))$$
$$= [t\xi'(t\phi_{\text{sph}}(b_{k-1}))]\,|_{t=0}^{t=1}\cdot\sqrt{\frac{\underline{q}_{\text{sph}}}{\mathbb{E}[h^2]+\xi'(\underline{q}_{\text{sph}})}}$$
$$= \psi_{\text{sph}}(b_{k-1})\sqrt{\frac{\underline{q}_{\text{sph}}}{\mathbb{E}[h^2]+\xi'(\underline{q}_{\text{sph}})}}.$$

Finally the first term gives energy contribution

$$h\langle \boldsymbol{m}_{\text{sph}}^k\rangle_N \simeq \mathbb{E}\left[h\left(h+Z\sqrt{\xi'(b_{k-1})}\right)\right]\sqrt{\frac{\underline{q}_{\text{sph}}}{\mathbb{E}[h^2]+\xi'(\underline{q}_{\text{sph}})}}$$
$$= \mathbb{E}[h^2]\sqrt{\frac{\underline{q}_{\text{sph}}}{\mathbb{E}[h^2]+\xi'(\underline{q}_{\text{sph}})}}.$$

Since $\lim_{k\to\infty} b_{k-1} = \xi'(\underline{q}_{\text{sph}})$ and $\psi_{\text{sph}}(\xi'(\underline{q}_{\text{sph}})) = \xi'(\underline{q}_{\text{sph}})$ we conclude

$$\lim_{k\to\infty}\operatorname*{p-lim}_{N\to\infty}\frac{H_N(\boldsymbol{m}_{\text{sph}}^k)}{N} = \sqrt{\underline{q}_{\text{sph}}(\mathbb{E}[h^2]+\xi'(\underline{q}_{\text{sph}}))}.$$

□

Proof of Lemma 4.5.1. The result follows from the preceding lemmas. □

4.5.2 Proof of Theorem 18

It follows from our algorithm that $GS_{\rm sph}(\xi,\mathcal{L}_h) \ge \underline{q}_{\rm sph}\xi''(\underline{q}_{\rm sph})^{1/2} + \int_{\underline{q}_{\rm sph}}^1 \xi''(\underline{q}_{\rm sph})^{1/2}{\rm d}q$. We now characterize the models in which equality holds, which coincide with those exhibiting no overlap gap. Moreover we give an alternate proof of the lower bound for $GS(\xi,\mathcal{L}_h)_{\rm sph}$ which shows that equality holds exactly in no overlap gap models.

Theorem 18. *Suppose ξ and $\mathcal{L}_h$ satisfy $\mathbb{E}[h^2]+\xi'(1) < \xi''(1)$, and let $\underline{q}_{sph} \in (0,1)$ be the unique solution to $\mathbb{E}[h^2]+\xi'(\underline{q}_{sph}) = \underline{q}_{sph}\xi''(\underline{q}_{sph})$. Then the spherical spin glass with parameters $\xi,\mathcal{L}_h$ has no overlap gap if and only if $\xi''(q)^{-1/2}$ is concave on $q \in [\underline{q}_{sph},1]$, in which case α is supported on $[\underline{q}_{sph},1]$ and takes the explicit form*

$$\alpha(s) = \begin{cases} 0, & s \in [0,\underline{q}_{sph}) \\ \frac{\xi'''(s)}{2\xi''(s)^{3/2}}, & s \in [\underline{q}_{sph},1]. \end{cases}$$

Moreover the ground-state energy satisfies

$$GS_{sph}(\xi,\mathcal{L}_h) \ge \underline{q}_{sph}\sqrt{\xi''(\underline{q}_{sph})} + \int_{\underline{q}_{sph}}^1 \sqrt{\xi''(q)}{\rm d}q$$

with equality if and only if no overlap gap occurs.

Proof. We use the results and notation of [CS17]. If $\xi''(q)^{-1/2}$ is concave on $[\underline{q}_{\rm sph},1]$ then the proof of Proposition 2 in [CS17] applies verbatim to show that the support of α is $[\underline{q}_{\rm sph},1]$. In fact it explicitly shows $\alpha(s) = \frac{\xi'''(s)}{2\xi''(s)^{3/2}}$ for $s \in [\underline{q}_{\rm sph},1]$).

In the other direction, we show that if no overlap gap holds and $\mathbb{E}[h^2]+\xi'(1) < \xi''(1)$, then $\xi''(q)^{-1/2}$ is concave on $[\underline{q}_{\rm sph},1]$. we use the statement and notation of [CS17, Theorem 2]. Assume α is supported on the interval $[\underline{q}_{\rm sph},1]$. The last condition in [CS17, Theorem 2] states that $g(u) = \int_u^1 \bar{g}(s){\rm d}s = 0$ for all $u \in [\underline{q}_{\rm sph},1]$, and therefore $\bar{g}(s) = 0$ for $s \in [\underline{q}_{\rm sph},1]$, where

$$\bar{g}(s) \equiv \xi'(s) + h^2 - \int_0^s \frac{dq}{\left(L - \int_0^q \alpha(r)dr\right)^2}.$$

Setting $s = \underline{q}_{\rm sph}$ yields $\mathbb{E}[h^2]+\xi'(\underline{q}_{\rm sph}) = \underline{q}_{\rm sph}L^{-2}$, i.e. $L = \sqrt{\frac{\underline{q}_{\rm sph}}{\mathbb{E}[h^2]+\xi'(\underline{q}_{\rm sph})}}$. Differentiating, all $s \ge \underline{q}_{\rm sph}$ satisfy

$$\xi''(s) = \frac{1}{(L - \int_{\underline{q}_{\rm sph}}^s \alpha(r)dr)^2}. \tag{4.5.7}$$

Taking $s = \underline{q}_{\rm sph}$ in (4.5.7) shows $L = \xi''(\underline{q}_{\rm sph})^{-1/2}$, hence $\mathbb{E}[h^2] + \xi'(\underline{q}_{\rm sph}) = \underline{q}_{\rm sph}\xi''(\underline{q}_{\rm sph})$. Rearranging (4.5.7) yields

$$L - \int_{\underline{q}_{\rm sph}}^{s} \alpha(r)dr = \xi''(s)^{-1/2}, \qquad s \geq \underline{q}_{\rm sph}.$$

As α must be non-decreasing based on [CS17, Equation (9)] it follows that $\xi''(s)^{-1/2}$ is concave on $s \in [\underline{q}_{\rm sph}, 1]$. This completes the proof of the first equivalence. We turn to the value of $GS_{\rm sph}(\xi, \mathcal{L}_h)$, first computing

$$\begin{aligned}\mathbb{E}[h^2] + \xi'(1) &= \underline{q}\xi''(\underline{q}_{\rm sph}) + \int_{\underline{q}_{\rm sph}}^{1} \xi''(\underline{q}_{\rm sph})\mathrm{d}q \\ &= \int_0^{\underline{q}_{\rm sph}} \xi''(\underline{q}_{\rm sph})\mathrm{d}q + \int_{\underline{q}_{\rm sph}}^{1} \xi''(\underline{q}_{\rm sph})\mathrm{d}q.\end{aligned}$$

Letting $L > \int_0^1 \alpha(s)\mathrm{d}s$ and let $a(q) = \int_0^q \alpha(s)\mathrm{d}s$, we find

$$\begin{aligned}2\mathcal{Q}(L,\alpha) &= (\mathbb{E}[h^2] + \xi'(1))L - \int_0^1 \xi''(q)a(q)\mathrm{d}q + \int_0^1 \frac{\mathrm{d}q}{L - a(q)} \\ &= \int_0^{\underline{q}_{\rm sph}} \big(\xi''(\underline{q}_{\rm sph})L - \xi''(q)a(q)\big)\mathrm{d}q + \int_0^{\underline{q}_{\rm sph}} \frac{\mathrm{d}q}{L - a(q)} \\ &+ \int_{\underline{q}_{\rm sph}}^1 \Big(\xi''(q)\big(L - a(q)\big) + \frac{1}{L - a(q)}\Big)\mathrm{d}q.\end{aligned}$$

Since ξ'' is increasing, AM-GM shows the second-to-last line is at most

$$\int_0^{\underline{q}_{\rm sph}} \xi''(\underline{q})\big(L - a(q)\big)\mathrm{d}q + \int_0^{\underline{q}_{\rm sph}} \frac{\mathrm{d}q}{L - a(q)} \geq 2\int_0^{\underline{q}_{\rm sph}} \sqrt{\xi''(\underline{q}_{\rm sph})}\mathrm{d}q = 2\underline{q}\sqrt{\xi''(\underline{q}_{\rm sph})}, \tag{4.5.8}$$

and similarly

$$\int_{\underline{q}_{\rm sph}}^1 \Big(\xi''(q)\big(L - a(q)\big) + \frac{1}{L - a(q)}\Big)\mathrm{d}q \geq 2\int_{\underline{q}_{\rm sph}}^1 \sqrt{\xi''(q)}\mathrm{d}q. \tag{4.5.9}$$

Combining, we conclude the lower bound on $GS_{\rm sph}(\xi, \mathcal{L}_h)$. Moreover for equality to hold in (4.5.8) and (4.5.9) we must have

$$\xi''(q)^{-1/2} = \begin{cases} L - a(q), \ \forall q \in [0, \underline{q}_{\rm sph}], \\ L - a(q), \ \forall q \in [\underline{q}_{\rm sph}, 1). \end{cases}$$

The first equality forces $\alpha(s) = 0$ on $[0, \underline{q}_{\rm sph})$ and $L = \xi''(q_0)^{-1/2}$, while the second equality implies

$\alpha(q) = -\frac{\mathrm{d}}{\mathrm{d}s}\xi''(q)^{-1/2}$ for all $q \in [\underline{q}_{\mathrm{sph}}, 1]$. Taken together this means that equality in the GS_{sph} lower bound implies no overlap gap, completing the proof. □

4.6 Incremental AMP Proofs

We will prove Lemma 4.6.3 which generalizes Lemma 4.3.6 to the setting of branching AMP and describes the limiting Gaussian processes $N^{\delta}_{\ell,a}, Z^{\delta}_{\ell,a}$. We recall the setup of Section 4.4 and in particular continue to use the value $q_B \in (\underline{q}, 1)$ to define the time $\ell^{\delta}_{q_B}$ at which $Z^{\delta}_{\ell^{\delta}_{q_B},1} = Z^{\delta}_{\ell^{\delta}_{q_B},2}$ last holds. For the branching setting we slightly generalize the filtration (4.3.15) to

$$\mathcal{F}^{\delta}_{\ell} = \sigma\left((Z^{\delta}_{k,a}, N^{\delta}_{k,a})_{0 \le k \le \ell, a \in \{1,2\}}\right).$$

Crucially note that we restrict here to $k \ge 0$, i.e. we do not include the preparatory iterates with negative index. We remark that if we consider all the IAMP iterates $(Z^{\delta}_{\ell,a}, N^{\delta}_{\ell,a})$ together in the linear order given by $(\ell, a) \to 2\ell + a$, then these are iterates of a standard AMP algorithm since each iterate depends only on the previous ones. Moreover it is easy to see that the Onsager correction terms are not affected by this rewriting. Therefore we may continue to use state evolution in the natural way even though we do not think of the iterates as actually being totally ordered.

Lemma 4.6.1. *In branching IAMP, $\mathcal{F}^{\delta}_{\ell}$ is jointly independent of the iterates $(Z^{-j})_{J_\ell < j \le K}$ for*

$$J_\ell \equiv \max\left(\{\ell\} \cup \{k_{i,a} + \ell - \ell^{\delta}_{q_i} : \ell^{\delta}_{q_i} \le \ell, a \in \{1,2\}\}\right).$$

Proof. We proceed by induction over ℓ, the base case $\ell = 0$ following from Proposition 4.2.4. Because the random variables $Z^{\ell}_{k,a}$ form a Gaussian process it suffices to verify that

$$\mathbb{E}\left[Z^{\delta}_{\ell,a} Z^{-j}\right] = 0$$

holds whenever $j > J_\ell$. By state evolution,

$$\mathbb{E}\left[Z^{\delta}_{\ell,a} Z^{-j}\right] = \xi'\left(\left[N^{\delta}_{\ell-1,a} Z^{-j-1}\right]\right).$$

By definition $N^{\delta}_{\ell-1,a}$ is $\mathcal{F}^{\delta}_{\ell-1}$-measurable. Since $\xi'(0) = 0$ it suffices to show that $\mathcal{F}^{\delta}_{\ell-1}$ is independent of Z^{-j-1}. By the inductive hypothesis, this holds if $j + 1 > J_{\ell-1}$. This in turn follows from the easy-to-verify fact that $J_\ell - 1 \ge J_{\ell-1}$, completing the proof. □

Corollary 4.6.2. *Let $G^{\delta}_{q_j,a}$ be the state evolution limit of $\mathbf{g}^{(q_j,a)}$ for each $(j,a) \in [m] \times [2]$. Then the law of $(G^{\delta}_{q_i,1}, G^{\delta}_{q_i,2})$ conditioned on $\mathcal{F}^{\delta}_{\ell^{\delta}_{q_i}-1}$ is $\mathsf{N}(0, I_2)$.*

Proof. Since $k_{i,1} \neq k_{i,2}$ it follows from Proposition 4.2.4 that $(G^\delta_{q_i,1}, G^\delta_{q_i,2}) \sim \mathsf{N}(0, I_2)$ holds as an unconditional law. Since we chose the values $k_{i,a}$ such that $k_{i,a} - \ell^\delta_{q_i} > k_{j,a'} - \ell^\delta_{q_j} > 0$ for any $i > j$ and $a, a' \in \{1, 2\}$, it follows that $k_{i,a} > J_{\ell^\delta_{q_i}-1}$. Applying Lemma 4.6.1 now concludes the proof. □

Lemma 4.6.3. *The sequences* $(Z^\delta_{\underline{\ell},a}, Z^\delta_{\underline{\ell}+1,a}, \dots)$ *and* $(N^\delta_{\underline{\ell},a}, N^\delta_{\underline{\ell}+1,a}, \dots)$ *satisfy for* $\ell \geq \underline{\ell}$*:*

$$\mathbb{E}[(Z^\delta_{\ell+1,a} - Z^\delta_{\ell,a})Z^\delta_{j,a}] = 0, \quad \text{for all } \underline{\ell} + 1 \leq j \leq \ell \tag{4.6.1}$$

$$\mathbb{E}[(Z^\delta_{\ell+1,a} - Z^\delta_{\ell,a})^2 | \mathcal{F}^\delta_\ell] = \xi'(q^\delta_{\ell+1}) - \xi'(q^\delta_\ell) \tag{4.6.2}$$

$$\mathbb{E}[(Z^\delta_{\ell+1,1} - Z^\delta_{\ell,1})(Z^\delta_{\ell+1,2} - Z^\delta_{\ell,2}) | \mathcal{F}^\delta_\ell] = \left(\xi'(q^\delta_{\ell+1}) - \xi'(q^\delta_\ell)\right) \cdot 1_{\ell < \ell^\delta_{q_B}} \tag{4.6.3}$$

$$\mathbb{E}[(Z^\delta_{\ell,a})^2] = \xi'(q^\delta_\ell) \tag{4.6.4}$$

$$\mathbb{E}[(N^\delta_{\ell+1,a} - N^\delta_{\ell,a}) | \mathcal{F}^\delta_\ell] = 0 \tag{4.6.5}$$

$$\mathbb{E}[(N^\delta_{\ell+1,a} - N^\delta_{\ell,a})^2 | \mathcal{F}^\delta_\ell] = \delta \tag{4.6.6}$$

$$\mathbb{E}[(N^\delta_{\ell+1,1} - N^\delta_{\ell,1})(N^\delta_{\ell+1,2} - N^\delta_{\ell,2}) | \mathcal{F}^\delta_\ell] = \delta \cdot 1_{\ell < \ell^\delta_{q_B}} \tag{4.6.7}$$

$$\mathbb{E}[(N^\delta_{\ell,a})^2] = q^\delta_{\ell+1}. \tag{4.6.8}$$

Proof. We recall that $(Z^\delta_{\ell,a})_{\ell \geq \underline{\ell}, a \in \{1,2\}}$ is a Gaussian process, which means we can ignore the conditioning on $\mathcal{F}^\delta_\ell$ in proving Equation (4.6.2). First we check that Equations (4.6.4) and (4.6.8) hold for $\ell = \underline{\ell}$. For Equation (4.6.8),

$$\mathbb{E}[(N^\delta_{\underline{\ell},a})^2] = (1 + \varepsilon_0)^2 \mathbb{E}[(M^{\underline{\ell}})^2] = (1 + \varepsilon_0)^2 q = q + \delta = q^\delta_1.$$

For Equation (4.6.4),

$$\mathbb{E}[(Z^\delta_{\underline{\ell},a})^2] = \xi'\left(\mathbb{E}[(M^{\underline{\ell}-1})^2]\right) = \xi'(q).$$

Observe now that if Equations (4.6.1), (4.6.2), (4.6.5), (4.6.6) hold for $\underline{\ell} \leq \ell \leq k$ then so do Equations (4.6.4) and (4.6.8), as

$$\mathbb{E}[(N^\delta_{\ell+1,a})^2] = \mathbb{E}[(N^\delta_{\ell+1,a} - N^\delta_{\ell,a})^2] + 2 \cdot \mathbb{E}[(N^\delta_{\ell+1,a} - N^\delta_{\ell,a})N^\delta_{\ell,a}] + \mathbb{E}[(N^\delta_{\ell,a})^2]$$

and similarly for $\mathbb{E}[(Z^\delta_{\ell+1,a})^2]$. Therefore to show the six identities (4.6.1), (4.6.2), (4.6.5), (4.6.6), (4.6.4) and (4.6.8) it suffices to check the base case $\ell = \underline{\ell}$ for Equations (4.6.1), (4.6.2), (4.6.5), (4.6.6) and to perform an inductive step to show these four identities for $\ell = k + 1$, assuming all six of these equations as inductive hypotheses for $\ell \leq k$. We turn to doing this, and finally show Equations (4.6.3), 4.6.7 at the end.

Base Case for Equations (4.6.1)**,** (4.6.2)**,** (4.6.5)**,** (4.6.6)**.** Note that here, none of the perturbations $\mathbf{g}^{(q_i,a)}$ appear yet. We begin with Equation (4.6.1):

$$\begin{aligned}\mathbb{E}\left[\left(Z^{\delta}_{\underline{\ell}+1,a} - Z^{\delta}_{\underline{\ell},a}\right) Z^{\delta}_{\underline{\ell},a}\right] &= \xi'\left(\mathbb{E}\left[N^{\delta}_{\underline{\ell},a} M^{\underline{\ell}-1}\right]\right) - \xi'\left(\mathbb{E}\left[M^{\underline{\ell}-1} M^{\underline{\ell}-1}\right]\right) \\ &= \xi'\left((1+\varepsilon_0)\,\mathbb{E}\left[M^{\underline{\ell}} M^{\underline{\ell}-1}\right]\right) - \xi'(q) \\ &= \xi'((1+\varepsilon_0)\phi(a_{\underline{\ell}-1})) - \xi'(q) \\ &= \xi'(q) - \xi'(q) \\ &= 0.\end{aligned}$$

This means $\mathbb{E}[Z^{\delta}_{\underline{\ell}+1,a} | Z^{\delta}_{\underline{\ell},a}] = Z^{\delta}_{\underline{\ell},a}$. Hence

$$\mathbb{E}\left[\left(Z^{\delta}_{\underline{\ell}+2,a} - Z^{\delta}_{\underline{\ell}+1,a}\right) Z^{\delta}_{\underline{\ell}+1,a}\right] = \xi'\left(\mathbb{E}[N^{\delta}_{\underline{\ell}+1} N^{\delta}_{\underline{\ell},a}]\right) - \xi'\left(\mathbb{E}[N^{\delta}_{\underline{\ell},a} N^{\delta}_{\underline{\ell},a}]\right).$$

To see that the above expression vanishes, it suffices to show that

$$\mathbb{E}[(N^{\delta}_{\underline{\ell}+1,a} - N^{\delta}_{\underline{\ell},a}) N^{\delta}_{\underline{\ell},a}] = 0.$$

This follows since we just showed $\mathbb{E}[Z^{\delta}_{\underline{\ell}+1,a} | Z^{\delta}_{\underline{\ell},a}] = Z^{\delta}_{\underline{\ell},a}$ and we have

$$\mathbb{E}[(N^{\delta}_{\underline{\ell}+1,a} - N^{\delta}_{\underline{\ell},a}) N^{\delta}_{\underline{\ell},a}] = \mathbb{E}[u^{\delta}_{\underline{\ell},a}(X^{\delta}_{\underline{\ell},a})(Z^{\delta}_{\underline{\ell}+1,a} - Z^{\delta}_{\underline{\ell},a})] = \mathbb{E}[u^{\delta}_{\underline{\ell},a}(Z^{\delta}_{\underline{\ell},a})(Z^{\delta}_{\underline{\ell}+1,a} - Z^{\delta}_{\underline{\ell},a})]$$

Next we verify the base case for Equation (4.6.2). Using the base case of Equation (4.6.1) in the first step we compute:

$$\begin{aligned}\mathbb{E}\left[\left(Z^{\delta}_{\underline{\ell}+1,a} - Z^{\delta}_{\underline{\ell},a}\right)^2\right] &= \mathbb{E}\left[\left(Z^{\delta}_{\underline{\ell}+1,a}\right)^2\right] - \mathbb{E}\left[\left(Z^{\delta}_{\underline{\ell},a}\right)^2\right] \\ &= \xi'\left(\mathbb{E}\left[(N^{\delta}_{\underline{\ell},a})^2\right]\right) - \xi'(q) \\ &= (1+\varepsilon_0)^2 q - \xi'(q) \\ &= \xi'\left(\frac{q^2}{\phi(a_{\underline{\ell}-1})^2}\right) - \xi'(q) \\ &= \xi'(q+\delta) - \xi'(q) \\ &= \xi'(q^{\delta}_{\underline{\ell}+1}) - \xi'(q).\end{aligned}$$

Continuing, we verify the base case for Equation (4.6.5). First note that

$$\begin{aligned}\mathbb{E}\left[\left(N^{\delta}_{\underline{\ell}+1,a} - N^{\delta}_{\underline{\ell},a}\right) | \mathcal{F}^{\delta}_{\underline{\ell}}\right] &= \mathbb{E}\left[u^{\delta}_{\underline{\ell}}(X^{\delta}_{\underline{\ell},a})(Z^{\delta}_{\underline{\ell}+1,a} - Z^{\delta}_{\underline{\ell},a}) \,|\, \mathcal{F}^{\delta}_{\underline{\ell}}\right] \\ &= 0.\end{aligned}$$

The last line holds because $X^\delta_{\underline{\ell},a}$ is $\mathcal{F}^\delta_{\underline{\ell}}$-measurable and $\mathbb{E}[Z^\delta_{\underline{\ell}+1,a} - Z^\delta_{\underline{\ell},a}|\mathcal{F}^\delta_{\underline{\ell}}] = 0$ as deduced above. Finally for Equation (4.6.6) using the martingale property again we obtain:

$$\begin{aligned}\mathbb{E}\left[\left(N^\delta_{\underline{\ell}+1,a} - N^\delta_{\underline{\ell},a}\right)^2\right] &= \mathbb{E}\left[(u^\delta_{\underline{\ell}}(X^\delta_{\underline{\ell},a}))^2(Z^\delta_{\underline{\ell}+1,a} - Z^\delta_{\underline{\ell},a})^2\right] \\ &= \mathbb{E}\left[\frac{\delta}{\xi'(q^\delta_{\underline{\ell}+1,a}) - \xi'(q^\delta_{\underline{\ell}})}\left(\xi'(q^\delta_{\underline{\ell}+1}) - \xi'(q^\delta_{\underline{\ell}})\right)\right] \\ &= \delta.\end{aligned}$$

Here the second line follows from the definition of u^δ_ℓ, and we can multiply the two expectations because $\mathbb{E}[(Z^\delta_{\underline{\ell}+1,a} - Z^\delta_{\underline{\ell},a})^2|\mathcal{F}^\delta_{\underline{\ell},a}]$ is constant while the other term is $\mathcal{F}^\delta_{\underline{\ell},a}$ measurable.

Inductive step We now induct, assuming all 6 identities (4.6.1), (4.6.2), (4.6.5), (4.6.6), (4.6.4) and (4.6.8) up to ℓ and showing Equations (4.6.1), (4.6.2), (4.6.5), (4.6.6) for $\ell+1$. We begin with Equation (4.6.1). Let $\underline{\ell}+1 \le j \le \ell$. State evolution implies

$$\mathbb{E}\left[\left(Z^\delta_{\ell+1,a} - Z^\delta_{\ell,a}\right) Z^\delta_{j,a}\right] = \xi'\left(\mathbb{E}\left[N^\delta_{\ell,a}N^\delta_{j-1,a}\right]\right) - \xi'\left(\mathbb{E}\left[N^\delta_{\ell-1,a}N^\delta_{j-1,a}\right]\right).$$

To show this equals 0 we must show

$$\mathbb{E}\left[N^\delta_{\ell,a}N^\delta_{j-1,a}\right] = \mathbb{E}\left[N^\delta_{\ell-1,a}N^\delta_{j-1,a}\right].$$

When $\ell = \ell^\delta_{q_i}$ for some $i \in [m]$ this follows from Corollary 4.6.2. Assuming $\ell \neq \ell^\delta_{q_i}$ for all i, the difference between the left and right sides is

$$\mathbb{E}[u^\delta_{\ell-1}(X^\delta_{\ell-1,a})(Z^\delta_{\ell,a} - Z^\delta_{\ell-1,a})N^\delta_{j-1,a}].$$

Since $N^\delta_{j-1,a}$ is $\mathcal{F}^\delta_{\ell-1}$ measurable and $\mathbb{E}[Z^\delta_{\ell,a}|\mathcal{F}^\delta_{\ell-1}] = Z^\delta_{\ell-1,a}$ holds by inductive hypothesis, we conclude the inductive step for Equation (4.6.1).

We continue to Equation (4.6.2). Using Equation (4.6.1) just proven in the first step we get

$$\begin{aligned}\mathbb{E}\left[\left(Z^\delta_{\ell+1,a} - Z^\delta_{\ell,a}\right)^2\right] &= \mathbb{E}\left[\left(Z^\delta_{\ell+1,a}\right)^2 - \left(Z^\delta_{\ell,a}\right)^2\right] \\ &= \xi'\left(\mathbb{E}[N^\delta_{\ell,a}N^\delta_{\ell,a}]\right) - \xi'\left(\mathbb{E}[N^\delta_{\ell-1,a}N^\delta_{\ell-1,a}]\right) \\ &= \xi'\left(q^\delta_{\ell+1}\right) - \xi'\left(q^\delta_{\ell}\right)\end{aligned}$$

Next we show Equation (4.6.5) continues to hold. If $\ell+1 = \ell^\delta_{q_i}$ for some $i \in [m]$ again follows from Corollary 4.6.2. When $\ell+1 \neq \ell^\delta_{q_i}$ for all i, it follows from the definition of the sequence $N^\delta_{\ell,a}$ and the just proven fact that $(Z^\delta_{\ell,a})_{\ell \ge \underline{\ell}+1}$ forms a martingale sequence. Finally we show Equation (4.6.6)

continues to hold inductively. Again for $\ell+1=\ell^{\delta}_{q_i}$ it follows from Corollary 4.6.2, and otherwise by definition

$$\mathbb{E}[(u^{\delta}_{\ell})^2]=\frac{\delta}{\xi'(q^{\delta}_{\ell})-\xi'(q^{\delta}_{\ell-1})}.$$

Moreover what we showed before implies $\mathbb{E}[(Z^{\delta}_{\ell+1,a}-Z^{\delta}_{\ell,a})^2|\mathcal{F}^{\delta}_{\ell}]=\xi'(q^{\delta}_{\ell})-\xi'(q^{\delta}_{\ell-1})$. Applying these observations to the identity

$$\mathbb{E}[(N^{\delta}_{\ell+1,a}-N^{\delta}_{\ell})^2|\mathcal{F}_{\ell}]=(u^{\delta}_{\ell})^2\mathbb{E}\left[(Z^{\delta}_{\ell+1,a}-Z^{\delta}_{\ell,a})^2\right]$$

implies Equation (4.6.6) continues to hold.

Equations (4.6.3) **and** (4.6.7) Finally we consider (4.6.3) and (4.6.7). For $\ell<\ell^{\delta}_{q_i}$ they follow directly from (4.6.2), (4.6.6). For $\ell=\ell^{\delta}_{q_i}$, (4.6.7) is trivial while (4.6.3) immediately follows from state evolution. For $\ell>\ell^{\delta}_{q_i}$, (4.6.7) follows from the inductive hypothesis and the computation

$$\mathbb{E}[(N^{\delta}_{\ell+1,1}-N^{\delta}_{\ell,1})(N^{\delta}_{\ell+1,2}-N^{\delta}_{\ell,2})|\mathcal{F}^{\delta}_{\ell}]=(u^{\delta}_{\ell})^2\mathbb{E}\left[(Z^{\delta}_{\ell+1,1}-Z^{\delta}_{\ell,1})(Z^{\delta}_{\ell+1,2}-Z^{\delta}_{\ell,2})|\mathcal{F}^{\delta}_{\ell}\right]=0.$$

Finally for $\ell>\ell^{\delta}_{q_i}$, (4.6.3) follows from the expansion

$$\begin{aligned}\mathbb{E}[(Z^{\delta}_{\ell+1,1}-Z^{\delta}_{\ell,1})(Z^{\delta}_{\ell+1,2}-Z^{\delta}_{\ell,2})]=\xi'\left(\mathbb{E}[N^{\delta}_{\ell,1}N^{\delta}_{\ell,2}]\right)-\xi'\left(\mathbb{E}[N^{\delta}_{\ell-1,1}N^{\delta}_{\ell,2}]\right)\\-\xi'\left(\mathbb{E}[N^{\delta}_{\ell-1,1}N^{\delta}_{\ell,2}]\right)+\xi'\left(\mathbb{E}[N^{\delta}_{\ell-1,1}N^{\delta}_{\ell-1,2}]\right)\end{aligned}$$

and the fact that all 4 terms on the right hand side are equal thanks to (4.6.5), (4.6.7). □

4.6.1 Diffusive Scaling Limit

We begin with the following slight generalization of Lemma 4.3.7 which allows for the additional perturbation steps of branching IAMP but still considers only a single sample path.

Lemma 4.6.4. *Fix $\bar{q}\in(\underline{q},1)$ and an index a. There exists a coupling between the families of triples $\{(Z^{\delta}_{\ell,a},X^{\delta}_{\ell,a},N^{\delta}_{\ell,a})\}_{\ell\geq 0}$ and $\{(Z_t,X_t,N_t)\}_{t\geq 0}$ such that the following holds. For some $\delta_0>0$ and constant $C>0$, for every $\delta\leq\delta_0$ and $\ell\geq\underline{\ell}$ with $q_\ell\leq\bar{q}$ we have*

$$\max_{\underline{\ell}\leq j\leq\ell}\mathbb{E}\left[\left(X^{\delta}_{j,a}-X_{q_j}\right)^2\right]\leq C\delta, \tag{4.6.9}$$

$$\max_{\underline{\ell}\leq j\leq\ell}\mathbb{E}\left[\left(N^{\delta}_{j,a}-N_{q_j}\right)^2\right]\leq C\delta. \tag{4.6.10}$$

Proof. We prove the scaling limits for X^{δ}_{ℓ} and N^{δ}_{ℓ} separately, inducting over ℓ in each proof. We suppress the index a as it is irrevelant.

Scaling limit for X_ℓ^δ We begin by checking the claim for $\ell = \underline{\ell}$. Recalling that $\int_0^{q_{\underline{\ell}+1}} \sqrt{\xi''(t)}\mathrm{d}B_t = Z_{\underline{\ell}+1}$, we have

$$\begin{aligned}
\mathbb{E}\left[\left(X_{\underline{\ell}}^\delta - X_q\right)^2\right] &= \mathbb{E}\left[\left(Z_{\underline{\ell}}^\delta - \int_0^{q_{\underline{\ell}+1}} \sqrt{\xi''(t)}\mathrm{d}B_t\right)^2\right] \\
&\le 2\,\mathbb{E}\left[\left(Z_{\underline{\ell}}^\delta - Z_{\underline{\ell}+1}^\delta\right)^2\right] + 2\,\mathbb{E}\left[\left(\int_q^{q_{\underline{\ell}+1}} \sqrt{\xi''(t)}\mathrm{d}B_t\right)^2\right] \\
&= 4(\xi'(q_{\underline{\ell}+1}) - \xi'(q)) \\
&\le C\delta.
\end{aligned}$$

We continue using a standard self-bounding argument. Let $\ell \ge \underline{\ell}+1$ such that $q_\ell \le \bar{q}$. Define $\Delta_\ell^X = X_\ell^\delta - X_{q_\ell}$. Then

$$\begin{aligned}
\Delta_\ell^X - \Delta_{\ell-1}^X &= \int_{q_{\ell-1}^\delta}^{q_\ell^\delta} \big(v(q_{\ell-1}^\delta; X_\ell^\delta) - v(t; X_t)\big)\mathrm{d}t + Z_\ell^\delta - Z_{\ell-1}^\delta - \int_{q_{\ell-1}^\delta}^{q_\ell^\delta} \sqrt{\xi''(s)}\mathrm{d}B_s \\
&= \int_{q_{\ell-1}^\delta}^{q_\ell^\delta} \big(v(q_{\ell-1}^\delta; X_\ell^\delta) - v(t; X_t)\big)\mathrm{d}t \\
&= \int_{q_{\ell-1}^\delta}^{q_\ell^\delta} \big(v(q_{\ell-1}^\delta; X_\ell^\delta) - v(q_{\ell-1}^\delta; X_t)\big)\mathrm{d}t + \int_{q_{\ell-1}^\delta}^{q_\ell^\delta} \big(v(q_{\ell-1}^\delta; X_t) - v(t; X_t)\big)\mathrm{d}t.
\end{aligned}$$

The first term just above is at most $C\int_{q_{j-1}^\delta}^{q_j^\delta} |X_j^\delta - X_t|\mathrm{d}t$ since v is Lipschitz in space uniformly for $t \in [0,1]$. For the second term we estimate

$$\begin{aligned}
&\sum_{k=\underline{\ell}+1}^{\ell} \int_{q_{k-1}^\delta}^{q_k^\delta} \big|v(q_{k-1}^\delta; X_t) - v(t; X_t)\big|\mathrm{d}t \\
&\quad\le \sum_{k=\underline{\ell}+1}^{\ell} \int_{q_{k-1}^\delta}^{q_k^\delta} \Big\{\big|v(q_{k-1}^\delta; X_t) - v(t; X_t)\big| + \big|v(t; X_t) - v(q_k^\delta; X_t)\big|\Big\}\mathrm{d}t \\
&\quad\le \delta \sum_{k=\underline{\ell}+1}^{\ell} \sup_{q_{k-1}^\delta \le t \le q_k^\delta} \Big\{\big|v(q_{k-1}^\delta; X_t) - v(t; X_t)\big| + \big|v(t; X_t) - v(q_k^\delta; X_t)\big|\Big\} \\
&\quad\le \delta \sup_{t_1,\cdots,t_k} \sum_{k=\underline{\ell}+1}^{\ell} \Big\{\big|v(q_{k-1}^\delta; X_{t_k}) - v(t_k; X_{t_k})\big| + \big|v(t_k; X_{t_k}) - v(q_k^\delta; X_{t_k})\big|\Big\} \\
&\quad\le C\delta,
\end{aligned}$$

where the last inequality follows since $|\partial_t(v(t,x))|, |\partial_x(v(t,x))$ are uniformly bounded for $t \in [0,\bar{q}], x \in$

$\mathbb{R}$. Combining the bounds and summing over j, we find

$$\left|\Delta_\ell^X\right| \le |\Delta_{\underline{\ell}}^X| + \sum_{j=\underline{\ell}+1}^{\ell} \left|\Delta_j^X - \Delta_{j-1}^X\right| \le C \sum_{j=\underline{\ell}+1}^{\ell} \int_{q_{j-1}^\delta}^{q_j^\delta} |X_j^\delta - X_t| \mathrm{d}t + 2C\delta.$$

Squaring and taking expectations,

$$\begin{aligned}\mathbb{E}\left[(\Delta_\ell^X)^2\right] &\le 2C^2\,\mathbb{E}\Big(\sum_{j=\underline{\ell}+1}^{\ell} \int_{q_{j-1}^\delta}^{q_j^\delta} |X_j^\delta - X_t| \mathrm{d}t\Big)^2 + 10C^2\delta^2 \\ &\le 2C^2(\ell - \underline{\ell})\delta \sum_{j=\underline{\ell}+1}^{\ell} \int_{q_{j-1}^\delta}^{q_j^\delta} \mathbb{E}\,|X_j^\delta - X_t|^2 \mathrm{d}t + 10C^2\delta^2.\end{aligned}$$

Furthermore, $\mathbb{E}\,|X_j^\delta - X_t|^2 \le 2\,\mathbb{E}\,|X_j^\delta - X_{q_j^\delta}|^2 + 2\,\mathbb{E}\,|X_{q_j^\delta} - X_t|^2$. It is clear that $\mathbb{E}\,|X_t - X_s|^2 \le C|t-s|$ for all t, s, as ξ'' is bounded on $[0,1]$. Therefore

$$\mathbb{E}\left[(\Delta_\ell^X)^2\right] \le 4C^2(\ell - \underline{\ell})\delta^2 \sum_{j=\underline{\ell}+1}^{\ell} \mathbb{E}\left[(\Delta_j^X)^2\right] + 4C^3(\ell - \underline{\ell})\delta \sum_{j=\underline{\ell}+1}^{\ell} \int_{q_{j-1}^\delta}^{q_j^\delta} \delta \mathrm{d}t + 10C^2\delta^2.$$

The middle term is proportional to $(\ell - \underline{\ell})^2\delta^3$. Using $(\ell - \underline{\ell})\delta \le 1$ we obtain that for δ smaller than an absolute constant, it holds that

$$\mathbb{E}\left[(\Delta_\ell^X)^2\right] \le C\delta \sum_{j=\underline{\ell}+1}^{\ell-1} \mathbb{E}\left[(\Delta_j^X)^2\right] + C\delta,$$

for a different absolute constant C. This implies $\mathbb{E}\left[(\Delta_\ell^X)^2\right] \le C\delta$ as desired.

Scaling limit for N_ℓ^δ Again we begin by checking that $\ell = \underline{\ell}$. We compute:

$$\begin{aligned}\mathbb{E}\left[\left(N_{\underline{\ell}}^\delta - N_q\right)^2\right] &= \mathbb{E}\left[\left((1+\varepsilon_0)\partial_x\Phi_{\gamma_*}(q, X_{\underline{\ell}}) - \partial_x\Phi_{\gamma_*}(q, X_q)\right)^2\right] \\ &\le 2\varepsilon_0^2\,\mathbb{E}\left[\left(\partial_x\Phi_{\gamma_*}(q, X_{\underline{\ell}})\right)^2\right] + 2\,\mathbb{E}\left[\left(\partial_x\Phi_{\gamma_*}(q, X_{\underline{\ell}}) - \partial_x\Phi_{\gamma_*}(q, X_q)\right)^2\right] \\ &= C\varepsilon_0^2 + C\,\mathbb{E}\left[\left(X_{\underline{\ell}}^\delta - X_q\right)^2\right] \\ &\le C\delta.\end{aligned}$$

Here we have used again the inequality $(x-z)^2 \le 2(x-y)^2 + 2(y-z)^2$ and the fact that derivatives of Φ_{γ_*} are bounded, as well as $\varepsilon_0^2 \le \delta/q$. At the end we use the bound on $\mathbb{E}\left[\left(X_{\underline{\ell}}^\delta - X_q\right)^2\right]$ shown

in the previous part of this proof. Next we turn to $\ell \geq \underline{\ell} + 1$. We have

$$\begin{aligned}(N^\delta_{j+1} - N_{q^\delta_{j+1}}) - (N^\delta_j - N_{q^\delta_j}) &= u^\delta_j(X^\delta_j)(Z^\delta_{j+1} - Z^\delta_j) - \int_{q^\delta_j}^{q^\delta_{j+1}} \sqrt{\xi''(t)} u(t, X_t) \mathrm{d}B_t \\ &= \int_{q^\delta_j}^{q^\delta_{j+1}} \sqrt{\xi''(t)} \left(u^\delta_j(X^\delta_j) - u(t, X_t)\right) \mathrm{d}B_t.\end{aligned}$$

and so

$$\mathbb{E}\left[\left(N^\delta_\ell - N_{q^\delta_\ell}\right)^2\right] \leq 2 \cdot \mathbb{E}\left[\left(N^\delta_{\underline{\ell}} - N_q\right)^2\right] + 2 \cdot \mathbb{E}\left[\left(\sum_{j=\underline{\ell}}^{\ell-1} \int_{q^\delta_j}^{q^\delta_{j+1}} \sqrt{\xi''(t)} \left(u^\delta_j(X^\delta_j) - u(t, X_t)\right) \mathrm{d}B_t\right)^2\right] \tag{4.6.11}$$

$$\leq 2C\delta + 2\sum_{j=\underline{\ell}}^{\ell-1} \int_{q^\delta_j}^{q^\delta_{j+1}} \mathbb{E}\left[\left(u^\delta_j(X^\delta_j) - u(t, X_t)\right)^2\right] \xi''(t) \mathrm{d}t.$$

Recall that $u^\delta_j(x) = u(q^\delta_j; x)/\Sigma^\delta_j$ for $j \geq 1$ where Σ^δ_j is given by

$$(\Sigma^\delta_j)^2 = \frac{\xi'(q^\delta_{j+1}) - \xi'(q^\delta_j)}{\delta} \mathbb{E}[u(q^\delta_j; X^\delta_j)^2].$$

We first show the bound

$$\left|(\Sigma^\delta_j)^2 - 1\right| \leq C\sqrt{\delta} \tag{4.6.12}$$

for δ small enough, which is of independent interest. Since u is bounded and ξ''' is bounded on $[0, 1]$, we have

$$\left|(\Sigma^\delta_j)^2 - \xi''(q^\delta_j) \mathbb{E}[u(q^\delta_j; X^\delta_j)^2]\right| \leq C\delta.$$

Observe now that

$$\mathbb{E}\left[\left|X^\delta_j - X_{q^\delta_j}\right|\right] \leq \sqrt{\delta} + \frac{\mathbb{E}\left[\left|X^\delta_j - X_{q^\delta_j}\right|^2\right]}{\sqrt{\delta}} \leq C\sqrt{\delta}.$$

Since u is Lipschitz in space and bounded, this implies

$$\left|(\Sigma^\delta_j)^2 - \xi''(q^\delta_j) \mathbb{E}[u(q^\delta_j; X_{q^\delta_j})^2]\right| \leq C\sqrt{\delta}.$$

Since $\mathbb{E}[N_t^2] = t$ for all $t \in [0, 1]$ and $t \mapsto u(t, X_t)$ is a.s. continuous, Lebesgue's differentiation theorem implies that for all $t \in [0, 1]$,

$$\xi''(t) \mathbb{E}[u(t; X_t)^2] = 1,$$

and hence $\big|(\Sigma_j^\delta)^2 - 1| \le C\sqrt{\delta}$ for δ smaller than some absolute constant. This implies the bound $|u_j^\delta(X_j^\delta) - u(q_j^\delta; X_j^\delta)| \le C\big|\frac{1}{\Sigma_j^\delta} - 1\big| \le C\sqrt{\delta}$. Now, going back to Eq. (4.6.11), we have

$$\begin{aligned}\mathbb{E}\left[\left(N_\ell^\delta - N_{q_\ell^\delta}\right)^2\right] &\le 2\sum_{j=\underline{\ell}}^{\ell-1}\int_{q_j^\delta}^{q_{j+1}^\delta}\mathbb{E}\left[\left(u_j^\delta(X_j^\delta) - u(q_j^\delta; X_j^\delta)\right)^2\right]\xi''(t)\mathrm{d}t \\ &\quad + 2\sum_{j=\underline{\ell}}^{\ell-1}\int_{q_j^\delta}^{q_{j+1}^\delta}\mathbb{E}\left[\left(u(q_j^\delta; X_j^\delta) - u(t, X_t)\right)^2\right]\xi''(t)\mathrm{d}t\end{aligned}$$

From what we just established the first term is at most $C(\ell - \underline{\ell})\delta^2 \le C\delta$. To estimate the second term we compute:

$$\begin{aligned}&\sum_{j=\underline{\ell}}^{\ell-1}\int_{q_j^\delta}^{q_{j+1}^\delta}\mathbb{E}\left[\left(u(q_j^\delta; X_j^\delta) - u(t, X_t)\right)^2\right]\xi''(t)\mathrm{d}t \\ &\qquad \le C\sum_{j=\underline{\ell}}^{\ell-1}\int_{q_j^\delta}^{q_{j+1}^\delta}\mathbb{E}\left[\left(u(q_j^\delta; X_j^\delta) - u(q_j^\delta, X_{q_j^\delta})\right)^2\right]\mathrm{d}t \\ &\qquad + C\sum_{j=\underline{\ell}}^{\ell-1}\int_{q_j^\delta}^{q_{j+1}^\delta}\mathbb{E}\left[\left(u(q_j^\delta; X_{q_j^\delta}) - u(q_j^\delta, X_t)\right)^2\right]\mathrm{d}t \\ &\qquad + C\sum_{j=\underline{\ell}}^{\ell-1}\int_{q_j^\delta}^{q_{j+1}^\delta}\mathbb{E}\left[\left(u(q_j^\delta; X_t) - u(t, X_t)\right)^2\right]\mathrm{d}t \\ &\qquad = I + II + III.\end{aligned}$$

Since u is Lipschitz in space, we obtain $I \le C(\ell - \underline{\ell})\delta^2$. From $\mathbb{E}[|X_t - X_s|^2] \le C|t - s|$, we obtain $II \le C(\ell - \underline{\ell})\delta^2$. Finally, since u is Lipschitz in time uniformly in space and $(\ell - \underline{\ell})\delta \le 1$, it follows that $III \le C\delta$. Altogether we obtain

$$\mathbb{E}\left[\left(N_\ell^\delta - N_{q_\ell^\delta}\right)^2\right] \le C\delta$$

concluding the proof. □

We now extend Lemmas 4.3.7, 4.6.4 to describe the joint scaling limit of multiple branches, which become independent at the branching time. Let $(B_t^a)_{t\in[0,1],a\in\{1,2\}}$ be standard Brownian motions with $B_t^1 = B_t^2$ for $t \le q_B$ and with independent increments after time q_B. Couple B_t^a with $(Z_{\ell,a}^\delta)_{\ell\ge 0}$ via

$$Z_{j,a}^\delta = \int_0^{q_j^\delta}\sqrt{\xi''(t)}\mathrm{d}B_t^a$$

and natural filtration $(\mathcal{F}_t)_{t\in[0,1]}$ with $\mathcal{F}_t = \sigma\left((B_s^1, B_s^2)_{s\le t}\right)$. We consider for $a \in \{1,2\}$ the SDE

$$dX_t^a = \gamma_*(t)\partial_x\Phi_{\gamma_*}(t, X_t^a)\mathrm{d}t + \sqrt{\xi''(t)}\mathrm{d}B_t^a$$

with initial condition $X_0^a = 0$, and define

$$N_t^a \equiv \partial_x\Phi_{\gamma_*}(q, X_{\underline{q}}^a) + \int_{\underline{q}}^t \sqrt{\xi''(s)}u(s, X_s^a)\mathrm{d}B_s^a = \partial_x\Phi_{\gamma_*}(t, X_t^a),$$
$$Z_t^a \equiv \int_0^t \sqrt{\xi''(s)}\mathrm{d}B_s^a.$$

Lemma 4.6.5. *Fix $\bar{q} \in (\underline{q}, 1)$. There exists a coupling between the families of triples $\{(Z_{\ell,a}^\delta, X_{\ell,a}^\delta, N_{\ell,a}^\delta)\}_{\ell\ge0,a\in\{1,2\}}$ and $\{(Z_t^a, X_t^a, N_t^a)\}_{t\ge0,a\in\{1,2\}}$ such that the following holds. For some $\delta_0 > 0$ and constant $C > 0$, for every $\delta \le \delta_0$ and $\ell \ge \underline{\ell}$ with $q_\ell \le \bar{q}$ we have*

$$\max_{\underline{\ell}\le j\le\ell} \mathbb{E}\left[\left(X_{j,a}^\delta - X_{q_j}^a\right)^2\right] \le C\delta,$$
$$\max_{\underline{\ell}\le j\le\ell} \mathbb{E}\left[\left(N_{j,a}^\delta - N_{q_j}^a\right)^2\right] \le C\delta.$$

Proof. We generate the desired "grand coupling" by starting with (B_t^1, B_t^2) as above, generating (Z_t^1, Z_t^2), and then setting $Z_{j,a}^\delta = Z_{q_j^\delta}^a$ for each $a \in \{1,2\}$ and $j \le \overline{\ell}$ as in the coupling of Lemma 4.6.4. It follows from Lemma 4.6.3 that this results in the correct law for $(Z_{j,a}^\delta)_{j\in\mathbb{N},a\in\{1,2\}}$. Now, all 3 continuous-time functions in the coupling of Lemma 4.6.4 are determined almost surely by Z_t. Furthermore all 3 discrete-time functions are determined almost surely by the sequence Z_j^δ. Therefore the coupling just constructed between $\{Z_{\ell,a}^\delta\}_{\ell\ge0,a\in\{1,2\}}$ and $\{Z_t^a\}_{t\ge0,a\in\{1,2\}}$ automatically extends to a coupling of $\{(Z_{\ell,a}^\delta, X_{\ell,a}^\delta, N_{\ell,a}^\delta)\}_{\ell\ge0,a\in\{1,2\}}$ and $\{(Z_t^a, X_t^a, N_t^a)\}_{t\ge0,a\in\{1,2\}}$. Since the two a-marginals of the coupling just constructed both agree with that of Lemma 4.6.4, the claimed approximation estimates carry over as well, concluding the proof. □

4.6.2 The Energy Gain of Incremental AMP

Here we prove Lemma 4.3.9, stated for the branching case.

Lemma 4.6.6.

$$\lim_{\bar{q}\to1}\lim_{\underline{\ell}\to\infty}\operatorname*{p\text{-}lim}_{N\to\infty}\frac{H_N\left(\boldsymbol{n}^{\overline{\ell},a}\right) - H_N\left(\boldsymbol{n}^{\underline{\ell},a}\right)}{N} = \int_{\underline{q}}^1 \xi''(t)\mathbb{E}\left[u\left(t, X_t\right)\right]\mathrm{d}t. \tag{4.6.13}$$

Proof. We give the main part of the proof for the ordinary (non-branching) version of the algorithm

and explain at the end why the same arguments apply in the branching case. Recall that $\delta = \delta(\underline{\ell}) \to 0$ as $\underline{\ell} \to \infty$, which we will implicitly use throughout the proof. Observe also that $\langle h, \boldsymbol{n}^{\overline{\ell}} - \boldsymbol{n}^{\underline{\ell}} \rangle_N \simeq 0$ because the values $(N_\ell^\delta)_{\ell \geq \underline{\ell}}$ form a martingale sequence. Therefore it suffices to compute the in-probability limit of $\frac{\widetilde{H}_N\left(\boldsymbol{n}^{\overline{\ell}}\right) - \widetilde{H}_N\left(\boldsymbol{n}^{\underline{\ell}}\right)}{N}$. The key is to write

$$\frac{\widetilde{H}_N\left(\boldsymbol{n}^{\overline{\ell}}\right) - \widetilde{H}_N\left(\boldsymbol{n}^{\underline{\ell}}\right)}{N} = \sum_{\ell=\underline{\ell}}^{\overline{\ell}-1} \frac{\widetilde{H}_N\left(\boldsymbol{n}^{\ell+1}\right) - \widetilde{H}_N\left(\boldsymbol{n}^{\ell}\right)}{N}$$

and use a Taylor series approximation of the summand. In particular for $F \in C^3(\mathbb{R})$, applying Taylor's approximation theorem twice yields

$$\begin{aligned} F(1) - F(0) &= aF'(0) + \frac{1}{2}F''(0) + O(\sup_{a\in[0,1]} |F'''(a)|) \\ &= F'(0) + \frac{1}{2}(F'(a) - F'(0)) + O(\sup_{a\in[0,1]} |F'''(a)|) \\ &= \frac{1}{2}(F'(1) + F'(0)) + O(\sup_{a\in[0,1]} |F'''(a)|). \end{aligned}$$

Assuming $\sup_\ell \frac{|\boldsymbol{n}^\ell|}{\sqrt{N}} \leq 2$, which holds with high probability, we apply this estimate with $F(a) = \widetilde{H}_N\left((1-a)\boldsymbol{n}^\ell + a\boldsymbol{n}^{\ell+1}\right)$. The result is:

$$\begin{aligned} &\left|\widetilde{H}_N\left(\boldsymbol{n}^{\ell+1}\right) - \widetilde{H}_N\left(\boldsymbol{n}^{\ell}\right) - \frac{1}{2}\left\langle \nabla\widetilde{H}_N(\boldsymbol{n}^\ell) + \nabla\widetilde{H}_N(\boldsymbol{n}^{\ell+1}), \boldsymbol{n}^{\ell+1} - \boldsymbol{n}^\ell \right\rangle\right| \\ &\qquad \leq O\left(\sup_{|\boldsymbol{v}|\leq 2\sqrt{N}} \left\|\nabla^3 \widetilde{H}_N(\boldsymbol{v})\right\|_{\text{inj}}\right) |\boldsymbol{n}^{\ell+1} - \boldsymbol{n}^\ell|^3. \end{aligned}$$

Here $\|T\|_{\text{inj}} = \sup_{\|\boldsymbol{x}\|=1} \langle T, \boldsymbol{x}^{\otimes 3}\rangle$ denotes the injective tensor norm on $(\mathbb{R}^N)^{\otimes 3}$. Proposition 4.2.1 implies that

$$\sup_{|\boldsymbol{v}|\leq 2\sqrt{N}} \left\|\nabla^3 \widetilde{H}_N(\boldsymbol{v})\right\|_{\text{inj}} \leq O(N^{-1/2}$$

with high probability. On the other hand p-$\lim_{N\to\infty} |\boldsymbol{n}^{\ell+1} - \boldsymbol{n}^\ell| = \sqrt{\delta N}$ for each $\underline{\ell} \leq \ell \leq \overline{\ell} - 1$. Summing and recalling that $\overline{\ell} - \underline{\ell} \leq \delta^{-1}$ yields the high-probability estimate

$$\begin{aligned} &\sum_{\ell=\underline{\ell}}^{\overline{\ell}-1} \left|\widetilde{H}_N\left(\boldsymbol{n}^{\ell+1}\right) - \widetilde{H}_N\left(\boldsymbol{n}^{\ell}\right) - \frac{1}{2}\left\langle \nabla\widetilde{H}_N(\boldsymbol{n}^\ell) + \nabla\widetilde{H}_N(\boldsymbol{n}^{\ell+1}), \boldsymbol{n}^{\ell+1} - \boldsymbol{n}^\ell \right\rangle\right| \\ &\quad \leq \sum_{\ell=\underline{\ell}}^{\overline{\ell}-1} O\left(\sup_{|x|\leq 2\sqrt{N}} \left\|\nabla^3 \widetilde{H}_N(x)\right\|_{\text{inj}}\right) \cdot \sup_\ell |\boldsymbol{n}^{\ell+1} - \boldsymbol{n}^\ell|^3 \\ &\quad \leq O(N\sqrt{\delta}). \end{aligned}$$

Because $\underline{\ell} \to \infty$ implies $\delta \to 0$ this term vanishes in the limit, and it remains to show

$$\lim_{\bar{q}\to 1}\lim_{\underline{\ell}\to\infty}\operatorname*{p-lim}_{N\to\infty}\sum_{\ell=\underline{\ell}}^{\bar{\ell}-1}\frac{1}{2}\left\langle\nabla\widetilde{H}_N(\boldsymbol{n}^\ell)+\nabla\widetilde{H}_N(\boldsymbol{n}^{\ell+1}),\boldsymbol{n}^{\ell+1}-\boldsymbol{n}^\ell\right\rangle_N=\int_q^1\xi''(t)\mathbb{E}\left[u\left(t,X_t\right)\right]\mathrm{d}t.$$

Next, observe by (4.3.13) that:

$$\nabla\widetilde{H}_N(\boldsymbol{n}^\ell)=\boldsymbol{z}^{\ell+1}-\sum_{j=0}^{\ell}d_{\ell,j}\boldsymbol{n}^{j-1}. \tag{4.6.14}$$

Passing to the limiting Gaussian process $(Z_k^\delta)_{k\in\mathbb{Z}^+}$ via state evolution, and ignoring for now the constant number of branching updates,

$$\begin{aligned}\operatorname*{p-lim}_{N\to\infty}\left\langle\nabla\widetilde{H}_N(\boldsymbol{n}^\ell),\boldsymbol{n}^{\ell+1}-\boldsymbol{n}^\ell\right\rangle_N&=\mathbb{E}\left[Z_{\ell+1}^\delta(N_{\ell+1}^\delta-N_\ell^\delta)\right]-\sum_{j=0}^{\ell}d_{\ell,j}\mathbb{E}\left[N_{j-1}^\delta(N_{\ell+1}^\delta-N_\ell^\delta)\right],\\ \operatorname*{p-lim}_{N\to\infty}\left\langle\nabla\widetilde{H}_N(\boldsymbol{n}^{\ell+1}),\boldsymbol{n}^{\ell+1}-\boldsymbol{n}^\ell\right\rangle_N&=\mathbb{E}\left[Z_{\ell+2}^\delta(N_{\ell+1}^\delta-N_\ell^\delta)\right]-\sum_{j=0}^{\ell+1}d_{\ell+1,j}\mathbb{E}\left[N_{j-1}^\delta(N_{\ell+1}^\delta-N_\ell^\delta)\right].\end{aligned}$$

As $(N_k^\delta)_{k\geq\mathbb{Z}^+}$ is a martingale process, it follows that the right-hand expectations all vanish. Similarly it holds that

$$\begin{aligned}\mathbb{E}[Z_{\ell+2}^\delta(N_{\ell+1}^\delta-N_\ell^\delta)]&=\mathbb{E}[Z_{\ell+1}^\delta(N_{\ell+1}^\delta-N_\ell^\delta)]\\ \mathbb{E}[Z_\ell^\delta(N_{\ell+1}^\delta-N_\ell^\delta)]&=0.\end{aligned}$$

Rewriting and using Lemma 4.6.3 in the last step,

$$\begin{aligned}\operatorname*{p-lim}_{N\to\infty}\frac{1}{2}\left\langle\nabla\widetilde{H}_N(\boldsymbol{n}^\ell)+\nabla\widetilde{H}_N(\boldsymbol{n}^{\ell+1}),\boldsymbol{n}^{\ell+1}-\boldsymbol{n}^\ell\right\rangle_N&=\mathbb{E}[(Z_{\ell+1}^\delta-Z_\ell^\delta)(N_{\ell+1}^\delta-N_\ell^\delta)]\\ &=\mathbb{E}[u_\ell^\delta(X_\ell^\delta)(Z_{\ell+1}^\delta-Z_\ell^\delta)^2]\\ &=\mathbb{E}\left[\mathbb{E}[u_\ell^\delta(X_\ell^\delta)(Z_{\ell+1}^\delta-Z_\ell^\delta)^2|\mathcal{F}_\ell^\delta]\right]\\ &=(\xi'(q_{\ell+1}^\delta)-\xi'(q_\ell^\delta))\cdot\mathbb{E}[u_\ell^\delta(X_\ell^\delta)]\\ &=(\xi'(q_{\ell+1}^\delta)-\xi'(q_\ell^\delta))\cdot\frac{\mathbb{E}[u_{q_\ell^\delta}(X_\ell^\delta)]}{\Sigma_\ell^\delta}.\end{aligned}$$

Recalling (4.6.12), the fact that $u_t(x)$ is uniformly Lipschitz in x for $t\in[0,\bar{q}]$, the fact that $\xi'(q_{\ell+1}^\delta)-\xi'(q_\ell^\delta)=\delta\xi''(q_\ell^\delta)+O(\delta^2)$ and the coupling of Lemma 4.6.4, it follows that

$$\operatorname*{p-lim}_{N\to\infty}\frac{1}{2}\left\langle\nabla\widetilde{H}_N(\boldsymbol{n}^\ell)+\nabla\widetilde{H}_N(\boldsymbol{n}^{\ell+1}),\boldsymbol{n}^{\ell+1}-\boldsymbol{n}^\ell\right\rangle_N=\delta\xi''(q_\ell^\delta)\mathbb{E}[u_{q_\ell^\delta}(X_{q_\ell^\delta})]+O_{\bar{q}}(\delta^{3/2}).$$

Summing over ℓ and using continuity of $u_t(x)$ in t, it follows that

$$\lim_{\underline{\ell}\to\infty} \operatorname*{p-lim}_{N\to\infty} \frac{\widetilde{H}_N(\boldsymbol{n}^{\overline{\ell}}) - \widetilde{H}_N(\boldsymbol{n}^{\underline{\ell}})}{N} = \int_{\underline{q}}^{\bar{q}} \xi''(t)\mathbb{E}[u(t, X_t)]\mathrm{d}t.$$

Sending $\bar{q} \to 1$ now concludes the proof when there are no branching steps. Extending the proof to cover branching steps is not difficult and we explain it now. Everything up to (4.6.14) is still valid, and if the number $|Q|$ of branching steps is m, then the full analysis applies to all but m terms. However the simple uniform bound

$$\begin{aligned}\widetilde{H}_N(\boldsymbol{n}^{\ell+1}) - \widetilde{H}_N(\boldsymbol{n}^{\ell}) &\le \|\boldsymbol{n}^{\ell+1} - \boldsymbol{n}^{\ell}\| \cdot \sup_{\|\boldsymbol{x}\|_N \le 1+\frac{\eta}{2}} \|\nabla \widetilde{H}_N(\boldsymbol{x})\| \\ &\le O(\sqrt{N\delta}) \cdot O(\sqrt{N}) \\ &\le O(N\sqrt{\delta})\end{aligned}$$

holds with high probability. Here we have used Proposition 4.2.6 to deduce $\|\boldsymbol{n}^{\ell}\|_N, \|\boldsymbol{n}^{\ell+1}\|_N \le 1 + o(1)$ with high probability, and also Proposition 4.2.1 and Equation (4.6.6). Therefore all telescoping terms, branching or not, uniformly contribute $O(N\sqrt{\delta})$ energy in probability. As a result, even when a constant number of non-branching terms are replaced by branching terms, the same analysis applies up to error $O(N\sqrt{\delta})$, yielding the same asymptotic energy for branching $\underline{q}$-IAMP and completing the proof. □

Chapter 5

Tight Lipschitz Hardness for Optimizing Mean-Field Spin Glasses

5.1 Introduction

For each $p \in 2\mathbb{N}$, let $\mathbf{G}^{(p)} \in \left(\mathbb{R}^N\right)^{\otimes p}$ be an independent p-tensor with i.i.d. $\mathcal{N}(0,1)$ entries. Let $h \geq 0$, and set $\boldsymbol{h} = (h, \dots, h) \in \mathbb{R}^N$. Fix a sequence $(\gamma_p)_{p \in 2\mathbb{N}}$ with $\gamma_p \geq 0$ and $\sum_{p \in 2\mathbb{N}} 2^p \gamma_p^2 < \infty$. The mixed even p-spin Hamiltonian H_N is defined by

$$H_N(\boldsymbol{\sigma}) = \langle \boldsymbol{h}, \boldsymbol{\sigma} \rangle + \widetilde{H}_N(\boldsymbol{\sigma}), \quad \text{where} \tag{5.1.1}$$

$$\widetilde{H}_N(\boldsymbol{\sigma}) = \sum_{p \in 2\mathbb{N}} \frac{\gamma_p}{N^{(p-1)/2}} \langle \mathbf{G}^{(p)}, \boldsymbol{\sigma}^{\otimes p} \rangle. \tag{5.1.2}$$

We consider inputs $\boldsymbol{\sigma}$ in either the sphere $S_N = \{\boldsymbol{\sigma} \in \mathbb{R}^N : \sum_{i=1}^N \sigma_i^2 = N\}$ or the cube $\Sigma_N = \{-1, 1\}^N$. These define, respectively, the *spherical* and *Ising* mixed p-spin glass models. The coefficients γ_p are customarily encoded in the *mixture function* $\xi(x) = \sum_{p \in 2\mathbb{N}} \gamma_p^2 x^p$. Note that $\widetilde{H}_N$ is equivalently described as the Gaussian process with covariance

$$\mathbb{E}\, \widetilde{H}_N(\boldsymbol{\sigma}^1) \widetilde{H}_N(\boldsymbol{\sigma}^2) = N\xi(\langle \boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2 \rangle / N).$$

Our purpose is to shed light on a discrepancy between the in-probability limiting maximum

values

$$\mathsf{OPT}^{\mathrm{Sp}}_{\xi,h} = \operatorname*{p-lim}_{N\to\infty} \frac{1}{N} \max_{\boldsymbol{\sigma}\in S_N} H_N(\boldsymbol{\sigma}), \qquad \mathsf{OPT}^{\mathrm{Is}}_{\xi,h} = \operatorname*{p-lim}_{N\to\infty} \frac{1}{N} \max_{\boldsymbol{\sigma}\in \Sigma_N} H_N(\boldsymbol{\sigma})$$

and the maximum *efficiently computable* values of H_N over the same sets. We will write $\mathsf{OPT}^{\mathrm{Sp}} = \mathsf{OPT}^{\mathrm{Sp}}_{\xi,h}$ and $\mathsf{OPT}^{\mathrm{Is}} = \mathsf{OPT}^{\mathrm{Is}}_{\xi,h}$ when ξ, h are clear from context. The values $\mathsf{OPT}^{\mathrm{Sp}}$ and $\mathsf{OPT}^{\mathrm{Is}}$ are given by the celebrated Parisi formula [Par79] which was proved for even models by [Tal06d, Tal06a] and in more generality by [Pan14]. While most often stated as a formula for the limiting free energy at inverse temperature β, the asymptotic maximum can be recovered as a $\beta \to \infty$ limit of the Parisi formula. Restricting for concreteness to the Ising case (we will state the analogous result for the spherical case in Section 5.2), the result can be expressed in the following form due to Auffinger and Chen [AC17b].

Define the function space

$$\mathscr{U} = \left\{ \zeta : [0,1) \to \mathbb{R}_{\geq 0} : \zeta \text{ is right-continuous and nondecreasing}, \int_0^1 \zeta(t)\mathrm{d}t < \infty \right\}. \tag{5.1.3}$$

For $\zeta \in \mathscr{U}$, define $\Phi_\zeta : [0,1] \times \mathbb{R} \to \mathbb{R}$ to be the solution of the following *Parisi PDE*.

$$\partial_t \Phi_\zeta(t,x) + \frac{1}{2}\xi''(t)\left(\partial_{xx}\Phi_\zeta(t,x) + \zeta(t)(\partial_x \Phi_\zeta(t,x))^2\right) = 0 \tag{5.1.4}$$

$$\Phi_\zeta(1,x) = |x|. \tag{5.1.5}$$

Existence and uniqueness properties for this PDE are well established and are reviewed in Subsection 5.6.1. The Parisi functional $\mathsf{P}^{\mathrm{Is}} = \mathsf{P}^{\mathrm{Is}}_{\xi,h} : \mathscr{U} \to \mathbb{R}$ is given by

$$\mathsf{P}^{\mathrm{Is}}(\zeta) = \Phi_\zeta(0,h) - \frac{1}{2}\int_0^1 t\xi''(t)\zeta(t)\ \mathrm{d}t. \tag{5.1.6}$$

Theorem 20 ([AC17b, Theorem 1])**.** *The following identity holds.*

$$\mathsf{OPT}^{\mathrm{Is}} = \inf_{\zeta\in\mathscr{U}} \mathsf{P}^{\mathrm{Is}}(\zeta). \tag{5.1.7}$$

The infimum over $\zeta \in \mathscr{U}$ is achieved at a unique $\zeta_* \in \mathscr{U}$ as shown in [AC17b, CHL18], which can be obtained as an appropriately renormalized zero-temperature limit of the corresponding minimizers in the positive temperature Parisi formula. These positive temperature minimizers roughly correspond to cumulative distribution functions for the overlap $\langle \boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2\rangle/N$ of two replicas $\boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2$ sampled from the Gibbs measure $e^{\beta H_N}/Z_N(\beta)$; this is why the functions ζ considered in the Parisi formula are nondecreasing.

We recall the main result of the previous Chapter. For a function $f : \mathbb{R} \to \mathbb{R}$ and interval J, let

$\|f\|_{\mathrm{TV}(J)}$ denote the total variation of f on J, expressed as the supremum over partitions:

$$\|f\|_{\mathrm{TV}(J)} = \sup_n \sup_{t_0<t_1<\cdots<t_n, t_i\in J} \sum_{i=1}^n |f(t_i) - f(t_{i-1})|.$$

Let $\mathscr{L} \supseteq \mathscr{U}$ denote the set of functions given by

$$\mathscr{L} = \left\{ \begin{array}{ll} \zeta : [0,1) \to \mathbb{R}_{\geq 0} : & \zeta \text{ right-continuous}, \|\xi'' \cdot \zeta\|_{\mathrm{TV}[0,t]} < \infty \text{ for all } t \in [0,1), \\ & \displaystyle\int_0^1 \xi''(t)\zeta(t)\ \mathrm{d}t < \infty \end{array} \right\}. \quad (5.1.8)$$

It turns out (see Subsection 5.6.1) that the definition of P^{Is} above extends from $\mathscr{U}$ to $\mathscr{L}$. Therefore we may define $\mathsf{ALG}^{\mathrm{Is}} = \mathsf{ALG}^{\mathrm{Is}}_{\xi,h}$ by

$$\mathsf{ALG}^{\mathrm{Is}} = \inf_{\zeta\in\mathscr{L}} \mathsf{P}^{\mathrm{Is}}(\zeta). \quad (5.1.9)$$

Note that $\mathsf{ALG}^{\mathrm{Is}} \leq \mathsf{OPT}^{\mathrm{Is}}$ trivially holds. We have $\mathsf{ALG}^{\mathrm{Is}} = \mathsf{OPT}^{\mathrm{Is}}$ if the infimum in (5.1.9) is attained by some $\zeta \in \mathscr{U}$, and otherwise $\mathsf{ALG}^{\mathrm{Is}} < \mathsf{OPT}^{\mathrm{Is}}$. The following result was shown in the previous Chapter.

Theorem 21. *Assume there exists $\zeta_* \in \mathscr{L}$ such that $\mathsf{P}^{\mathrm{Is}}(\zeta_*) = \mathsf{ALG}^{\mathrm{Is}}$. Then for any $\varepsilon > 0$, there exists an efficient AMP algorithm $\mathcal{A} : \mathcal{H}_N \to C_N$ such that*

$$\mathbb{P}[H_N(\mathcal{A}(H_N))/N \geq \mathsf{ALG}^{\mathrm{Is}} - \varepsilon] \geq 1 - o(1), \quad c = c(\varepsilon) > 0.$$

Recall that the AMP algorithms of the previous Chapter use only a constant number of queries of $\nabla H_N(\cdot)$; this results in computation time linear in the description length of H_N when ξ is a polynomial, assuming oracle access to ζ_* and the function Φ_{ζ_*}. Because of the natural condition under which ALG and OPT are equal, one might conjecture that the aforementioned AMP algorithms achieve the best asymptotic energy possible for efficient algorithms.

This belief was also aligned with results on the "critical point complexity" of pure spherical spin glasses with $\xi(x) = x^p$ and $h = 0$. In this case, the analogous value $\mathsf{ALG}^{\mathrm{Sp}}$ is the one obtained by [Sub21] and coincides with the onset of exponentially many bounded index critical points, as established in [ABAČ13, Sub17]. In this case almost all local optima have energy value $\mathsf{ALG}^{\mathrm{Sp}} \pm o(1)$ with high probability, which suggests from another direction that exceeding the energy $\mathsf{ALG}^{\mathrm{Sp}}$ might be computationally intractable. On the other hand, it is not clear whether this threshold coincides with $\mathsf{ALG}^{\mathrm{Sp}}$ beyond the pure case.

It unfortunately seems difficult to establish any limitations on the power of general polynomial-time algorithms for such a task. However one might still hope to characterize the power of natural classes of algorithms that include gradient descent and AMP. To this end, we define the following

distance on the space $\mathcal{H}_N$ of Hamiltonians H_N. We identify H_N with its disorder coefficients $(\mathbf{G}^{(p)})_{p\in 2\mathbb{N}}$, which we concatenate (in an arbitrary but fixed order) into an infinite vector $\mathbf{g}(H_N)$. We equip $\mathcal{H}_N$ with the (possibly infinite) distance

$$\|H_N - H'_N\|_2 = \|\mathbf{g}(H_N) - \mathbf{g}(H'_N)\|_2.$$

Let $B_N = \{\boldsymbol{\sigma} \in \mathbb{R}^N : \sum_{i=1}^N \sigma_i^2 \le N\}$ and $C_N = [-1,1]^N$ be the convex hulls of S_N and Σ_N, which we equip with the standard $\|\cdot\|_2$ distance. A consequence of our main result is that no suitably *Lipschitz* function $\mathcal{A} : \mathcal{H}_N \to C_N$ can surpass the asymptotic value $\mathsf{ALG}^{\mathrm{Is}}$. (And similarly in the spherical case for $\mathcal{A} : \mathcal{H}_N \to B_N$ and an analogous $\mathsf{ALG}^{\mathrm{Sp}}$.)

Theorem 22. *Let $\tau, \varepsilon > 0$ be constants. For N sufficiently large, any τ-Lipschitz $\mathcal{A} : \mathcal{H}_N \to C_N$ satisfies*

$$\mathbb{P}\left[H_N(\mathcal{A}(H_N))/N \ge \mathsf{ALG}^{\mathrm{Is}} + \varepsilon\right] \le \exp(-cN), \quad c = c(\xi, h, \varepsilon, \tau) > 0.$$

Note that the Lipschitz condition $\|\mathcal{A}(H_N) - \mathcal{A}(H'_N)\|_2 \le \tau \|H_N - H'_N\|_2$ holds vacuously when the latter distance is infinite.

The IAMP algorithms of the previous Chapter are $O(1)$-Lipschitz in the sense above[1]. While the approach of [Sub21] is not Lipschitz, its performance is captured by AMP as explained in [AMS21, Remark 2.2].[2] Hence in tandem with these constructive results, Theorem 22 identifies the exact asymptotic value achievable by Lipschitz functions $\mathcal{A} : \mathcal{H}_N \to C_N$ (assuming the existence of a minimizer $\zeta_* \in \mathscr{L}$ as required in Theorem 21). We also give an analogous result for spherical spin glasses, in which there is no question of existence of a minimizer on the algorithmic side. Let us remark that the rate e^{-cN} in Theorem 22 is best possible up to the value of c, being achieved even for the trivial algorithm $\mathcal{A}(H_N) = (1, 1, \ldots, 1)$ which ignores its input entirely.

Abstractly, the assumption that $\mathcal{A}$ is Lipschitz is geometrically natural and brings us near the well-studied setting of Lipschitz selection [Shv84, PY95, Shv02, FS18]. Here one is given a metrized family $\mathcal{S}$ of subsets inside a metric space X. The goal is to find a function $f : \mathcal{S} \to X$ with the selection property that $f(S) \in S$ for all $S \in \mathcal{S}$, and such that f has a small Lipschitz constant. Indeed a Lipschitz function $\mathcal{A} : \mathcal{H}_N \to C_N$ achieving energy E is almost the same as a Lipschitz selector for the level sets

$$S_E(H_N) = \{\boldsymbol{\sigma} \in C_N : H_N(\boldsymbol{\sigma})/N \ge E\}$$

metrized by the norm on $\mathcal{H}_N$ defined above (and leaving aside the fact that $S_E(H_N)$ may not determine H_N). Of course we can only hope for $\mathcal{A}(H_N) \in S_E(H_N)$ to hold with high probability, since $S_E(H_N)$ is empty with small but positive probability. See the next Chapter for a problem in *online* Lipschitz selection.

[1] Technically we rounded the output to the discrete set Σ_N at the end, making the algorithm discontinuous. Removing the rounding step yields Lipschitz maps $\mathcal{A} : \mathcal{H}_N \to C_N$ with the same performance.

[2] We also outline a similar impossibility result for a family of variants of [Sub21] in Subsection 5.3.7.

Many natural optimization algorithms satisfy the Lipschitz property above on a set $K_N \subseteq \mathcal{H}_N$ of inputs with $1 - \exp(-\Omega(N))$ probability; this suffices just as well for Theorem 22 thanks to the Kirszbraun extension theorem (see Subsection C.1.1). As explained in Section C.1, algorithms with this property include the following examples, all run for a constant (i.e. dimension-independent) number of iterations or amount of time.

- Gradient descent and natural variants thereof;
- Approximate message passing;
- More general "higher-order" optimization methods with access to $\nabla^k H_N(\cdot)$ for constant k;
- Langevin dynamics for the Gibbs measure $e^{\beta H_N}$ with suitable reflecting boundary conditions and any positive constant β.

In fact we will not require the full Lipschitz assumption on $\mathcal{A}$, but only a consequence that we call overlap concentration. Roughly speaking, overlap concentration of $\mathcal{A}$ means that given any fixed correlation between the disorder coefficients of H_N^1 and H_N^2, the overlap $\langle \mathcal{A}(H_N^1), \mathcal{A}(H_N^2) \rangle / N$ tightly concentrates around its mean. This property holds automatically for τ-Lipschitz $\mathcal{A}$ thanks to concentration of measure on Gaussian space. It also might plausibly be satisfied for some discontinuous algorithms such as the Glauber dynamics.

5.1.1 Further Background

We now describe some other results on algorithmically optimizing spin glass Hamiltonians. First, in the worst case over the disorder $\mathbf{G}^{(p)}$, achieving any constant approximation ratio to the true optimum value is known to be quasi-NP hard even for degree 2 polynomials [ABE$^+$05, BBH$^+$12]. For the Sherrington-Kirkpatrick model with $\xi(t) = t^2/2$ on the cube, it was recently shown to be NP-hard on average to compute the *exact* value of the partition function [GK21b]. Of course, these computational hardness results demand much stronger guarantees than the approximate optimization with high probability that we consider.

Another important line of work, alluded to above, has studied the *complexity* of the landscape of H_N on the sphere, defined as the exponential growth rate for the number of local optima and saddle points of finite-index at a given energy level. These are understood to serve as barriers to efficient optimization, and a non-rigorous study was undertaken in [CLR03, CLR05, Par06] followed by a great deal of recent progress in [ABAČ13, ABA13, Sub17, McK21, Kiv21, SZ21]. Notably because the true maximum value of H_N is nothing but its largest critical value, the first moment results of [ABAČ13] combined with the second moment results of [Sub17] gave an alternate self-contained proof of the Parisi formula for the ground state in pure spherical models. In a related

spirit, [Cha09, DEZ15, CS17, CHL18] have shown that mixed even p-spin Hamiltonians typically contain exponentially many well-separated near-global maxima.

Other works such as [CK94, BCKM98, BADG06, BAGJ20, CCM21] have studied natural algorithms such as Langevin and Glauber dynamics on short (independent of N) time scales. These approaches yield (often non-rigorous) predictions for the energy achieved after a fixed amount of time. However these predictions involve complicated systems of differential equations, and to the best of our knowledge it is not known how to cleanly describe the long-time limiting energy achieved. Let us also mention the recent results of [EKZ21, AJK$^+$21] showing that the Glauber dynamics for the Sherrington-Kirkpatrick model mix rapidly at high temperature. By contrast the problem of optimization considered in this work is related to the *low* temperature behavior of the model.

5.1.2 The Overlap Gap Property as a Barrier to Algorithms

Optimizing a spin glass Hamiltonian is one example of a *random optimization problem*, where one aims to find an input achieving a large value for a random objective function. These problems include finding a large independent set in a random graph, the Number Partitioning Problem, and constraint satisfaction problems (CSPs) such as random k-SAT and q-coloring a random graph.

Like spin glass optimization, random optimization problems often have information-computation gaps, where the maximum objective that exists is larger than the maximum objective that can be found by any known efficient algorithm. Since the early 2000s, there has emerged a large body of evidence that suggests that these gaps are inherent. This evidence has also produced heuristics predicting the optimal objective achieved by efficient algorithms.

We will focus on a line of work linking the failure of algorithms to phase transitions in the problem's solution geometry. One version of this connection was proposed in [ACO08, COE15] based on a *shattering* phase transition: at large constraint density the solution space breaks into exponentially many small components, for suitable random instances of k-SAT, q-coloring, and maximum independent set. Shattering defeats local search heuristics, suggesting that polynomial-time algorithms should not succeed. Other predictions based on the *clustering*, *condensation* [KMRT$^+$07] and *freezing* [ZK07] transitions have also been suggested.

Another recent line of work [GS14, RV17a, GS17, CGPR19, GJ21, GJW20a, Wei22, GK21a, BH21, GJW21] on the Overlap Gap Property (OGP) has made substantial progress on *rigorously* linking solution geometry clustering in random optimization problems to algorithmic hardness. A survey can be found in [Gam21]. Initiated by Gamarnik and Sudan in [GS14], this line of work formalizes clustering as an "overlap gap," the absence of a pair of solutions with medium overlap, and proves that this condition implies failure of various classes of stable algorithms. In its original form, an OGP argument consists of two parts. First, it shows that above some constraint density or

objective value, with high probability there exists no pair of solutions with medium overlap. Then, it argues that a stable algorithm solving the problem can be used to construct such a solution pair, implying that such an algorithm cannot exist if the OGP holds. An important difference from the predictions above is that the shattering, clustering, and freezing transitions describe properties of a *typical* solution, while an OGP requires that solution pairs with medium overlap *do not exist at all*.

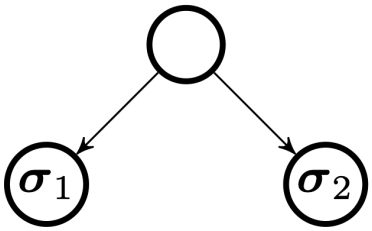
Over many problems, a pattern has emerged where the classic OGP argument shows the failure of stable algorithms above an intermediate objective value, smaller than the maximum objective that exists but larger than the maximum objective that algorithms can find. To improve the value beyond which stable algorithms are proven to fail, subsequent works have considered "multi-OGPs": enhancements of OGP that use more complex forbidden structures involving more than two solutions. The strategy is similar: one shows that an appropriately chosen structure of solutions does not occur with high probability, and that a putative stable algorithm solving the problem can be used to construct this structure. This technique improves on the classic OGP if the structure becomes forbidden at a lower objective value than the classic OGP forbidden structure.

Multi-OGPs have the potential to show nearly-tight algorithmic hardness for stable algorithms. For maximum independent sets on $G(N, d/N)$ in the limit $N \to \infty$ followed by $d \to \infty$, [Wei22] proved by a multi-OGP that *low degree polynomial* algorithms cannot attain any objective asymptotically larger than the believed algorithmic limit. Previously [RV17a] showed the same impossibility result for the more restricted class of *local* algorithms. Similarly, [BH21] used a multi-OGP to prove that low degree polynomials fail to solve random k-SAT at a clause density a constant factor above where algorithms are known to succeed.

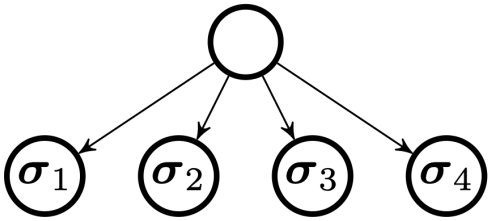
For *pure* spherical and Ising p-spin glasses where $h = 0$ and $p \geq 4$ is even, $\mathsf{ALG} < \mathsf{OPT}$ always holds (recall (5.1.7), (5.1.9)). In such models, [GJW20a] showed using a (2-solution) OGP that low degree polynomials cannot achieve some objective strictly smaller than OPT, extending a similar hardness result of [GJ21] for approximate message passing. [GJW21] extended the conclusions of [GJW20a] to Boolean circuits of depth less than $\frac{\log n}{2 \log \log n}$. As pointed out in [Sel21b, Section 6], these results extend in the Ising case to any mixed even model where $\mathsf{ALG}^{\mathrm{Is}} < \mathsf{OPT}^{\mathrm{Is}}$. In this chapter, we will use a multi-OGP to show that overlap concentrated algorithms cannot optimize mixed even spherical or Ising spin glasses to any objective larger than ALG.

The design of our multi-OGP is a significant departure from previous work. Previous OGPs and multi-OGPs all use one of the following three forbidden structures, see Figure 5.1.

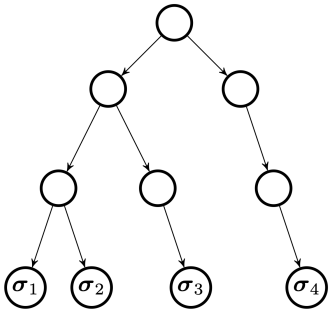
- Classic OGP: two solutions with medium overlap [GS14, CGPR19, GJ21, GJW20a, GJW21].
- Star OGP: several solutions with approximately equal pairwise overlap [RV17a, GS17, GK21a].
- Ladder OGP: several solutions, where the i-th solution ($i \geq 2$) has medium "multi-overlap" with the first $i - 1$ solutions, for a problem-specific notion of multi-overlap [Wei22, BH21].



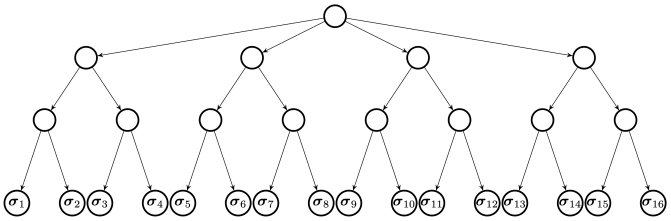
(a) **Classic OGP**: $\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2$ have medium overlap.



(b) **Star OGP**: many solutions, medium overlaps.



(c) **Ladder OGP**: medium "multi-overlaps" between $\boldsymbol{\sigma}_i$ and $\{\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_{i-1}\}$.



(d) **Branching OGP**: many solutions in an ultrametric tree.

Figure 5.1: Schematics of forbidden structures in OGP arguments.

In contrast, the forbidden structure in our multi-OGP is an arbitrarily complicated ultrametric branching tree of solutions. We call this the *Branching OGP*. Informally, the Branching OGP is the condition that for any fixed $\varepsilon > 0$, no constellation of configurations with a certain ultrametric overlap structure has average energy $\mathsf{ALG} + \varepsilon$. The definition involves a family of "ultrametrically correlated" Hamiltonians, with one input in the constellation per Hamiltonian.

We establish this branching OGP as follows. Using a version of the Guerra-Talagrand interpolation, which we take to zero temperature, we derive an upper bound for the maximum average energy of configurations arranged into the desired structure. This upper bound is a multi-dimensional analogue of the Parisi formula, and depends on an essentially arbitrary increasing function $\zeta : [0, 1] \to \mathbb{R}^+$ (which we are free to minimize over). We show that for a symmetric branching tree, the resulting estimate can be upper bounded by $\mathsf{P}(\kappa\zeta)$. Here P is the Parisi functional P^{Is} or its spherical analogue P^{Sp}, and κ is a decreasing piecewise-constant function that depends on the tree. By making the tree branch rapidly, the function κ can be arranged to decrease as rapidly as desired. As a result, the functions $\kappa\zeta$ are dense in the space $\mathscr{L}$. Thus, we may choose a tree and ζ such that $\mathsf{P}(\kappa\zeta)$ is arbitrarily close to ALG.

Roughly speaking, we show that an overlap concentrated $\mathcal{A}$ allows the construction of an arbitrary ultrametric constellation of outputs. Consequently, if $\mathcal{A}$ outputs points with energy at least $\mathsf{ALG}+\varepsilon$, then $\mathcal{A}$ run on the appropriate family of ultrametrically correlated Hamiltonians will output the forbidden structure above, a contradiction. Some additional complications are created by the fact that $\mathbb{E}[\mathcal{A}(H_N)]$ may be arbitrary, and that $\mathcal{A}(H_N)$ may be in the interior of C_N (or in the spherical case, B_N). The former issue requires us to control the maximum average energy of ultrametric constellations of points that all have approximately a fixed overlap with $\mathbb{E}[\mathcal{A}(H_N)]$. We deal with the latter issue by composing $\mathcal{A}$ with an additional phase that grows each output of $\mathcal{A}$ into its own ultrametric tree of points in Σ_N (or S_N), so that the resulting set of points has the forbidden ultrametric structure.

We also show that the full strength of the branching OGP is necessary to establish Lipschitz hardness at all objectives above ALG, in the sense that any less complex ultrametric structure fails to be forbidden at an energy bounded away from ALG. More precisely, consider a spherical model ξ without external field; we restrict to this case for convenience. Consider a fixed ultrametric overlap structure of inputs, whose corresponding rooted tree (cf. Subsection 5.7.2) does not contain a full depth-D binary tree. We prove that if $\mathsf{ALG}^{\mathrm{Sp}} < \mathsf{OPT}^{\mathrm{Sp}}$, with high probability there exists a constellation of inputs with this overlap structure where each input achieves energy at least $\mathsf{ALG}^{\mathrm{Sp}} + \varepsilon_{\xi,D}$, for a constant $\varepsilon_{\xi,D} > 0$ depending only on ξ, D.

Remark 5.1.1. To our knowledge, this is the first hardness result in any natural random optimization problem that is tight in the strong sense of characterizing the exact point ALG where hardness occurs. The aforementioned hardness results for maximum independent set on $G(N, d/N)$ are tight in the sense of matching the best algorithms within a $1 + o_d(1)$ factor in the limit $d \to \infty$. In fact, prior to this work, all outstanding *predictions* for the algorithmic threshold in any random optimization problem have only matched the best algorithms within a $1 + o_d(1)$ factor in the large-degree limit. Consequently we believe that the branching OGP elucidates the fundamental reason for algorithmic hardness and may provide a framework for exact algorithmic thresholds in other random optimization problems.

Remark 5.1.2. The significance of ultrametricity in mean-field spin glasses began with [Par79] and has played an enormous role in guiding the mathematical understanding of the low temperature regime in works such as [Rue87, Pan13a, Jag17, CS21]. Ultrametricity also appears naturally in the context of optimization algorithms. Indeed in [Sub21, Remark 6], [AM20, Section 3.4] and [Sel21b, Theorem 4] it was realized that the aforementioned algorithms achieving asymptotic energy ALG are capable of more. Namely, they can construct arbitrary ultrametric constellations of solutions (subject to a suitable diameter upper bound), each with energy ALG. Our proof via branching OGP establishes a sharp converse — the existence of essentially arbitrary ultrametric configurations at a given energy level is *equivalent* to achievability by Lipschitz $\mathcal{A}$.

Remark 5.1.3. Since the algorithm of Subag in [Sub21] uses the top eigenvector of the Hessian $\nabla^2 H_N(\boldsymbol{x})$ for various $\boldsymbol{x} \in B_N$, it is not Lipschitz in H_N in the sense we require. However a different branching OGP argument shows that a stylized class of algorithms which includes a natural variant of Subag's approach is also incapable of achieving energy $\mathsf{ALG} + \varepsilon$. This argument uses only a single Hamiltonian, constructing a branching tree structure using the internal randomness of the algorithm. In this sense, it bears resemblance to the original OGP analysis of [GS14]. An outline is given in Subsection 5.3.7.

5.2 The Optimal Energy of Overlap Concentrated Algorithms

5.2.1 Overlap Concentrated Algorithms

For any $p \in [0,1]$, we may construct two correlated copies $H_N^{(1)}, H_N^{(2)}$ of H_N as follows. Construct three i.i.d. Hamiltonians $\widetilde{H}_N^{[0]}, \widetilde{H}_N^{[1]}, \widetilde{H}_N^{[2]}$ with mixture ξ, as in (5.1.2). Let

$$\widetilde{H}_N^{(1)} = \sqrt{p}\widetilde{H}_N^{[0]} + \sqrt{1-p}\widetilde{H}_N^{[1]} \qquad \text{and} \qquad \widetilde{H}_N^{(2)} = \sqrt{p}\widetilde{H}_N^{[0]} + \sqrt{1-p}\widetilde{H}_N^{[2]}$$

and define

$$H_N^{(1)}(\boldsymbol{\sigma}) = \langle \boldsymbol{h}, \boldsymbol{\sigma} \rangle + \widetilde{H}_N^{(1)}(\boldsymbol{\sigma}) \qquad \text{and} \qquad H_N^{(2)}(\boldsymbol{\sigma}) = \langle \boldsymbol{h}, \boldsymbol{\sigma} \rangle + \widetilde{H}_N^{(2)}(\boldsymbol{\sigma}).$$

We say the pair of Hamiltonians $H_N^{(1)}, H_N^{(2)}$ is p-correlated. Note that pairs of corresponding entries in $\mathbf{g}^{(1)} = \mathbf{g}(H_N^{(1)})$ and $\mathbf{g}^{(2)} = \mathbf{g}(H_N^{(2)})$ are Gaussian with covariance $\left[\begin{smallmatrix} 1 & p \\ p & 1 \end{smallmatrix}\right]$.

We will determine the maximum energy attained by algorithms $\mathcal{A}_N : \mathcal{H}_N \to B_N$ or $\mathcal{A}_N : \mathcal{H}_N \to C_N$ (always assumed to be measurable) obeying the following overlap concentration property.

Definition 5.2.1. *Let $\lambda, \nu > 0$. An algorithm $\mathcal{A} = \mathcal{A}_N$ is (λ, ν) overlap concentrated if for any $p \in [0,1]$ and p-correlated Hamiltonians $H_N^{(1)}, H_N^{(2)}$,*

$$\mathbb{P}\left[\left| R\left(\mathcal{A}(H_N^{(1)}), \mathcal{A}(H_N^{(2)})\right) - \mathbb{E}\, R\left(\mathcal{A}(H_N^{(1)}), \mathcal{A}(H_N^{(2)})\right)\right| \geq \lambda\right] \leq \nu. \tag{5.2.1}$$

5.2.2 The Spherical Zero-Temperature Parisi Functional

We introduce a Parisi functional P^{Sp} for the spherical setting, analogous to the Parisi functional P^{Is} for the Ising setting introduced in (5.1.6). Similarly to Theorem 20, Auffinger and Chen [AC17a], see also [CS17], characterize the ground state energy of the spherical spin glass by a variational formula in terms of this Parisi functional. Recall the set $\mathscr{U}$ defined in (5.1.3). Let

$$\mathscr{V}(\xi) = \left\{ (B, \zeta) \in \mathbb{R}^+ \times \mathscr{U} : B > \int_0^1 \xi''(t)\zeta(t)\, \mathrm{d}t \right\}.$$

Define the spherical Parisi functional $\mathsf{P}^{\mathrm{Sp}} = \mathsf{P}^{\mathrm{Sp}}_{\xi,h} : \mathscr{V}(\xi) \to \mathbb{R}$ by

$$\mathsf{P}^{\mathrm{Sp}}(B,\zeta) = \frac{1}{2}\left[\frac{h^2}{B_\zeta(0)} + \int_0^1 \left(\frac{\xi''(t)}{B_\zeta(t)} + B_\zeta(t)\right)\,\mathrm{d}t\right], \tag{5.2.2}$$

where for $t \in [0,1]$

$$B_\zeta(t) = B - \int_t^1 \xi''(q)\zeta(q)\,\mathrm{d}q. \tag{5.2.3}$$

Theorem 23 ([AC17a, Theorem 10])**.** *The following identity holds.*

$$\mathsf{OPT}^{\mathrm{Sp}} = \inf_{(B,\zeta)\in\mathscr{V}(\xi)} \mathsf{P}^{\mathrm{Sp}}(B,\zeta). \tag{5.2.4}$$

The infimum is attained at a unique $(B_*,\zeta_*) \in \mathscr{V}(\xi)$*.*

5.2.3 Main Results

We defined $\mathsf{ALG}^{\mathrm{Is}}$ in (5.1.9) by a non-monotone extension of the variational formula in (5.1.7). We can similarly define $\mathsf{ALG}^{\mathrm{Sp}}$ by a non-monotone extension of (5.2.4). Recall the set $\mathscr{L}$ defined in (5.1.8). Let $\mathscr{K}(\xi) \supseteq \mathscr{V}(\xi)$ denote the set

$$\mathscr{K}(\xi) = \left\{(B,\zeta) \in \mathbb{R}^+ \times \mathscr{L} : B > \int_0^1 \xi''(t)\zeta(t)\,\mathrm{d}t\right\}.$$

The Parisi functional P^{Sp} can clearly be defined on $\mathscr{K}(\xi)$. We define $\mathsf{ALG}^{\mathrm{Sp}} = \mathsf{ALG}^{\mathrm{Sp}}_{\xi,h}$ by

$$\mathsf{ALG}^{\mathrm{Sp}} = \inf_{(B,\zeta)\in\mathscr{K}(\xi)} \mathsf{P}^{\mathrm{Sp}}(B,\zeta). \tag{5.2.5}$$

Note that $\mathsf{ALG}^{\mathrm{Sp}} \le \mathsf{OPT}^{\mathrm{Sp}}$ trivially.

We are now ready to state the main result of this work. We will show that for any mixed even spherical or Ising spin glass, no overlap concentrated algorithm can attain an energy level above the algorithmic thresholds $\mathsf{ALG}^{\mathrm{Sp}}$ and $\mathsf{ALG}^{\mathrm{Is}}$ with nontrivial probability.

Theorem 24 (Main Result)**.** *Consider a mixed even Hamiltonian* H_N *with model* (ξ,h)*. Let* $\mathsf{ALG} = \mathsf{ALG}^{\mathrm{Sp}}$ *(resp.* $\mathsf{ALG}^{\mathrm{Is}}$*). For any* $\varepsilon > 0$ *there are* $\lambda, c, N_0 > 0$ *depending only on* ξ, h, ε *such that the following holds for any* $N \ge N_0$ *and any* $\nu \in [0,1]$*. For any* (λ,ν) *overlap concentrated* $\mathcal{A} = \mathcal{A}_N : \mathcal{H}_N \to B_N$ *(resp.* C_N*),*

$$\mathbb{P}\left[\frac{1}{N}H_N\left(\mathcal{A}(H_N)\right) \ge \mathsf{ALG} + \varepsilon\right] \le \exp(-cN) + 3(\nu/\lambda)^c.$$

Remark 5.2.1. If $\mathcal{A}$ is τ-Lipschitz, (λ,ν) overlap concentration holds with $\nu = \exp(-c_{\lambda,\tau}N)$ by

concentration of measure on Gaussian space, see Proposition C.1.2. Hence in this case the probability on the right-hand side above is exponentially small in N. The same property holds when $\mathcal{A}$ is τ-Lipschitz on a set of inputs with $1 - \exp(-\Omega(N))$ probability, see Proposition C.1.3.

In tandem with Theorem 21 and its spherical analogue Theorem 25 below, Theorem 37 exactly characterizes the maximum energy attained by overlap concentrated algorithms (again with the caveat on the algorithmic side in the Ising case that a minimizer $\gamma_* \in \mathscr{L}$ exists in Theorem 21). We will see in Section C.1 that the algorithms in these two theorems are overlap concentrated.

Theorem 25 ([AMS21, Sel21b])**.** *For any $\varepsilon > 0$, there exists an efficient and $O_\varepsilon(1)$-Lipschitz AMP algorithm $\mathcal{A} : \mathcal{H}_N \to B_N$ such that*

$$\mathbb{P}[H_N(\mathcal{A}(H_N))/N \geq \mathsf{ALG}^{\mathrm{Sp}} - \varepsilon] \geq 1 - \exp(-cN), \quad c = c(\varepsilon) > 0.$$

In the case of the spherical spin glass, the value of $\mathsf{ALG}^{\mathrm{Sp}}$ is explicit, and is given by the following proposition. We will prove this proposition in Appendix C.3.

Proposition 5.2.2. *If $h^2 + \xi'(1) \geq \xi''(1)$, then*

$$\mathsf{ALG}^{\mathrm{Sp}} = (h^2 + \xi'(1))^{1/2},$$

and the infimum in (5.2.5) *is uniquely attained by $B = (h^2 + \xi'(1))^{1/2}$, $\zeta = 0$. Otherwise,*

$$\mathsf{ALG}^{\mathrm{Sp}} = \widehat{q}\xi''(\widehat{q})^{1/2} + \int_{\widehat{q}}^1 \xi''(q)^{1/2}\, \mathrm{d}q$$

where $\widehat{q} \in [0, 1)$ is the unique number satisfying $h^2 + \xi'(\widehat{q}) = \widehat{q}\xi''(\widehat{q})$. If $h > 0$, the infimum in (5.2.5) *is uniquely attained by $B = \xi''(1)^{1/2}$ and*

$$\zeta(q) = \mathbb{I}\{q \geq \widehat{q}\}\frac{\xi'''(q)}{2\xi''(q)^{3/2}} = -\mathbb{I}\{q \geq \widehat{q}\}\frac{\mathrm{d}}{\mathrm{d}q}\xi''(q)^{-1/2}. \tag{5.2.6}$$

If $h = 0$, the infimum is not attained. It is achieved by $B = \xi''(1)^{1/2}$ and ζ given by (5.2.6) *in the limit as $\widehat{q} \to 0^+$.*[3]

Note that $\mathsf{ALG}^{\mathrm{Sp}} = \mathsf{OPT}^{\mathrm{Sp}}$ if and only if the infimum in (5.2.5) is attained at a pair $(B, \zeta) \in \mathscr{V}(\xi)$. Thus, Proposition 5.2.2 implies that $\mathsf{ALG}^{\mathrm{Sp}} = \mathsf{OPT}^{\mathrm{Sp}}$ if and only if $h^2 + \xi'(1) \geq \xi''(1)$ or $\xi''(q)^{-1/2}$ is concave on $[\widehat{q}, 1]$. In the former case, the model is replica symmetric at zero temperature; in the latter case it is full replica symmetry breaking on $[\widehat{q}, 1]$ at zero temperature. Interestingly, in the case $h^2 + \xi'(1) > \xi''(1)$, [Fyo13, BČNS21] showed that H_N has "trivial complexity": no critical points on S_N with high probability except for the unique global maximizer and minimizer.

[3]When $h = 0$, we cannot take $\widehat{q} = 0$ in (5.2.6) because then $B = \int_0^1 \xi''(q)\zeta(q)\, \mathrm{d}q$, so $(B, \zeta) \notin \mathscr{K}(\xi)$.

In the important case of the pure p-spin model, with $h = 0$ and $\xi(x) = x^p$ for $p \geq 4$ even,

$$\mathsf{ALG}^{\mathrm{Sp}} = \int_0^1 \xi''(q)^{1/2}\, \mathrm{d}q = 2\sqrt{\frac{p-1}{p}}.$$

This coincides with the threshold $E_\infty(p)$ identified in [ABAČ13]. As conjectured in [ABAČ13] and proved in [Sub17], with high probability an overwhelming majority of local maxima of H_N on S_N have energy value $E_\infty(p) \pm o(1)$. This suggests that it may be computationally intractable to achieve energy at least $E_\infty(p) + \varepsilon$ for any $\varepsilon > 0$; our results confirm this hypothesis for overlap concentrated algorithms.

Remark 5.2.2. Our results generalize with no changes in the proofs to arbitrary external fields $\boldsymbol{h} = (h_1, \dots, h_N)$ which are independent of $\widetilde{H}_N$ — one only needs to replace h^2 by $\frac{\|\boldsymbol{h}\|^2}{N}$ in (5.2.2) and replace $\Phi(0, h)$ by $\frac{1}{N}\sum_{i=1}^N \Phi(0, h_i)$ in (5.1.6). This includes for instance the natural case of Gaussian external field $\boldsymbol{h} \sim \mathcal{N}(0, I_N)$. Here $\mathcal{A}$ can depend arbitrarily on $\boldsymbol{h}$ as long as overlap concentration holds conditionally on $\boldsymbol{h}$.

5.2.4 Notation and Preliminaries

We generally use ordinary lower-case letters $(x, y, \dots)$ for scalars and bold lower-case $(\boldsymbol{x}, \boldsymbol{y}, \dots)$ for vectors. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$, we denote the ordinary inner product by $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^N x_i y_i$ and the normalized inner product by $R(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{N}\langle \boldsymbol{x}, \boldsymbol{y} \rangle$. We associate with these inner products the norms $\|\boldsymbol{x}\|_2^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle$ and $\|\boldsymbol{x}\|_N^2 = R(\boldsymbol{x}, \boldsymbol{x})$. There is no confusion between the $\|\cdot\|_N$ norm and the ℓ_p norm, which will not appear in this chapter. We use the standard notations $O(\cdot), \Omega(\cdot), o(\cdot)$ to indicate asymptotic behavior in N.

Ensembles of scalars over an index set $\mathbb{L}$ are denoted with an arrow $(\vec{x}, \vec{y}, \dots)$, and the entry of $\vec{x}$ indexed by $u \in \mathbb{L}$ is denoted $x(u)$. Similarly, ensembles of vectors are written in bold and with an arrow $(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}}, \dots)$, and the entry of $\vec{\boldsymbol{x}}$ indexed by $u \in \mathbb{L}$ are denoted $\boldsymbol{x}(u)$. Sequences of scalars parametrizing these ensembles are also denoted with an arrow, for example $\vec{k} = (k_1, \dots, k_D)$.

We reiterate that $S_N = \{\boldsymbol{x} \in \mathbb{R}^N : \sum_{i=1}^N \boldsymbol{x}_i^2 = N\}$ and $\Sigma_N = \{-1, 1\}^N$, and that $B_N = \{\boldsymbol{x} \in \mathbb{R}^N : \sum_{i=1}^N \boldsymbol{x}_i^2 \leq N\}$ and $C_N = [-1, 1]^N$ are their convex hulls. The space of Hamiltonians H_N is denoted $\mathcal{H}_N$. We identify each Hamiltonian H_N with its disorder coefficients $(\mathbf{G}^{(p)})_{p \in 2\mathbb{N}}$, which we concatenate into a vector $\mathbf{g} = \mathbf{g}(H_N)$.

For any tensor $A_p \in (\mathbb{R}^N)^{\otimes p}$, where $p \geq 1$, we define the operator norm

$$\|A_p\|_{\mathrm{op}} = \frac{1}{N} \max_{\boldsymbol{\sigma}^1, \dots, \boldsymbol{\sigma}^p \in S_N} \left| \langle A_p, \boldsymbol{\sigma}^1 \otimes \cdots \otimes \boldsymbol{\sigma}^p \rangle \right|.$$

Note that when $p = 1$, $\|A_p\|_{\mathrm{op}} = \|A_p\|_N$. The following proposition shows that with exponentially

high probability, the operator norms of all constant-order gradients of H_N are bounded and $O(1)$-Lipschitz. We will prove this proposition in Appendix C.2.

Proposition 5.2.3. *For fixed model (ξ, h) and $r \in [1, \sqrt{2})$, there exists a constant $c > 0$, sequence $(K_N)_{N\geq 1}$ of sets $K_N \subseteq \mathcal{H}_N$, and sequence of constants $(C_k)_{k\geq 1}$ independent of N, such that the following properties hold.*

1. $\mathbb{P}[H_N \in K_N] \geq 1 - e^{-cN}$;
2. *If $H_N \in K_N$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$ satisfy $\|\boldsymbol{x}\|_N, \|\boldsymbol{y}\|_N \leq r$, then*

$$\left\|\nabla^k H_N(\boldsymbol{x})\right\|_{\mathrm{op}} \leq C_k, \tag{5.2.7}$$

$$\left\|\nabla^k H_N(\boldsymbol{x}) - \nabla^k H_N(\boldsymbol{y})\right\|_{\mathrm{op}} \leq C_{k+1}\|\boldsymbol{x} - \boldsymbol{y}\|_N. \tag{5.2.8}$$

Organization. The rest of this chapter is structured as follows. In Section 5.3, we formulate Proposition 5.3.2, which establishes our main branching OGP, and prove Theorem 37 assuming this proposition. Sections 5.4 through 5.6 prove Proposition 5.3.2 using a many-replica version of the Guerra-Talagrand interpolation. Section 5.7 shows that (for spherical models with $h = 0$) the full strength of our branching OGP is necessary to show tight algorithmic hardness. Section C.1 shows that approximately Lipschitz algorithms are overlap concentrated, and that natural optimization algorithms including gradient descent, AMP, and Langevin dynamics are approximately Lipschitz.

5.3 Proof of Main Impossibility Result

In this section, we prove Theorem 37 assuming Proposition 5.3.2, which establishes the main OGP. Throughout, we fix a model (ξ, h) and $\varepsilon > 0$. Let H_N be a Hamiltonian (5.1.1) with model (ξ, h). Let $\lambda > 0$ be a constant we will set later, and let $\mathcal{A} : \mathcal{H}_N \to B_N$ (resp. C_N) be (λ, ν) overlap concentrated.

5.3.1 The Correlation Function

We define the correlation function $\chi : [0, 1] \to \mathbb{R}$ by

$$\chi(p) = \mathbb{E}\, R\left(\mathcal{A}(H_N^{(1)}), \mathcal{A}(H_N^{(2)})\right), \tag{5.3.1}$$

where $H_N^{(1)}, H_N^{(2)}$ are p-correlated copies of H_N. The following proposition establishes several properties of correlation functions, which we will later exploit.

Proposition 5.3.1. *The correlation function χ has the following properties.*

(i) For all $p \in [0,1]$, $\chi(p) \in [0,1]$.

(ii) χ is either strictly increasing or constant on $[0,1]$.

(iii) For all $p \in [0,1]$, $\chi(p) \le (1-p)\chi(0) + p\chi(1)$.

We call any $\chi : [0,1] \to \mathbb{R}$ satisfying the conclusions of Proposition 5.3.1 a *correlation function*.

Proof. In this proof, we will write $\mathcal{A}(\mathbf{g})$ to mean $\mathcal{A}(H_N)$ for the Hamiltonian H_N with disorder coefficients $\mathbf{g} = \mathbf{g}(H_N)$. We introduce the Fourier expansion of $\mathcal{A}$. For each nonnegative integer j, let He_j denote the j-th univariate Hermite polynomial. These are defined by $\mathrm{He}_0(x) = 1$ and for $n \ge 0$,

$$\mathrm{He}_{n+1}(x) = x\mathrm{He}_n(x) - \mathrm{He}_n'(x).$$

Recall that the renormalized Hermite polynomials $\widetilde{\mathrm{He}}_n = \frac{1}{\sqrt{n!}}\mathrm{He}_n$ form an orthonormal basis of $L^2(\mathbb{R})$ with the standard Gaussian measure, i.e. they form a complete basis and satisfy

$$\mathop{\mathbb{E}}_{g\sim\mathcal{N}(0,1)} \widetilde{\mathrm{He}}_n(g)\widetilde{\mathrm{He}}_m(g) = \mathbb{I}[n=m].$$

For each multi-index $\alpha = (\alpha_1, \alpha_2, \dots,)$ of nonnegative integers that are eventually zero, define the multivariate Hermite polynomial

$$\widetilde{\mathrm{He}}_\alpha(\mathbf{g}) = \prod_i \widetilde{\mathrm{He}}_{\alpha_i}(\mathbf{g}_i),$$

These polynomials form an orthonormal basis of $L^2(\mathbb{R}^{\mathbb{N}})$ with the standard Gaussian measure, see e.g. [LMP15, Theorem 8.1.7]. Hence for each $1 \le i \le N$, we can write

$$\mathcal{A}_i(\mathbf{g}) = \sum_\alpha \widehat{\mathcal{A}}_i(\alpha)\widetilde{\mathrm{He}}_\alpha(\mathbf{g}) \qquad \text{where} \qquad \widehat{\mathcal{A}}_i(\alpha) = \mathbb{E}\left[\mathcal{A}_i(\mathbf{g})\widetilde{\mathrm{He}}_\alpha(\mathbf{g})\right].$$

For each multi-index α, let $|\alpha| = \sum_{i\ge1}\alpha_i$. For each nonnegative integer j, introduce the Fourier weight

$$W_j = \frac{1}{N}\sum_{i=1}^N \sum_{|\alpha|=j} \widehat{\mathcal{A}}_i(\alpha)^2 \ge 0.$$

For $i = 1, 2$, let $\mathbf{g}^{(i)} = \mathbf{g}(H_N^{(i)})$. Let T_p denote the Ornstein-Uhlenbeck operator. We compute that

$$\begin{aligned}\chi(p) &= \frac{1}{N}\mathbb{E}\left\langle \mathcal{A}(\mathbf{g}^{(1)}), \mathcal{A}(\mathbf{g}^{(2)})\right\rangle = \frac{1}{N}\mathbb{E}\left\langle \mathcal{A}(\mathbf{g}), T_p\mathcal{A}(\mathbf{g})\right\rangle = \frac{1}{N}\mathbb{E}\left\|T_{\sqrt{p}}\mathcal{A}(\mathbf{g})\right\|_2^2 \\ &= \frac{1}{N}\sum_{i=1}^N \left\|T_{\sqrt{p}}\mathcal{A}_i(\mathbf{g})\right\|_2^2 = \frac{1}{N}\sum_{i=1}^N\sum_\alpha p^{|\alpha|}\widehat{\mathcal{A}}_i(\alpha)^2 = \sum_{j\ge0} p^j W_j.\end{aligned}$$

It is now clear that $0 \le \chi(p) \le \chi(1)$. Since $\chi(1) = \mathbb{E}\left\|\mathcal{A}(H_N)\right\|_N^2 \le 1$, this proves the first claim.

The second claim follows because $\chi(p)$ is strictly increasing unless $W_j = 0$ for all $j \geq 1$, in which case $\chi(p)$ is constant. Finally, the last claim follows since χ is manifestly convex. □

5.3.2 Hierarchically Correlated Hamiltonians

Here we define the hierarchically organized ensemble of correlated Hamiltonians that will play a central role in our proofs of impossibility. Let D be a nonnegative integer and $\vec{k} = (k_1, \ldots, k_D)$ for positive integers $k_1, \ldots, k_D$. For each $0 \leq d \leq D$, let $V_d = [k_1] \times \cdots \times [k_d]$ denote the set of length d sequences with j-th element in $[k_j]$. The set V_0 consists of the empty tuple, which we denote $\emptyset$. Let $\mathbb{T}(\vec{k})$ denote the depth D tree rooted at $\emptyset$ with depth d vertex set V_d, where $u \in V_d$ is the parent of $v \in V_{d+1}$ if u is an initial substring of v. For nodes $u^1, u^2 \in \mathbb{T}(\vec{k})$, let

$$u^1 \wedge u^2 = \max\left\{d \in \mathbb{Z}_{\geq 0} : u^1_{d'} = u^2_{d'} \text{ for all } 1 \leq d' \leq d\right\},$$

where the set on the right-hand side always contains 0 vacuously. This is the depth of the least common ancestor of u^1 and u^2. Let $\mathbb{L}(\vec{k}) = V_D$ denote the set of leaves of $\mathbb{T}(\vec{k})$. When $\vec{k}$ is clear from context, we denote $\mathbb{T}(\vec{k})$ and $\mathbb{L}(\vec{k})$ by $\mathbb{T}$ and $\mathbb{L}$. Finally, let $K = |\mathbb{L}| = \prod_{d=1}^{D} k_d$.

Let sequences $\vec{p} = (p_0, p_1, \ldots, p_D)$ and $\vec{q} = (q_0, q_1, \ldots, q_D)$ satisfy

$$0 = p_0 \leq p_1 \leq \cdots \leq p_D = 1,$$
$$0 \leq q_0 < q_1 < \cdots < q_D = 1.$$

The sequence $\vec{p}$ controls the correlation structure of our ensemble of Hamiltonians, while the sequence $\vec{q}$ controls the overlap structure that we will require the inputs to these Hamiltonians to have.

We now construct an ensemble of Hamiltonians $(H_N^{(u)})_{u \in \mathbb{L}}$, such that each $H_N^{(u)}$ is marginally distributed as H_N and each pair of Hamiltonians $H_N^{(u^1)}, H_N^{(u^2)}$ is $p_{u^1 \wedge u^2}$-correlated. For each $u \in \mathbb{T}$, including non-leaf nodes, let $\widetilde{H}_N^{[u]}$ be an independent copy of $\widetilde{H}_N$, generated by (5.1.2). For each $u \in \mathbb{L}$, we construct

$$H_N^{(u)}(\boldsymbol{\sigma}) = \langle \boldsymbol{h}, \boldsymbol{\sigma} \rangle + \widetilde{H}_N^{(u)}(\boldsymbol{\sigma}), \quad \text{where}$$
$$\widetilde{H}_N^{(u)} = \sum_{d=1}^{D} \sqrt{p_d - p_{d-1}} \cdot \widetilde{H}_N^{[(u_1, \ldots, u_d)]}. \tag{5.3.2}$$

It is clear that this ensemble has the stated properties. Consider a state space of K-tuples

$$\vec{\boldsymbol{\sigma}} = (\boldsymbol{\sigma}(u))_{u \in \mathbb{L}} \in (\mathbb{R}^N)^K.$$

We define a grand Hamiltonian on this state space by

$$\mathcal{H}_N^{\vec{k},\vec{p}}(\vec{\boldsymbol{\sigma}}) \equiv \sum_{u\in\mathbb{L}} H_N^{(u)}(\boldsymbol{\sigma}(u)).$$

We will denote this by $\mathcal{H}_N$ when $\vec{k},\vec{p}$ are clear from context. For states $\vec{\boldsymbol{\sigma}}^1,\vec{\boldsymbol{\sigma}}^2 \in (\mathbb{R}^N)^K$, define the overlap matrix $R = R(\vec{\boldsymbol{\sigma}}^1,\vec{\boldsymbol{\sigma}}^2) \in \mathbb{R}^{K\times K}$ by

$$R_{u^1,u^2} = R(\boldsymbol{\sigma}^1(u^1),\boldsymbol{\sigma}^2(u^2))$$

for all $u^1,u^2 \in \mathbb{L}$. We now define an overlap matrix $Q = Q^{\vec{k},\vec{q}} \in \mathbb{R}^{K\times K}$; we will control the maximum energy of $\mathcal{H}_N$ over inputs $\vec{\boldsymbol{\sigma}}$ with approximately this self-overlap. Let Q have rows and columns indexed by $u^1,u^2 \in \mathbb{L}$ and entries

$$Q_{u^1,u^2} = q_{u^1\wedge u^2}.$$

Fix a point $\boldsymbol{m} \in \mathbb{R}^N$ such that $\|\boldsymbol{m}\|_N^2 = q_0$, which we will later take to be $\boldsymbol{m} = \mathbb{E}[\mathcal{A}(H_N)]$. For a tolerance $\eta \in (0,1)$, define the band

$$B(\boldsymbol{m},\eta) = \left\{\boldsymbol{\sigma}\in\mathbb{R}^N : |R(\boldsymbol{\sigma},\boldsymbol{m}) - q_0| \le \eta\right\}.$$

Define the sets of points in S_N^K and Σ_N^K with self-overlap approximately Q and overlap with $\boldsymbol{m}$ approximately q_0 by

$$\begin{aligned}\mathcal{Q}^{\mathrm{Sp}}(Q,\boldsymbol{m},\eta) &= \left\{\vec{\boldsymbol{\sigma}} \in (S_N\cap B(\boldsymbol{m},\eta))^K : \|R(\vec{\boldsymbol{\sigma}},\vec{\boldsymbol{\sigma}}) - Q\|_\infty \le \eta\right\},\\ \mathcal{Q}^{\mathrm{Is}}(Q,\boldsymbol{m},\eta) &= \left\{\vec{\boldsymbol{\sigma}} \in (\Sigma_N\cap B(\boldsymbol{m},\eta))^K : \|R(\vec{\boldsymbol{\sigma}},\vec{\boldsymbol{\sigma}}) - Q\|_\infty \le \eta\right\}.\end{aligned}$$

Let χ be a correlation function (recall Proposition 5.3.1). We say $\vec{p} = (p_0,\dots,p_D)$ and $\vec{q} = (q_0,\dots,q_D)$ are χ-*aligned* if the following properties hold for all $0 \le d \le D$.

- If $q_d \le \chi(1)$, then $\chi(p_d) = q_d$.
- If $q_d > \chi(1)$, then $p_d = 1$.

The following proposition controls the expected maximum energy of the grand Hamiltonian constrained on the sets $\mathcal{Q}^{\mathrm{Sp}}(Q,\boldsymbol{m},\eta)$ and $\mathcal{Q}^{\mathrm{Is}}(Q,\boldsymbol{m},\eta)$, and is the main ingredient in our proof of impossibility. We defer the proof of this proposition to Sections 5.4 through 5.6.

Proposition 5.3.2. *For any mixed even model (ξ,h) and $\varepsilon > 0$, there exists a small constant $\eta_0 \in (0,1)$ and large constants $N_0,K_0 > 0$, dependent only on ξ,h,ε, such that for all $N \ge N_0$ the following holds.*

Let $\mathsf{ALG} = \mathsf{ALG}^{\mathrm{Sp}}$ *(resp.* $\mathsf{ALG}^{\mathrm{Is}}$*). For any correlation function χ and vector $\boldsymbol{m} \in \mathbb{R}^N$ with*

$\|\boldsymbol{m}\|_N^2 = \chi(0)$, *there exist* $D, \vec{k}, \vec{p}, \vec{q}, \eta$ *as above such that* $\vec{p}$ *and* $\vec{q}$ *are* χ*-aligned,* $\eta \geq \eta_0$, $K \leq K_0$, *and*

$$\frac{1}{N}\,\mathbb{E}\max_{\vec{\boldsymbol{\sigma}} \in \mathcal{Q}(\eta)} \mathcal{H}_N(\vec{\boldsymbol{\sigma}}) \leq K(\mathsf{ALG} + \varepsilon),$$

where $\mathcal{Q}(\eta) = \mathcal{Q}^{\mathrm{Sp}}(Q, \boldsymbol{m}, \eta)$ *(resp.* $\mathcal{Q}^{\mathrm{Is}}(Q, \boldsymbol{m}, \eta)$*).*

5.3.3 Extending a Branching Tree to S_N and Σ_N

To account for the possibility that $\mathcal{A}$ outputs solutions in B_N (resp. C_N) not in S_N (resp. Σ_N), we will show that a branching tree of solutions in B_N (resp. C_N) output by $\mathcal{A}$ can always be extended into a branching tree of solutions in S_N (resp. Σ_N), with only a small cost to the energies attained.

Consider χ-aligned $\vec{p}, \vec{q}$ as above. Let $\underline{D} \leq D$ be the smallest integer such that $p_{\underline{D}} = 1$. Define $\vec{\underline{k}} = (k_1, \ldots, k_{\underline{D}})$, $\vec{\underline{p}} = (p_0, \ldots, p_{\underline{D}})$, and $\vec{\underline{q}} = (q_0, \ldots, q_{\underline{D}})$. Let $\underline{\mathbb{L}} = V_{\underline{D}}$ denote the nodes of $\mathbb{T}$ at depth $\underline{D}$, and let $\underline{K} = |\underline{\mathbb{L}}| = \prod_{d=1}^{\underline{D}} k_d$.

Consider an analogous state space of $\underline{K}$-tuples

$$\vec{\underline{\boldsymbol{\sigma}}} = (\underline{\boldsymbol{\sigma}}(\underline{u}))_{\underline{u} \in \underline{\mathbb{L}}} \in (\mathbb{R}^N)^{\underline{K}}.$$

Define $\underline{Q} = \underline{Q}^{\vec{\underline{k}}, \vec{\underline{q}}} \in \mathbb{R}^{\underline{K} \times \underline{K}}$ analogously as the matrix indexed by $\underline{u}^1, \underline{u}^2 \in \underline{\mathbb{L}}$, where

$$\underline{Q}_{\underline{u}^1, \underline{u}^2} = q_{\underline{u}^1 \wedge \underline{u}^2} \wedge \chi(1).$$

Note that because $\vec{p}, \vec{q}$ are χ-aligned, $q_{\underline{D}-1} < \chi(1) \leq q_{\underline{D}}$. So, the right-hand side is $\chi(1)$ if $\underline{u}^1 \wedge \underline{u}^2 = \underline{D}$ (i.e. $\underline{u}^1 = \underline{u}^2$) and $q_{\underline{u}^1 \wedge \underline{u}^2}$ otherwise. The following sets capture the overlap structure of outputs of $\mathcal{A}$.

$$\underline{\mathcal{Q}}^{\mathrm{Sp}}(\underline{Q}, \boldsymbol{m}, \eta) = \left\{ \vec{\underline{\boldsymbol{\sigma}}} \in (B_N \cap B(\boldsymbol{m}, \eta))^{\underline{K}} : \left\| R(\vec{\underline{\boldsymbol{\sigma}}}, \vec{\underline{\boldsymbol{\sigma}}}) - \underline{Q} \right\|_\infty \leq \eta \right\},$$
$$\underline{\mathcal{Q}}^{\mathrm{Is}}(\underline{Q}, \boldsymbol{m}, \eta) = \left\{ \vec{\underline{\boldsymbol{\sigma}}} \in (C_N \cap B(\boldsymbol{m}, \eta))^{\underline{K}} : \left\| R(\vec{\underline{\boldsymbol{\sigma}}}, \vec{\underline{\boldsymbol{\sigma}}}) - \underline{Q} \right\|_\infty \leq \eta \right\}.$$

By the construction (5.3.2), for each $\underline{u} \in \underline{\mathbb{L}}$ the Hamiltonians

$$\left\{ H_N^{(u)} : u \in \mathbb{L} \text{ is a descendant of } \underline{u} \text{ in } \mathbb{T} \right\}$$

are equal almost surely. Let $H_N^{(\underline{u})}$ denote any representative from this set.

We next define the condition S_{eigen} which guarantees existence of a suitable "extension" $\vec{\boldsymbol{\sigma}}$ of $\vec{\underline{\boldsymbol{\sigma}}} = \left(\mathcal{A}(H_N^{(\underline{u})}) \right)_{\underline{u} \in \underline{\mathbb{L}}}$. First, given a subset $S \subseteq [N]$, denote by W_S the $|S|$ dimensional subspace spanned by the elementary basis vectors $\{e_s : s \in S\}$. Below, λ_j denotes the j-th largest eigenvalue and $(\cdot)|_{W_S}$ denotes restriction to the subspace W_S as a bilinear form, or equivalently $A|_{W_S} = P_{W_S} A P_{W_S}$,

where P_{W_S} is the projection onto W_S.

Definition 5.3.3. *For constants δ and K, let $S_{\text{eigen}}(\delta, K)$ denote the event that both of the below hold for all $\underline{u} \in \underline{\mathbb{L}}$.*

1. $\lambda_{2K+1}\left(\nabla^2 H_N^{(\underline{u})}(\boldsymbol{x})|_{W_S}\right) \geq 0$ *for all $S \subseteq [N]$ of size $|S| \geq \delta N$.*
2. $H_N^{(\underline{u})} \in K_N$, *for the K_N given by Proposition 5.2.3.*

We will use the following lemma, whose proof is deferred to Subsection 5.3.6.

Lemma 5.3.4. *Fix a model ξ, h, constants $\varepsilon, \eta > 0$, and $\vec{k}, \vec{q}$ as above. Let δ be sufficiently small depending on $\xi, h, \eta, \varepsilon$, and assume that $S_{\text{eigen}}(\delta, K)$ holds. For any $\underline{\vec{\boldsymbol{\sigma}}} \in \underline{\mathcal{Q}}(\eta/2)$, there exists $\vec{\boldsymbol{\sigma}} \in \mathcal{Q}(\eta)$ such that*

$$H_N^{(\underline{u})}(\boldsymbol{\sigma}(u)) \geq H_N^{(\underline{u})}(\underline{\boldsymbol{\sigma}}(\underline{u})) - N\varepsilon$$

whenever $\underline{u} \in \underline{\mathbb{L}}$ is an ancestor of $u \in \mathbb{L}$.

5.3.4 Completion of the Proof

We will now finish the proof of Theorem 37. Below we give the proof in the spherical setting; the Ising case follows verbatim up to replacing B_N by C_N and ALG^{Sp} by ALG^{Is} (since $C_N \subseteq B_N$).

Let $\mathsf{ALG} = \mathsf{ALG}^{\text{Sp}}$. Let χ be the correlation function of $\mathcal{A}$ defined in (5.3.1) and set $\boldsymbol{m} = \mathbb{E}[\mathcal{A}(H_N)]$. Note that $\|\boldsymbol{m}\|_N^2 = \chi(0)$ by definition. For small $\varepsilon/2 > 0$ there exist N_0, K_0, η_0 and $D, \vec{k}, \vec{p}, \vec{q}, \eta, K$ as in Proposition 5.3.2 such that

$$\frac{1}{N}\mathbb{E}\max_{\vec{\boldsymbol{\sigma}} \in \mathcal{Q}(\eta)} \mathcal{H}_N(\vec{\boldsymbol{\sigma}}) \leq K(\mathsf{ALG} + \varepsilon/2). \tag{5.3.3}$$

For $N \geq N_0$ let

$$\alpha_N = \mathbb{P}\left[\frac{1}{N} H_N(\mathcal{A}(H_N)) \geq \mathsf{ALG} + \varepsilon\right].$$

For each $\underline{u} \in \underline{\mathbb{L}}$, let $\underline{\boldsymbol{\sigma}}(\underline{u}) = \mathcal{A}(H_N^{(\underline{u})})$, and let $\underline{\vec{\boldsymbol{\sigma}}} = (\underline{\boldsymbol{\sigma}}(\underline{u}))_{\underline{u} \in \underline{\mathbb{L}}}$. We define the following events, where $\delta > 0$ is chosen so that Lemma 5.3.4 holds with parameters $\varepsilon/4, \eta, \vec{k}, \vec{q}$. In the statement of Theorem 37, we take $\lambda = \eta_0/4 \leq \eta/4$.

Define the following events.

$$\begin{aligned} S_{\text{solve}} &= \left\{ \frac{1}{N} H_N^{(\underline{u})}(\underline{\boldsymbol{\sigma}}(\underline{u})) \geq \mathsf{ALG} + \varepsilon \text{ for all } \underline{u} \in \underline{\mathbb{L}}(\vec{k}) \right\}, \\ S_{\text{overlap}} &= \{ \vec{\underline{\boldsymbol{\sigma}}} \in \underline{\mathcal{Q}}(\eta/2) \}, \\ S_{\text{eigen}} &= \{ S_{\text{eigen}}(\delta, K) \}, \\ S_{\text{ogp}} &= \left\{ \frac{1}{N} \max_{\vec{\boldsymbol{\sigma}} \in \mathcal{Q}(\eta)} \mathcal{H}_N(\vec{\boldsymbol{\sigma}}) < K(\mathsf{ALG} + 3\varepsilon/4) \right\}. \end{aligned}$$

Proposition 5.3.5. *With parameters as above,*

$$S_{\text{solve}} \cap S_{\text{overlap}} \cap S_{\text{eigen}} \cap S_{\text{ogp}} = \emptyset.$$

Proof. Suppose that the first three events hold. Then $\mathcal{A}$ outputs $\vec{\underline{\boldsymbol{\sigma}}} \in \underline{\mathcal{Q}}(\eta/2)$ such that for all $\underline{u} \in \underline{\mathbb{L}}$,

$$H_N^{(\underline{u})}(\underline{\boldsymbol{\sigma}}(\underline{u})) \geq \mathsf{ALG} + \varepsilon.$$

Lemma 5.3.4 now implies the existence of $\vec{\boldsymbol{\sigma}} \in \mathcal{Q}(\eta)$ such that for all $\underline{u} \in \underline{\mathbb{L}}$,

$$H_N^{(u)}(\boldsymbol{\sigma}(u)) \geq \mathsf{ALG} + 3\varepsilon/4.$$

This contradicts S_{ogp}. □

Proposition 5.3.6. *The following inequalities hold.*

(a) $\mathbb{P}(S_{\text{solve}}) \geq \alpha_N^K$.

(b) $\mathbb{P}(S_{\text{overlap}}) \geq 1 - K^2\nu - \frac{2K\nu}{\lambda}$.

(c) $\mathbb{P}(S_{\text{eigen}}) \geq 1 - \exp(-cN)$ *for* $c > 0$ *depending only on* ξ, h, ε.

(d) $\mathbb{P}(S_{\text{ogp}}) \geq 1 - 2\exp\left(-\frac{\varepsilon^2}{32\xi(1)}N\right)$.

We defer the proof of this proposition to after the proof of Theorem 37.

Proof of Theorem 37. Lemma 5.3.5 implies that $\mathbb{P}(S_{\text{solve}}) + \mathbb{P}(S_{\text{overlap}}) + \mathbb{P}(S_{\text{eigen}}) + \mathbb{P}(S_{\text{ogp}}) \leq 3$. Because $(K^2 + 2K)^{1/K} \leq 3$ for any positive integer K and $\lambda < 1$,

$$\begin{aligned} \alpha_N &\leq \left(K^2\nu + \frac{2K\nu}{\lambda}\right)^{1/K} + 2\exp\left(-\frac{\varepsilon^2}{32K\xi(1)}N\right) + e^{-cN/K} \\ &\leq 3\left(\frac{\nu}{\lambda}\right)^{1/K} + 2\exp\left(-\frac{\varepsilon^2}{32K\xi(1)}N\right) + e^{-cN/K}. \end{aligned}$$

Recall that $K \le K_0$ and K_0 is a constant depending only on ξ, h, ε. The proof is complete up to choosing an appropriate c in Theorem 37. □

5.3.5 Proofs of Probability Lower Bounds

In this section, we will prove Proposition 5.3.6. As preparation we first give two useful concentration lemmas. The first shows that $R(\mathcal{A}(H_N), \boldsymbol{m})$ concentrates around $\|\boldsymbol{m}\|_N^2$ for overlap concentrated algorithms with $\mathbb{E}[\mathcal{A}(H_N)] = \boldsymbol{m}$.

Lemma 5.3.7. *If $\mathcal{A} = \mathcal{A}_N$ is (λ, ν) overlap concentrated and $\mathbb{E}[\mathcal{A}(H_N)] = \boldsymbol{m}$, then*

$$\mathbb{P}\left[\left|R(\mathcal{A}(H_N), \boldsymbol{m}) - \|\boldsymbol{m}\|_N^2\right| > 2\lambda\right] \le \frac{2\nu}{\lambda}. \tag{5.3.4}$$

Proof. Define the convex function $\psi(t) = (|t - \|\boldsymbol{m}\|_N^2| - \lambda)_+$. Then by Jensen's inequality, for independent Hamiltonians H_N and H_N',

$$\mathbb{E}\left[\psi\left(R(\mathcal{A}(H_N), \boldsymbol{m})\right)\right] \le \mathbb{E}\left[\psi\left(R(\mathcal{A}(H_N), \mathcal{A}(H_N'))\right)\right].$$

Because $\mathcal{A}$ is (λ, ν) overlap concentrated, $\psi\left(R(\mathcal{A}(H_N), \mathcal{A}(H_N'))\right) = 0$ with probability at least $1 - \nu$. Moreover, $\psi\left(R(\mathcal{A}(H_N), \mathcal{A}(H_N'))\right) \le 2$ pointwise. So,

$$\mathbb{E}\left[\psi\left(R(\mathcal{A}(H_N), \boldsymbol{m})\right)\right] \le 2\nu.$$

By Markov's inequality,

$$\mathbb{P}\left[\left|R(\mathcal{A}(H_N), \boldsymbol{m}) - \|\boldsymbol{m}\|_N^2\right| > 2\lambda\right] = \mathbb{P}\left[\psi\left(R(\mathcal{A}(H_N), \boldsymbol{m})\right) > \lambda\right] \le \frac{2\nu}{\lambda}.$$

□

The next lemma shows subgaussian concentration for $\frac{1}{N}\max_{\vec{\boldsymbol{\sigma}} \in \mathcal{Q}(\eta)} \mathcal{H}_N(\vec{\boldsymbol{\sigma}})$.

Proposition 5.3.8. *The random variable*

$$Y = \frac{1}{N} \max_{\vec{\boldsymbol{\sigma}} \in \mathcal{Q}(\eta)} \mathcal{H}_N(\vec{\boldsymbol{\sigma}})$$

satisfies for all $t \ge 0$

$$\mathbb{P}[|Y - \mathbb{E}\, Y| \ge t] \le 2\exp\left(-\frac{Nt^2}{2K^2\xi(1)}\right).$$

Proof. For any $\vec{\boldsymbol{\sigma}} \in S_N^K$, by Cauchy-Schwarz the variance of $\mathcal{H}_N(\vec{\boldsymbol{\sigma}})$ is at most

$$\begin{aligned}\mathbb{E}\left[(\mathcal{H}_N(\vec{\boldsymbol{\sigma}}) - \mathbb{E}\,\mathcal{H}_N(\vec{\boldsymbol{\sigma}}))^2\right] &= \sum_{u^1,u^2\in\mathbb{L}} \mathbb{E}\,\widetilde{H}_N^{(u^1)}(\boldsymbol{\sigma}(u^1))\,\mathbb{E}\,\widetilde{H}_N^{(u^2)}(\boldsymbol{\sigma}(u^2)) \\ &\le K\sum_{u\in\mathbb{L}} \mathbb{E}\,\widetilde{H}_N^{(u)}(\boldsymbol{\sigma}(u))^2 \\ &= NK^2\xi(1).\end{aligned}$$

The result now follows from the Borell-TIS inequality ([Bor75, CIS76], or see [Zei15, Theorem 2]). Note that both the statement and proof of Borell-TIS hold for noncentered Gaussian processes with no modification. □

We now prove each part of Proposition 5.3.6 in turn.

Proof of Proposition 5.3.6(a). For $0 \le d \le \underline{D}$, let $1^d \in \mathbb{T}$ denote the node $(1,\dots,1)$ with d entries (so $1^0 = \emptyset$ is the root of $\mathbb{T}$), and let S_d be the event that $H_N^{(\underline{u})}(\underline{\boldsymbol{\sigma}}(\underline{u})) \ge \mathsf{ALG} + \varepsilon$ for all $\underline{u} \in \mathbb{L}$ descended from the node 1^d. Let $P_d = \mathbb{P}[S_d]$. Note that $P_{\underline{D}} = \alpha_N$. We will show $P_0 \ge \alpha_N^{\frac{K}{N}} \ge \alpha_N^K$ by showing that for all $1 \le d \le \underline{D}$,

$$P_{d-1} \ge P_d^{k_d}.$$

The result will then follow by induction.

Recall the construction (5.3.2) of the Hamiltonians $H_N^{(u)}$ in terms of i.i.d. Hamiltonians $(\widetilde{H}_N^{[u]})_{u\in\mathbb{T}}$. Conditioned on the Hamiltonians $\Omega_{d-1} = (\widetilde{H}_N^{[1^{d'}]})_{0\le d'\le d-1}$, let $f_d(\Omega_{d-1})$ denote the conditional probability of S_d. Note that

$$P_d = \mathbb{E}\, f_d(\Omega_{d-1}).$$

By symmetry of the k_d descendant subtrees of the node 1^{d-1},

$$P_{d-1} = \mathbb{E}\, f_d(\Omega_{d-1})^{k_d}.$$

Thus $P_{d-1} \ge P_d^{k_d}$ by Jensen's inequality. □

Proof of Proposition 5.3.6(b). By definition of χ, $\mathbb{E}\,R(\underline{\boldsymbol{\sigma}}(\underline{u}^1),\underline{\boldsymbol{\sigma}}(\underline{u}^2)) = \chi(p_{\underline{u}^1\wedge\underline{u}^2})$. If $\underline{u}^1 \wedge \underline{u}^2 < \underline{D}$, then $p_{\underline{u}^1\wedge\underline{u}^2} < 1$. Because $\vec{p},\vec{q}$ are χ-aligned, we have $\chi(p_{\underline{u}^1\wedge\underline{u}^2}) = q_{\underline{u}^1\wedge\underline{u}^2}$. If $\underline{u}^1 \wedge \underline{u}^2 = \underline{D}$, then $p_{\underline{u}^1\wedge\underline{u}^2} = 1$, so clearly $\chi(p_{\underline{u}^1\wedge\underline{u}^2}) = \chi(1)$. So, in all cases, $\mathbb{E}\,R(\underline{\boldsymbol{\sigma}}(\underline{u}^1),\underline{\boldsymbol{\sigma}}(\underline{u}^2)) = \underline{Q}_{\underline{u}^1,\underline{u}^2}$.

Using (5.2.1) and a union bound over $\underline{u}^1,\underline{u}^2 \in \mathbb{L}$, we have

$$\left\|R(\vec{\underline{\boldsymbol{\sigma}}},\vec{\underline{\boldsymbol{\sigma}}}) - \underline{Q}\right\|_\infty \le \lambda$$

with probability at least $1 - K^2\nu$. By Lemma 5.3.7 and a union bound, we have

$$\left| R(\mathcal{A}(H_N^{(u)}), \boldsymbol{m}) - \|\boldsymbol{m}\|_N^2 \right| \le 2\lambda$$

for all $\underline{u} \in \underline{\mathbb{L}}$ with probability at least $1 - \frac{2K\nu}{\lambda}$. Recall that $\lambda = \eta_0/4 \le \eta/4$. By a final union bound,

$$\mathbb{P}[\underline{\vec{\boldsymbol{\sigma}}} \in \underline{\mathcal{Q}}(\eta/2)] \ge 1 - K^2\nu - \frac{2K\nu}{\lambda}.$$

□

Proof of Proposition 5.3.6(c). We focus on a fixed $\underline{u} \in \underline{\mathbb{L}}$. The requirements $H_N^{(u)} \in K_N$ follow from Proposition 5.2.3. The uniform eigenvalue lower bound follows by union bounding over subspaces S and a net of points $\boldsymbol{x}$. In fact it follows from exactly the same proof as [Sel21a, Lemma 2.6] up to replacing each appearance of an eigenvalue λ_i to λ_{K+i}.

□

Proof of Proposition 5.3.6(d). By (5.3.3) and Proposition 5.3.8 with $t = K\varepsilon/4$,

$$\begin{aligned}\mathbb{P}\left[\frac{1}{N}\max_{\vec{\boldsymbol{\sigma}}\in\mathcal{Q}(\eta)} \mathcal{H}_N(\vec{\boldsymbol{\sigma}}) \ge K(\mathsf{ALG} + 3\varepsilon/4)\right] &\le \mathbb{P}\left[\frac{1}{N}\max_{\vec{\boldsymbol{\sigma}}\in\mathcal{Q}(\eta)} \mathcal{H}_N(\vec{\boldsymbol{\sigma}}) - \frac{1}{N}\mathbb{E}\max_{\vec{\boldsymbol{\sigma}}\in\mathcal{Q}(\eta)} \mathcal{H}_N(\vec{\boldsymbol{\sigma}}) \ge \frac{K\varepsilon}{4}\right] \\ &\le 2\exp\left(-\frac{\varepsilon^2}{32\xi(1)}N\right).\end{aligned}$$

□

5.3.6 Proof of Lemma 5.3.4

The spherical case of Lemma 5.3.4 follows from [Sub21, Remark 6] and does not require any of the axis-aligned subspace conditions. We therefore focus on the Ising case, which is a slight extension of the main result of [Sel21a].

Lemma 5.3.9. *Suppose $S_{\text{eigen}}(\delta, K)$ holds. Then for any $\boldsymbol{x} \in [-1,1]^N$ with $||\boldsymbol{x}||_N^2 \le 1 - \delta$, any $u \in \mathbb{L}$ and any subspace $W \subseteq \mathbb{R}^N$ of dimension $\dim(W) \ge N - K - 1$, there are mutually orthogonal vectors $\boldsymbol{y}^1, \dots, \boldsymbol{y}^K \in W \cap \boldsymbol{x}^\perp$ such that for each $i \in [K]$ the following hold where C_3 is as in Proposition 5.2.3.*

1. *$\boldsymbol{x} + \boldsymbol{y}^i \in [-1,1]^N$.*
2. *If $x_j \in \{-1,1\}$ then $y_j^i = 0$.*
3. *$H_N^{(u)}(\boldsymbol{x} + \boldsymbol{y}^i) - H_N^{(u)}(\boldsymbol{x}) \ge -\delta\|\boldsymbol{y}^i\|_2^2$.*

4. $\left\|\boldsymbol{y}^i\right\|_N \le \frac{\delta}{10C_3}$.

5. *If* $\|\boldsymbol{x}\|_N^2 < q_d$ *for some* $1 \le d \le D$, *then* $\left\|\boldsymbol{x}+\boldsymbol{y}^i\right\|_N^2 \le q_d$.

6. *At least one of the following three events holds.*

   (a) $\left\|\boldsymbol{y}^i\right\|_N = \frac{\delta}{10C_3}$.

   (b) $\boldsymbol{x}+\boldsymbol{y}^i$ *has strictly more* ± 1*-valued coordinates than* $\boldsymbol{x}$.

   (c) $||\boldsymbol{x}||_N^2 < q_d$ *and* $||\boldsymbol{x}+\boldsymbol{y}^i||_N^2 = q_d$ *for some* $1 \le d \le D$.

Proof. By the Markov inequality, $\boldsymbol{x}$ has a set S of at least $(1-||\boldsymbol{x}||_N^2)N$ coordinates not in $\{-1,1\}$. $S_{\text{eigen}}(\delta, K)$ and the Cauchy interlacing inequality imply

$$\lambda_K\left(\nabla^2 H_N^{(u)}(\boldsymbol{x})|_{W_S\cap W}\right) \ge \lambda_{2K+1}\left(\nabla^2 H_N^{(u)}(\boldsymbol{x})|_{W_S}\right) \ge 0.$$

Let $\boldsymbol{y}^1,\dots,\boldsymbol{y}^K \in W_S(\boldsymbol{x})\cap W$ be a corresponding choice of orthogonal eigenvectors, each satisfying

$$\left\langle \boldsymbol{y}^i, \nabla^2 H_N^{(u)}(\boldsymbol{x})\boldsymbol{y}^i\right\rangle \ge 0.$$

Since $\boldsymbol{y}^i$ and $-\boldsymbol{y}^i$ play symmetric roles we may assume without loss of generality that $\langle \nabla H_N^{(u)}(\boldsymbol{y}), \boldsymbol{y}^i\rangle \ge 0$. Replacing $\boldsymbol{y}^i$ by $t\boldsymbol{y}^i$ for suitable $t \in [0,1]$ if needed, we may ensure that Items 1, 2, 4, 5, and 6 above hold.

Since $S_{\text{eigen}}(\delta, K)$ implies that $\left\|\nabla^3 H_N^{(u)}\right\|_{\text{op}}$ is uniformly bounded by C_3, it follows that along the line segment $\boldsymbol{x}+[0,1]\boldsymbol{y}^i$ the Hessian of $H_N^{(u)}$ varies in operator norm by at most $\frac{\delta}{5}$. This combined with $\langle \nabla H_N^{(u)}(\boldsymbol{x}), \boldsymbol{y}^i\rangle \ge 0$ implies

$$H_N^{(u)}(\boldsymbol{x}+\boldsymbol{y}^i) \ge H_N^{(u)}(\boldsymbol{x}) - \delta\left\|\boldsymbol{y}^i\right\|_2^2.$$

This completes the proof. □

Proof of Lemma 5.3.4. Take

$$\delta < \frac{\min(\varepsilon,\eta,1-q_{D-1})^2}{16(C_1+C_3+1)}$$

sufficiently small, where C_1, C_3 are given by Proposition 5.2.3. Enumerate $\underline{u}^1,\dots,\underline{u}^{\underline{K}} \in \underline{\mathbb{L}}$. Assume the points $\boldsymbol{\sigma}(u)$ for descendants $u \in \mathbb{L}$ of $\underline{u}^1,\dots,\underline{u}^{j-1}$ have already been chosen and satisfy the conclusions of Lemma 5.3.4. We show how to define the points $\boldsymbol{\sigma}(u)$ for u a descendant of $\underline{u}^j$.

From the starting point $\boldsymbol{x}^{0,\underline{u}^j} = \boldsymbol{\sigma}(\underline{u}^j)$, we produce iterates $\boldsymbol{x}^{i,v}$ for $i \in \mathbb{N}$ and $v \in \mathbb{T}$ a descendant of $\underline{u}^j$, similarly to [Sub21] and [Sel21a, Proof of Theorem 1]. First let $d_0 = d_0(\underline{u}^j) \in [D]$ be such that $||\boldsymbol{x}^{0,\underline{u}^j}||_N^2 \in [q_{d_0-1}, q_{d_0})$, and set $\boldsymbol{x}^{0,v} = \boldsymbol{x}^{0,\underline{u}^j}$ for all depth d_0 descendants v of $\underline{u}^j$ if $d_0 > \underline{D}$.

Given a point $\boldsymbol{x}^{m,v}$ with v a descendant of $\underline{u}^j$, suppose that $||\boldsymbol{x}^{m,v}||_N^2 \in (q_{|v|-1}, q_{|v|} \wedge (1-\delta))$. Then take the subspace $W^\perp$ (which changes from iteration to iteration) to be the span of $\boldsymbol{m}$ as well as all currently defined leaves of the exploration tree (including $\boldsymbol{x}^{m,v}$ itself). Hence $\dim(W^\perp) \le K+1$ and so $\dim(W) \ge N-K-1$. (The resulting exploration tree can be constructed in arbitrary order; at any time it will have at most K leaves.)

Then there exists $\boldsymbol{y}^{m,v}$ satisfying the properties of Lemma 5.3.9 with subspace W and Hamiltonian $H_N^{(\underline{u}^j)}$. We update

$$\boldsymbol{x}^{m+1,v} = \boldsymbol{x}^{m,v} + \boldsymbol{y}^{m,v}, \quad v \in \mathbb{T}.$$

However if $||\boldsymbol{x}^{m,v}||_N^2 = q_{|v|}$, then we let $v^1, \dots, v^{k_{d+1}}$ be the children of v in $\mathbb{T}$ and generate $\boldsymbol{y}^{m,v^1}, \dots \boldsymbol{y}^{m,v^{k_{d+1}}}$ again using Lemma 5.3.9. We then define

$$\boldsymbol{x}^{m+1,v^j} = \boldsymbol{x}^{m,v} + \boldsymbol{y}^{m,v^j}, \quad j \in [k_{d+1}].$$

Continuing in this way, we eventually reach points $\boldsymbol{x}^{m+1,u}$ with $||\boldsymbol{x}^{m+1,u}||_N^2 \ge (1-\delta)$ for each $u \in \mathbb{L}$; indeed the last condition of Lemma 5.3.9 ensures that this eventually occurs for each $u \in \mathbb{L}$. We set $\boldsymbol{x}^u = \boldsymbol{x}^{m+1,u}$. Observe that by orthogonality of $\boldsymbol{x}^{m,v}$ and $\boldsymbol{y}^{m,v}$,

$$\begin{aligned} H_N^{(u)}(\boldsymbol{x}^{m+1,v}) &\ge H_N^{(u)}(\boldsymbol{x}^{m,v}) - N\delta \|\boldsymbol{y}^{m,v}\|_N^2 \\ &\ge H_N^{(u)}(\boldsymbol{x}^{m,v}) - N\delta \cdot \left(\left\|\boldsymbol{x}^{m+1,v}\right\|_N^2 - \|\boldsymbol{x}^{m,v}\|_N^2 \right). \end{aligned}$$

It follows by telescoping that (recall $\underline{u}^j \in \underline{\mathbb{L}}$ is an ancestor of $u \in \mathbb{L}$),

$$H_N^{(u)}(\boldsymbol{x}^u) \ge H_N^{(u)}(\boldsymbol{x}^{\underline{u}^j}) - N\delta \ge H_N^{(u)}(\boldsymbol{x}^{\underline{u}^j}) - N\varepsilon/2.$$

Since every update above is made orthogonally to all contemporaneous iterates, it is not difficult to see that the final iterates $(\boldsymbol{x}^u)_{u \in \mathbb{L}}$ satisfy the following.

- $R(\boldsymbol{x}^u, \boldsymbol{x}^u) \ge 1 - \delta \ge q_{u \wedge u} - \frac{\eta}{2}$.
- If $u^1 \ne u^2$ are both descendants of $\underline{u}^j \in \underline{\mathbb{L}}$ and $u^1 \wedge u^2 < d_0(\underline{u}^j)$, then

  $$R(\boldsymbol{x}^{u^1}, \boldsymbol{x}^{u^2}) = R(\boldsymbol{x}^{\underline{u}^j}, \boldsymbol{x}^{\underline{u}^j}) \le q_{\underline{D}} + \frac{\eta}{2} \le q_{u^1 \wedge u^2} + \frac{\eta}{2}.$$

  and

  $$R(\boldsymbol{x}^{u^1}, \boldsymbol{x}^{u^2}) \ge q_{d_0 - 1} \ge q_{u^1 \wedge u^2},$$

  hence $\left| R(\boldsymbol{x}^{u^1}, \boldsymbol{x}^{u^2}) - q_{u^1 \wedge u^2} \right| \le \eta/2$.
- Otherwise, $R(\boldsymbol{x}^{u^1}, \boldsymbol{x}^{u^2}) = q_{u^1 \wedge u^2}$.

Moreover all updates were also orthogonal to $\boldsymbol{m}$, so $|R(\boldsymbol{m}, \boldsymbol{x}^u)| \le \eta/2$ for all $u \in \mathbb{L}$.

Finally, to produce outputs in Σ_N, for each $u \in \mathbb{L}$ and $i \in [N]$ we independently round the coordinate x_i^u at random to $\sigma(u)_i \in \{-1, 1\}$ so that $\mathbb{E}[\boldsymbol{\sigma}(u)] = \boldsymbol{x}^u$. It is not difficult to see that

$$\mathbb{P}[|R(\boldsymbol{x}^{u^1}, \boldsymbol{x}^{u^2}) - R(\boldsymbol{\sigma}(u^1), \boldsymbol{\sigma}(u^2))| \ge \delta] \le e^{-c(\delta)N}$$

for each $u^1, u^2 \in \mathbb{L}$, and similarly for inner products with $\boldsymbol{m}$. We conclude that $\vec{\boldsymbol{\sigma}} \in \mathcal{Q}(\eta)$ holds with probability $1 - e^{-c(\delta)N}$ (since $\delta \le \eta/2$). Similarly $\|\boldsymbol{\sigma}(u) - \boldsymbol{x}^u\|_2^2$ is an independent sum of N terms each at most 1 and has expectation at most δ. It follows that

$$\mathbb{P}[\|\boldsymbol{\sigma}(u) - \boldsymbol{x}^u\|_N \ge 2\delta^{1/2}] \le e^{-c(\delta)N}.$$

Now using S_{eigen}, for every $(\underline{u}, u) \in \underline{\mathbb{L}} \times \mathbb{L}$ with $\underline{u}$ an ancestor of u,

$$\begin{aligned} H_N^{(u)}(\boldsymbol{\sigma}(u)) &\ge H_N^{(u)}(\boldsymbol{x}^u) - 2C_1\delta^{1/2}N \\ &\ge H_N^{(u)}(\boldsymbol{x}^u) - N\varepsilon/2 \\ &\ge H_N^{(u)}(\boldsymbol{x}^{\underline{u}}) - N\varepsilon \end{aligned}$$

holds with probability $1 - e^{-c(\delta)N}$. In particular, the above events hold simultaneously over all $(\underline{u}, u)$ with probability at least $\frac{1}{2}$ over the random rounding step. Hence there exists some $\vec{\boldsymbol{\sigma}}$ satisfying all desired conditions. This concludes the proof. □

5.3.7 A Different Class of Algorithms Capturing The Approach of Subag

The optimization algorithm of [Sub21] in the spherical setting can be summarized as follows. Starting from any $\boldsymbol{x}^1 \in B_N$ with $\|\boldsymbol{x}^1\|_N^2 = \delta$, repeatedly compute the maximum-eigenvalue unit eigenvector $\boldsymbol{v}^i \in \mathbb{R}^N$ of $P_{(\boldsymbol{x}^i)^\perp} \nabla^2 H_N(\boldsymbol{x}^i) P_{(\boldsymbol{x}^i)^\perp}$ (the Hessian of H_N at $\boldsymbol{x}^i$ restricted to the orthogonal complement of $\boldsymbol{x}^i$). Then, set

$$\boldsymbol{x}^{i+1} = \boldsymbol{x}^i + \boldsymbol{v}^i\sqrt{\delta N} \tag{5.3.5}$$

where the sign of $\boldsymbol{v}^i$ is chosen depending on the gradient $\nabla H_N(\boldsymbol{x}^i)$. By construction, $\|\boldsymbol{x}^i\|_N^2 = i\delta$, so if $\delta^{-1} = m \in \mathbb{N}$ then $\boldsymbol{x}^m \in S_N$. By uniformly lower bounding the maximum eigenvalue of the Hessians, [Sub21] showed that this algorithm obtains energy at least $(\mathsf{ALG}^{\text{Sp}} + o_\delta(1))N$ as $\delta \to 0$. Because the maximum eigenvalue is a discontinuous operation, our results do not apply to Subag's algorithm.

We consider the following variant. At each $\boldsymbol{x}^i$, let the subspace $W(\boldsymbol{x}^i)$ be the span of the top $\lfloor \delta N \rfloor$ eigenvectors of $P_{(\boldsymbol{x}^i)^\perp} \nabla^2 H_N(\boldsymbol{x}^i) P_{(\boldsymbol{x}^i)^\perp}$. Next, choose $\boldsymbol{v}^i$ *uniformly at random* from the unit sphere of $W(\boldsymbol{x}^i)$ and update using (5.3.5). This modified algorithm obeys the same guarantees as

that of [Sub21] by exactly the same proof.

More generally, we define the class of δ-*subspace random walk* algorithms for $\delta > 0$ with $\delta^{-1} = m \in \mathbb{N}$, only in the spherical setting for convenience, as follows. Given H_N, let $W(\boldsymbol{x}^i) \subseteq \mathbb{R}^N$ be an arbitrary (measurable in $(H_N, \boldsymbol{x})$) subspace of dimension $\lfloor \delta N \rfloor$. Starting from arbitrary $\boldsymbol{x}^1 \in B_N$ with $\|\boldsymbol{x}_1\|_N^2 = \delta$, repeatedly choose a uniformly random unit vector $\boldsymbol{v}^i \in W(\boldsymbol{x}^i)$ and define $\boldsymbol{x}^{i+1}$ via (5.3.5), leading to the output $\boldsymbol{\sigma} = \boldsymbol{x}^m$. Note that unlike in the rest of this chapter, here the output $\boldsymbol{x}^{i+1}$ is random even given H_N, i.e. $\boldsymbol{x}^{i+1} = \mathcal{A}(H_N, \omega)$ for some independent random variable ω. As we now outline, for $\delta \le \delta_0(\varepsilon)$ sufficiently small depending on ε, no δ-subspace random walk algorithm can achieve energy than $\mathsf{ALG}^{\mathrm{Sp}} + \varepsilon$ with non-negligible probability.

Fixing H_N and $\boldsymbol{x}^1$, for any $j \le m$ we may generate coupled outputs $\boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2$ as follows. First use shared iterates $\boldsymbol{x}^{i,1} = \boldsymbol{x}^{i,2} = \boldsymbol{x}^i$ for $i \le j$ and then proceed via

$$\boldsymbol{x}^{i+1,\ell} = \boldsymbol{x}^{i,\ell} + \boldsymbol{v}^{i,\ell}\sqrt{\delta N}, \quad \ell \in \{1,2\}$$

for independent update sequences $(\boldsymbol{v}^{j,1}, \dots, \boldsymbol{v}^{m-1,1})$ and $(\boldsymbol{v}^{j,2}, \dots, \boldsymbol{v}^{m-1,2})$. Finally output $\boldsymbol{\sigma}^\ell = \boldsymbol{x}^{m,\ell}$. It is not difficult to see that for N sufficiently large,

$$\mathbb{P}\left[\left|R(\boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2) - j\delta\right| > \eta/2\right] \le e^{-cN}$$

for some $c = c(\delta, \eta)$ thanks to the random directions of the updates $\boldsymbol{v}^{i,\ell}$. With $\mathbb{L}$ as in the earlier part of this section, we can now construct a branching tree of outputs $\boldsymbol{\sigma}(u)$ for $u \in \mathbb{L}$. As $\delta \to 0$, for appropriate $j_d = \lfloor q_d \delta^{-1} \rfloor$, the solution configuration $\vec{\boldsymbol{\sigma}}$ hence constructed satisfies

$$\mathbb{P}[\vec{\boldsymbol{\sigma}} \in \mathcal{Q}(\eta)] \le e^{-cN}$$

with $\boldsymbol{m}$ the zero vector. Because we consider a single Hamiltonian H_N, we use Proposition 5.3.2 with $\chi(p) \to 0$ for all $p < 1$. Since the statement is uniform in χ, this does not present any difficulties (we are essentially "defining" $\vec{p} = 1^D$ to be χ-aligned with arbitrary $\vec{q}$). Mimicking the proofs earlier in this section (including the argument in the proof of Proposition 5.3.6(a) which now uses Jensen's inequality on the randomness of $\mathcal{A}$), we obtain the following result.

Theorem 26. *Consider a mixed even Hamiltonian H_N with model (ξ, h). For any $\varepsilon > 0$ there are $\delta_0, c, N_0 > 0$ depending only on ξ, h, ε such that the following holds for any $N \ge N_0$ and $\delta < \delta_0$. For any δ-subspace random walk algorithm $\mathcal{A}$,*

$$\mathbb{P}\left[\frac{1}{N} H_N\left(\mathcal{A}(H_N, \omega)\right) \ge \mathsf{ALG}^{\mathrm{Sp}} + \varepsilon\right] \le \exp(-cN).$$

5.4 Guerra's Interpolation

In this section, we begin the proof of Proposition 5.3.2. We take either $\mathcal{Q}(\eta) = \mathcal{Q}^{\mathrm{Sp}}(Q, \boldsymbol{m}, \eta)$ or $\mathcal{Q}(\eta) = \mathcal{Q}^{\mathrm{Is}}(Q, \boldsymbol{m}, \eta)$ (recall $Q = Q^{\vec{k},\vec{q}}$); the proofs in this section apply uniformly to both cases. The goal of this section is to use Guerra's interpolation to upper bound the constrained free energy

$$F_N(\mathcal{Q}(\eta)) = \frac{1}{N} \log \mathbb{E} \int_{\mathcal{Q}(\eta)} \exp \mathcal{H}_N(\vec{\boldsymbol{\sigma}}) \, \mathrm{d}\mu^K(\vec{\boldsymbol{\sigma}}),$$

where μ is a (for now) arbitrary measure on S_N. In the sequel, we will take μ to be the uniform measure on S_N for spherical spin glasses, and the counting measure on Σ_N for Ising spin glasses. We develop a bound on $F_N(\mathcal{Q}(\eta))$ that holds for all $D, \vec{k}, \vec{p}, \vec{q}, \eta$, and will set these variables in the sequel to prove Proposition 5.3.2.

We will control this free energy by controlling the following related free energy. Let $\lambda \in \mathbb{R}$ be a constant we will set later. For all $\boldsymbol{\sigma} \in \mathbb{R}^N$, let $\pi(\boldsymbol{\sigma}) = \boldsymbol{\sigma} - \boldsymbol{m}$. We define the following modified grand Hamiltonian, where we add an external field $\lambda \boldsymbol{m}$ centered at $\boldsymbol{m}$:

$$\begin{aligned}\mathcal{H}_{N,\lambda}(\vec{\boldsymbol{\sigma}}) &= \mathcal{H}_N(\vec{\boldsymbol{\sigma}}) + \sum_{u \in \mathbb{L}} \langle \lambda \boldsymbol{m}, \pi(\boldsymbol{\sigma}(u)) \rangle \\ &= K \langle \boldsymbol{h}, \boldsymbol{m} \rangle + \sum_{u \in \mathbb{L}} \left[\langle \boldsymbol{h} + \lambda \boldsymbol{m}, \pi(\boldsymbol{\sigma}(u)) \rangle + \widetilde{H}_N^{(u)}(\boldsymbol{\sigma}(u)) \right].\end{aligned}$$

We define the free energy

$$F_{N,\lambda}(\mathcal{Q}(\eta)) = \frac{1}{N} \log \mathbb{E} \int_{\mathcal{Q}(\eta)} \exp \mathcal{H}_{N,\lambda}(\vec{\boldsymbol{\sigma}}) \, \mathrm{d}\mu^K(\vec{\boldsymbol{\sigma}}).$$

Since $\mathcal{Q}(\eta) \subseteq B(\boldsymbol{m}, \eta)^K$, we have $|\mathcal{H}_N(\vec{\boldsymbol{\sigma}}) - \mathcal{H}_{N,\lambda}(\vec{\boldsymbol{\sigma}})| \le NK|\lambda|\eta$ for all $\vec{\boldsymbol{\sigma}} \in \mathcal{Q}(\eta)$, and so

$$|F_N(\mathcal{Q}(\eta)) - F_{N,\lambda}(\mathcal{Q}(\eta))| \le K|\lambda|\eta. \tag{5.4.1}$$

Define the matrices $M^{\vec{k},\vec{p},1}, \dots, M^{\vec{k},\vec{p},D} \in \mathbb{R}^{K \times K}$, whose rows and columns are indexed by $\mathbb{L}$, by

$$M^{\vec{k},\vec{p},d}_{u^1,u^2} = \mathbb{I}\{u^1 \wedge u^2 \ge d\} p_{u^1 \wedge u^2}.$$

Further, define $M^{\vec{k},\vec{p},\vec{q}} : [q_0, 1) \to \mathbb{R}^{K \times K}$ as the piecewise constant matrix-valued function such that for $q \in [q_{d-1}, q_d)$, $M^{\vec{k},\vec{p},\vec{q}}(q) = M^{\vec{k},\vec{p},d}$. Define $\kappa^{\vec{k},\vec{p},\vec{q}} : [q_0, 1) \to \mathbb{R}$ by

$$\kappa^{\vec{k},\vec{p},\vec{q}}(q) = \frac{1}{K} \mathrm{Sum}(M^{\vec{k},\vec{p},\vec{q}}(q)),$$

where Sum denotes the sum of entries of a matrix. Explicitly, for $q \in [q_{d-1}, q_d)$,

$$\kappa^{\vec{k},\vec{p},\vec{q}}(q) = \sum_{j=d}^{D-1} \left[(k_{j+1} - 1) \prod_{\ell=j+2}^{D} k_\ell \right] p_j + p_D. \tag{5.4.2}$$

When $\vec{k}, \vec{p}, \vec{q}$ are clear, we will write $M^d = M^{\vec{k},\vec{p},d}$, $M(q) = M^{\vec{k},\vec{p},\vec{q}}(q)$ and $\kappa(q) = \kappa^{\vec{k},\vec{p},\vec{q}}(q)$. Consider a sequence

$$0 = \zeta_{-1} < \zeta_0 < \cdots < \zeta_D = 1,$$

which we identify with the piecewise constant CDF $\zeta : [q_0, 1) \to [0, 1]$, where for $x \in [q_d, q_{d+1})$,

$$\zeta(x) = \zeta_d, \tag{5.4.3}$$

corresponding to the discrete distribution $\zeta(\{q_d\}) = \zeta_d - \zeta_{d-1}$. We denote by $\mathcal{M}_{\vec{q}}$ the set of such CDFs ζ for a given $\vec{q}$.

Let $\mathbb{T}_D = \mathbb{N}^0 \cup \mathbb{N}^1 \cup \cdots \cup \mathbb{N}^D$ and for $\omega \in \mathbb{T}_D$, let $|\omega|$ denote the length of ω. Let $\emptyset$ denote the empty tuple. We think of $\mathbb{T}_D$ as a tree rooted at $\emptyset$, where the parent of any $\omega \neq \emptyset$ is the initial substring of ω with length $|\omega| - 1$. For $\alpha \in \mathbb{N}^D$, let $p(\alpha) = ((\alpha_1), (\alpha_1, \alpha_2), \ldots, (\alpha_1, \ldots, \alpha_D))$ denote the path of vertices from the root to α, not including the root. For $\alpha^1, \alpha^2 \in \mathbb{N}^D$, let $\alpha^1 \wedge \alpha^2$ denote the depth of the least common ancestor of α^1 and α^2. Recall the Ruelle cascades $(\nu_\alpha)_{\alpha \in \mathbb{N}^D}$ corresponding to $(\zeta_0, \zeta_1, \ldots, \zeta_{D-1})$ which were introduced in [Rue87], see also [Pan13b, Section 2.3].

For each increasing $\psi : [q_0, 1] \to \mathbb{R}_{\geq 0}$, we define a Gaussian process $g_\psi^{(u)}(\alpha)$ indexed by $(u, \alpha) \in \mathbb{L} \times \mathbb{N}^D$ as follows. Generate $\vec{\eta}_\emptyset \in \mathbb{R}^K$ by

$$\vec{\eta}_\emptyset = (\eta_\emptyset(u))_{u \in \mathbb{L}} \sim \mathcal{N}(0, M^1).$$

Furthermore, for each non-root $\omega \in \mathbb{T}_D$, independently generate $\vec{\eta}_\omega \in \mathbb{R}^K$ by

$$\vec{\eta}_\omega = (\eta_\omega(u))_{u \in \mathbb{L}} \sim \mathcal{N}(0, M^{|\omega|}).$$

Then, for each $u \in \mathbb{L}$, set

$$g_\psi^{(u)}(\alpha) = \eta_\emptyset(u)\psi(q_0)^{1/2} + \sum_{\omega \in p(\alpha)} \eta_\omega(u)(\psi(q_{|\omega|}) - \psi(q_{|\omega|-1}))^{1/2}.$$

This is the centered Gaussian process with covariance

$$\mathbb{E}\, g_\psi^{(u^1)}(\alpha^1) g_\psi^{(u^2)}(\alpha^2) = p_{u^1 \wedge u^2} \psi(q_{\alpha^1 \wedge \alpha^2} \wedge q_{u^1 \wedge u^2}),$$

where for $x, y \in \mathbb{R}$, $x \wedge y = \min(x, y)$. Generate N i.i.d. copies of the process $g_{\xi'}^{(u)}(\alpha)$, which we

denote $g_{\xi',i}^{(u)}(\alpha)$ for $i = 1, \dots, N$. Similarly, for the function

$$\theta(q) = (q - q_0)\xi'(q) - \xi(q) + \xi(q_0),$$

we generate N i.i.d. processes $g_{\theta,i}^{(u)}(\alpha)$ for $i = 1, \dots, N$. Note that for $q \in [q_0, 1)$,

$$\theta(q) = \int_{q_0}^{q} (\xi'(q) - \xi'(q'))\, \mathrm{d}q' \geq 0 \qquad \text{and} \qquad \theta'(q) = (q - q_0)\xi''(q) \geq 0,$$

so θ is nonnegative and increasing, as required. For $t \in [0, 1]$, define the interpolating Hamiltonian

$$\begin{aligned}\mathcal{H}_{N,\lambda,t}(\vec{\boldsymbol{\sigma}}, \alpha) = \sum_{u \in \mathbb{L}} \left[\sqrt{t} \widetilde{H}_N^{(u)}(\boldsymbol{\sigma}(u)) + \sqrt{1-t} \sum_{i=1}^{N} g_{\xi',i}^{(u)}(\alpha) \pi(\boldsymbol{\sigma}(u))_i + \sqrt{t} \sum_{i=1}^{N} g_{\theta,i}^{(u)}(\alpha) \right] \\ + K \langle \boldsymbol{h}, \boldsymbol{m} \rangle + \langle \boldsymbol{h} + \lambda \boldsymbol{m}, \pi(\boldsymbol{\sigma}(u)) \rangle \end{aligned} \tag{5.4.4}$$

and the interpolating free energy

$$\varphi(t) = \frac{1}{N} \mathbb{E} \log \sum_{\alpha \in \mathbb{N}^D} \nu_\alpha \int_{\mathcal{Q}(\eta)} \exp \mathcal{H}_{N,\lambda,t}(\vec{\boldsymbol{\sigma}}, \alpha)\, \mathrm{d}\mu^K(\vec{\boldsymbol{\sigma}}).$$

The following bound on $F_N(\mathcal{Q}(\eta))$ is the main result of this section.

Proposition 5.4.1. *The free energy $F_N(\mathcal{Q}(\eta))$ is upper bounded by*

$$F_N(\mathcal{Q}(\eta)) \leq \varphi(0) - \frac{K}{2} \int_{q_0}^{1} (q - q_0)\xi''(q)\kappa(q)\zeta(q)\, \mathrm{d}q + 3K^2\xi''(1)\eta + K|\lambda|\eta.$$

where $\zeta : [q_0, 1) \to [0, 1]$ is defined in (5.4.3).

Lemma 5.4.2 (Guerra's interpolation bound)**.** *For all $t \in [0, 1]$ and $\eta \in (0, 1)$,*

$$\varphi'(t) \leq 3K^2\xi''(1)\eta.$$

Proof. Let $\langle \cdot \rangle_t$ denote the average with respect to the Gibbs measure on $\mathcal{Q}(\eta) \times \mathbb{N}^D$ given by

$$G(\vec{\boldsymbol{\sigma}}, \alpha) \propto \nu_\alpha \exp \mathcal{H}_{N,\lambda,t}(\vec{\boldsymbol{\sigma}}, \alpha).$$

By Gaussian integration by parts [Pan13b, Lemma 1.4],

$$\begin{aligned}\varphi'(t) &= \frac{1}{N} \mathbb{E} \left\langle \frac{\partial \mathcal{H}_{N,\lambda,t}}{\partial t}(\vec{\boldsymbol{\sigma}}, \alpha) \right\rangle_t \\ &= \frac{1}{N} \mathbb{E} \left\langle \mathbb{E} \frac{\partial \mathcal{H}_{N,\lambda,t}}{\partial t}(\vec{\boldsymbol{\sigma}}^1, \alpha^1) \mathcal{H}_{N,\lambda,t}(\vec{\boldsymbol{\sigma}}^1, \alpha^1) - \mathbb{E} \frac{\partial \mathcal{H}_{N,\lambda,t}}{\partial t}(\vec{\boldsymbol{\sigma}}^1, \alpha^1) \mathcal{H}_{N,\lambda,t}(\vec{\boldsymbol{\sigma}}^2, \alpha^2) \right\rangle_t, \end{aligned} \tag{5.4.5}$$

where $(\vec{\boldsymbol{\sigma}}^1, \alpha^1)$ and $(\vec{\boldsymbol{\sigma}}^2, \alpha^2)$ are independent samples from the Gibbs measure. Recall (5.4.4). For any realizations $(\vec{\boldsymbol{\sigma}}^1, \alpha^1)$ and $(\vec{\boldsymbol{\sigma}}^2, \alpha^2)$,

$$
\begin{aligned}
&\frac{2}{N}\,\mathbb{E}\,\frac{\partial \mathcal{H}_{N,\lambda,t}}{\partial t}(\vec{\boldsymbol{\sigma}}^1, \alpha^1)\mathcal{H}_{N,t}(\vec{\boldsymbol{\sigma}}^2, \alpha^2) \\
&= \sum_{u^1,u^2 \in \mathbb{L}} p_{u^1 \wedge u^2} \left[\xi(R(\boldsymbol{\sigma}^1(u^1), \boldsymbol{\sigma}^2(u^2))) - R(\pi(\boldsymbol{\sigma}^1(u^1)), \pi(\boldsymbol{\sigma}^2(u^2)))\xi'(q_{\alpha^1 \wedge \alpha^2} \wedge q_{u^1 \wedge u^2}) + \theta(q_{\alpha^1 \wedge \alpha^2} \wedge q_{u^1 \wedge u^2})\right] \\
&= \sum_{u^1,u^2 \in \mathbb{L}} p_{u^1 \wedge u^2} \big[\xi(R(\boldsymbol{\sigma}^1(u^1), \boldsymbol{\sigma}^2(u^2))) \\
&\qquad - \big(R(\boldsymbol{\sigma}^1(u^1), \boldsymbol{\sigma}^2(u^2)) - R(\boldsymbol{\sigma}^1(u^1), \boldsymbol{m}) - R(\boldsymbol{\sigma}^2(u^2), \boldsymbol{m}) + R(\boldsymbol{m}, \boldsymbol{m})\big)\,\xi'(q_{\alpha^1 \wedge \alpha^2} \wedge q_{u^1 \wedge u^2}) \\
&\qquad + \theta(q_{\alpha^1 \wedge \alpha^2} \wedge q_{u^1 \wedge u^2})\big] \\
&= \sum_{u^1,u^2 \in \mathbb{L}} p_{u^1 \wedge u^2} \big[C\left(R(\boldsymbol{\sigma}^1(u^1), \boldsymbol{\sigma}^2(u^2)), q_{\alpha^1 \wedge \alpha^2} \wedge q_{u^1 \wedge u^2}\right) \\
&\qquad + \big(R(\boldsymbol{\sigma}^1(u^1), \boldsymbol{m}) + R(\boldsymbol{\sigma}^2(u^2), \boldsymbol{m}) - 2q_0\big)\,\xi'(q_{\alpha^1 \wedge \alpha^2} \wedge q_{u^1 \wedge u^2}) + \xi(q_0)\big],
\end{aligned}
$$

where

$$
C(x, y) = \xi(x) - \xi(y) - (x - y)\xi'(y) = \int_y^x \int_y^z \xi''(w)\ \mathrm{d}w\ \mathrm{d}z. \tag{5.4.6}
$$

Because $\boldsymbol{\sigma}^1(u^1), \boldsymbol{\sigma}^2(u^2) \in B(\boldsymbol{m}, \eta)$,

$$
\left|\big(R(\boldsymbol{\sigma}^1(u^1), \boldsymbol{m}) + R(\boldsymbol{\sigma}^2(u^2), \boldsymbol{m}) - 2q_0\big)\,\xi'(q_{\alpha^1 \wedge \alpha^2} \wedge q_{u^1 \wedge u^2})\right| \le 2\xi'(1)\eta.
$$

Hence using (5.4.5) and noting that $q_{\alpha^1 \wedge \alpha^1} = 1$, we obtain

$$
\begin{aligned}
\varphi'(t) \le \frac{1}{2} \sup_{\substack{\vec{\boldsymbol{\sigma}}^1, \vec{\boldsymbol{\sigma}}^2 \in \mathcal{Q}(\eta) \\ \alpha^1, \alpha^2 \in \mathbb{N}^D}} \sum_{u^1,u^2 \in \mathbb{L}} &\Big[C\left(R(\boldsymbol{\sigma}^1(u^1), \boldsymbol{\sigma}^1(u^2)), q_{u^1 \wedge u^2}\right) - C\left(R(\boldsymbol{\sigma}^1(u^1), \boldsymbol{\sigma}^2(u^2)), q_{\alpha^1 \wedge \alpha^2} \wedge q_{u^1 \wedge u^2}\right)\Big] \\
&+ 2K^2\xi'(1)\eta.
\end{aligned}
$$

By (5.4.6), $0 \le C(x, y) \le |x - y|^2\xi''(1)$. Since $|R(\boldsymbol{\sigma}^1(u^1), \boldsymbol{\sigma}^1(u^2)) - q_{u^1 \wedge u^2}| \le \eta$ for $\vec{\boldsymbol{\sigma}}^1 \in \mathcal{Q}(\eta)$,

$$
C\left(R(\boldsymbol{\sigma}^1(u^1), \boldsymbol{\sigma}^1(u^2)), q_{u^1 \wedge u^2}\right) \le \xi''(1)\eta^2.
$$

Moreover,

$$
C\left(R(\boldsymbol{\sigma}^1(u^1), \boldsymbol{\sigma}^2(u^2)), q_{\alpha^1 \wedge \alpha^2} \wedge q_{u^1 \wedge u^2}\right) \ge 0.
$$

So,

$$
\varphi'(t) \le \frac{1}{2}K^2\xi''(1)\eta^2 + 2K^2\xi'(1)\eta \le 3K^2\xi''(1)\eta.
$$

□

We will now evaluate $\varphi(1)$ to complete the proof of Proposition 5.4.1.

Lemma 5.4.3. *The following identity holds.*

$$\varphi(1) = F_{N,\lambda}(\mathcal{Q}(\eta)) + \frac{K}{2}\sum_{d=0}^{D-1}\kappa(q_d)\zeta_d(\theta(q_{d+1}) - \theta(q_d)).$$

Proof. It is clear that

$$\varphi(1) = F_{N,\lambda}(\mathcal{Q}(\eta)) + \frac{1}{N}\log\mathbb{E}\sum_{\alpha\in\mathbb{N}^D}\nu_\alpha \exp\sum_{u\in\mathbb{L}}\sum_{i=1}^{N} g_{\theta,i}^{(u)}(\alpha).$$

We will evaluate the last term by the recursive evaluation of Ruelle cascades. For $1 \le d \le D$, independently generate $\vec{\boldsymbol{\eta}}_d = (\boldsymbol{\eta}_d(u))_{u\in\mathbb{L}} \in (\mathbb{R}^N)^K$ by generating, independently for each $1 \le i \le N$,

$$(\vec{\boldsymbol{\eta}}_d)_i = (\boldsymbol{\eta}_d(u)_i)_{u\in\mathbb{L}} \sim \mathcal{N}(0, M^d).$$

(Because $\theta(q_0) = 0$, we will not need $\vec{\boldsymbol{\eta}}_0$, corresponding to the root $\emptyset$ of $\mathbb{T}_D$.) Let

$$X_D = \sum_{u\in\mathbb{L}}\sum_{i=1}^{N}\sum_{d=1}^{D}\boldsymbol{\eta}_d(u)_i\left(\theta(q_d) - \theta(q_{d-1})\right)^{1/2},$$

and for $0 \le d \le D-1$ let

$$X_d = \frac{1}{\zeta_d}\log\mathbb{E}_d \exp\zeta_d X_{d+1}, \tag{5.4.7}$$

where $\mathbb{E}_d$ denotes expectation with respect to $\vec{\boldsymbol{\eta}}_{d+1}$. By properties of Ruelle cascades [Pan13b, Theorem 2.9],

$$\frac{1}{N}\log\mathbb{E}\sum_{\alpha\in\mathbb{N}^D}\nu_\alpha\exp\sum_{u\in\mathbb{L}}\sum_{i=1}^{N} g_{\theta,i}^{(u)}(\alpha) = \frac{1}{N}X_0.$$

Here we use that the depth-zero term $\eta_\emptyset(u)\theta(q_0)^{1/2}$ of $g_\theta^{(u)}(\alpha)$ is zero because $\theta(q_0) = 0$. We now evaluate X_0 by (5.4.7). For each $1 \le d \le D$, $\sum_{u\in\mathbb{L}}\sum_{i=1}^{N}\boldsymbol{\eta}_d(u)_i$ has variance

$$\mathbb{E}\left(\sum_{u\in\mathbb{L}}\sum_{i=1}^{N}\boldsymbol{\eta}_d(u)_i\right)^2 = N\mathrm{Sum}(M^d) = NK\kappa(q_{d-1}).$$

So,

$$\begin{aligned}\frac{1}{\zeta_d}\log\mathbb{E}_d\exp\zeta_d\left(\sum_{u\in\mathbb{L}}\sum_{i=1}^{N}\boldsymbol{\eta}_{d+1}(u)_i\right)\left(\theta(q_{d+1}) - \theta(q_d)\right)^{1/2} &= \frac{1}{\zeta_d}\log\exp\left(\frac{NK}{2}\kappa(q_d)\zeta_d^2(\theta(q_{d+1}) - \theta(q_d))\right)\\ &= \frac{NK}{2}\kappa(q_d)\zeta_d(\theta(q_{d+1}) - \theta(q_d)).\end{aligned}$$

A straightforward induction argument using this computation gives

$$\frac{1}{N}X_0 = \frac{K}{2}\sum_{d=0}^{D-1}\kappa(q_d)\zeta_d(\theta(q_{d+1}) - \theta(q_d)),$$

completing the proof. □

Corollary 5.4.4. *For the distribution function $\zeta : [q_0, 1) \to [0, 1]$ defined in* (5.4.3)*,*

$$\varphi(1) = F_{N,\lambda}(\mathcal{Q}(\eta)) + \frac{K}{2}\int_{q_0}^{1}(q - q_0)\xi''(q)\kappa(q)\zeta(q)\ \mathrm{d}q$$

Proof. On each interval $[q_d, q_{d+1})$, the functions $\kappa(q)$ and $\zeta(q)$ are constant. Moreover, recall that $\theta'(q) = (q - q_0)\xi''(q)$. The result follows from Lemma 5.4.3. □

Proof of Proposition 5.4.1. By Lemma 5.4.2 and Corollary 5.4.4,

$$F_{N,\lambda}(\mathcal{Q}(\eta)) \le \varphi(0) - \frac{K}{2}\int_{q_0}^{1}(q - q_0)\xi''(q)\kappa(q)\zeta(q)\ \mathrm{d}q + 3K^2\xi''(1)\eta.$$

The result follows from (5.4.1). □

In the following two sections, we will use Proposition 5.4.1 to upper bound $F_N(\mathcal{Q}(\eta))$ in the spherical and Ising settings by estimating

$$\varphi(0) = KR(\boldsymbol{h}, \boldsymbol{m}) + \frac{1}{N}\log\mathbb{E}\sum_{\alpha\in\mathbb{N}^D}\nu_\alpha\int_{\mathcal{Q}(\eta)}\exp\sum_{u\in\mathbb{L}}\left[\langle\boldsymbol{h} + \lambda\boldsymbol{m}, \pi(\boldsymbol{\sigma}(u))\rangle + \sum_{i=1}^{N}g_{\xi,i}^{(u)}(\alpha)\pi(\boldsymbol{\sigma}(u))_i\right]\mathrm{d}\mu^K(\vec{\boldsymbol{\sigma}}). \tag{5.4.8}$$

In the spherical and Ising settings, μ is respectively the uniform measure on S_N and the counting measure on Σ_N. We denote $\varphi(0)$ in these settings by $\varphi^{\mathrm{Sp}}(0)$ and $\varphi^{\mathrm{Is}}(0)$. We will also denote F_N in these settings by F_N^{Sp} and F_N^{Is}.

5.5 Overlap-Constrained Upper Bound on the Spherical Grand Hamiltonian

In this section, we complete the proof of Proposition 5.3.2 in the spherical setting. Denote the expected overlap-constrained maximum energy of the grand Hamiltonian by

$$\mathsf{OPT}_N^{\mathrm{Sp}}(\mathcal{Q}(\eta)) = \frac{1}{N}\,\mathbb{E}\max_{\vec{\boldsymbol{\sigma}}\in\mathcal{Q}(\eta)}\mathcal{H}_N(\vec{\boldsymbol{\sigma}}).$$

Let $\underline{\mathscr{L}}$ and $\overline{\mathscr{L}}$ denote the subsets of $\mathscr{L}$ supported on $[0, q_0)$ and $[q_0, 1)$, respectively. The function κ defined in (5.4.2) is an element of $\overline{\mathscr{L}}$. Moreover (recall (5.4.3)) $\mathcal{M}_{\vec{q}} \subseteq \overline{\mathscr{L}}$. For $\beta > 0$ and $\zeta \in \mathcal{M}_{\vec{q}}$, let $\beta\kappa\zeta \in \overline{\mathscr{L}}$ denote the pointwise product $\beta\kappa\zeta(q) = \beta\kappa(q)\zeta(q)$. For any $\underline{\zeta} \in \underline{\mathscr{L}}$, let $\underline{\zeta} + \beta\kappa\zeta \in \mathscr{L}$ be the function

$$(\underline{\zeta} + \beta\kappa\zeta)(q) = \begin{cases} \underline{\zeta}(q) & q < q_0, \\ \beta\kappa\zeta(q) & q \geq q_0. \end{cases}$$

We will develop the following bound on $\mathsf{OPT}_N^{\mathrm{Sp}}(\mathcal{Q}(\eta))$ for all $D, \vec{k}, \vec{p}, \vec{q}, \eta, \beta$.

Proposition 5.5.1. *Let $\zeta \in \mathcal{M}_{\vec{q}}$ and $\underline{\zeta} \in \underline{\mathscr{L}}$ be arbitrary. Let $\beta > 0$ and suppose that $(B, \underline{\zeta} + \beta\kappa\zeta) \in \mathscr{K}(\xi)$, $B \geq \beta^{-1}$. There exists a constant C, depending only on ξ, h, such that for $N \geq C \log \max(K, 2)$,*

$$\mathsf{OPT}_N^{\mathrm{Sp}}(\mathcal{Q}(\eta)) \leq K\mathsf{P}^{\mathrm{Sp}}(B, \underline{\zeta} + \beta\kappa\zeta) + CK^2 \left(\beta\eta + B\eta + \frac{\log \frac{1}{\eta}}{\beta} + \frac{1}{\sqrt{N}} \right).$$

Crucially, in the input of the Parisi functional, the increasing function ζ is pointwise multiplied by κ, which (by selecting appropriate parameters $\vec{k}, \vec{p}, \vec{q}$) can be arranged to decrease as rapidly as desired. This multiplication by κ allows us to pass from increasing functions $\zeta \in \mathcal{M}_{\vec{q}}$ to arbitrary bounded variation functions, in the sense that $\beta\kappa\zeta$ can approximate any element of $\overline{\mathscr{L}}$. Consequently, $\underline{\zeta} + \beta\kappa\zeta$ can approximate any element of $\mathscr{L}$, and $\mathsf{P}^{\mathrm{Sp}}(B, \underline{\zeta} + \beta\kappa\zeta)$ can be made arbitrarily close to $\mathsf{ALG}^{\mathrm{Sp}}$. We will prove Proposition 5.3.2 by setting the parameters in Proposition 5.5.1 such that $(B, \underline{\zeta} + \beta\kappa\zeta)$ approximates the minimizer of P^{Sp} and the error term is small.

Our proof of Proposition 5.3.2 proceeds in three steps. In Subsection 5.5.1 we use the machinery of the previous section to prove Proposition 5.5.2, an upper bound on the free energy $F_N^{\mathrm{Sp}}(\mathcal{Q}(\eta))$. In Subsection 5.5.2, we take this bound to low temperature to prove Proposition 5.5.1. In Subsection 5.5.3, we complete the proof of Proposition 5.3.2 by setting appropriate parameters in Proposition 5.5.1.

5.5.1 The Free Energy Upper Bound

In this subsection, we will use Proposition 5.4.1 to upper bound $F_N^{\mathrm{Sp}}(\mathcal{Q}(\eta))$. We take μ to be the uniform measure on S_N. The main result of this subsection is the following upper bound on $F_N^{\mathrm{Sp}}(\mathcal{Q}(\eta))$, which holds for all $D, \vec{k}, \vec{p}, \vec{q}, \eta$.

Proposition 5.5.2. *Let $\zeta \in \mathcal{M}_{\vec{q}}$ and $\underline{\zeta} \in \underline{\mathscr{L}}$ be arbitrary. Suppose $(B, \underline{\zeta} + \kappa\zeta) \in \mathscr{K}(\xi)$, $B \geq 1$, and $N \geq 2$. Then,*

$$F_N^{\mathrm{Sp}}(\mathcal{Q}(\eta)) \leq K\mathsf{P}^{\mathrm{Sp}}(B, \underline{\zeta} + \kappa\zeta) + 3K^2\xi''(1)\eta + KB\eta.$$

The crux of this argument is to upper bound $\varphi^{\mathrm{Sp}}(0)$ so that we may apply Proposition 5.4.1. We

equip the state space $(\mathbb{R}^N)^K$ with the natural inner product

$$\langle \vec{\boldsymbol{y}}^1, \vec{\boldsymbol{y}}^2 \rangle = \sum_{u \in \mathbb{L}} \langle \boldsymbol{y}^1(u), \boldsymbol{y}^2(u) \rangle$$

and norm $\|\vec{\boldsymbol{y}}\|^2 = \langle \vec{\boldsymbol{y}}, \vec{\boldsymbol{y}} \rangle$. Generate $\vec{\boldsymbol{\eta}}_0 = (\boldsymbol{\eta}_0(u)) \in (\mathbb{R}^N)^K$ by generating, independently for each $1 \le i \le N$,

$$(\vec{\boldsymbol{\eta}}_0)_i = (\boldsymbol{\eta}_0(u)_i)_{u \in \mathbb{L}} \sim \mathcal{N}(0, M^1). \tag{5.5.1}$$

Similarly, for $1 \le d \le D$, independently generate $\vec{\boldsymbol{\eta}}_d = (\boldsymbol{\eta}_d(u))_{u \in \mathbb{L}} \in (\mathbb{R}^N)^K$ by generating, independently for each $1 \le i \le N$,

$$(\vec{\boldsymbol{\eta}}_d)_i = (\boldsymbol{\eta}_d(u)_i)_{u \in \mathbb{L}} \sim \mathcal{N}(0, M^d). \tag{5.5.2}$$

Let $\vec{\boldsymbol{m}} = (\boldsymbol{m}(u))_{u \in \mathbb{L}} \in (\mathbb{R}^N)^K$ and $\vec{\boldsymbol{h}} = (\boldsymbol{h}(u))_{u \in \mathbb{L}} \in (\mathbb{R}^N)^K$ satisfy $\boldsymbol{m}(u) = \boldsymbol{m}$ and $\boldsymbol{h}(u) = \boldsymbol{h}$ for all $u \in \mathbb{L}$. For $\vec{\boldsymbol{\sigma}} \in (\mathbb{R}^N)^K$, define $\pi(\vec{\boldsymbol{\sigma}}) = \vec{\boldsymbol{\sigma}} - \vec{\boldsymbol{m}}$. We define the following functions on $(\mathbb{R}^N)^K$. Let

$$\begin{aligned} G_D(\vec{\boldsymbol{y}}) &= \log \int_{\mathcal{Q}(\eta)} \exp \langle \vec{\boldsymbol{y}}, \pi(\vec{\boldsymbol{\sigma}}) \rangle \, \mathrm{d}\mu^K(\vec{\boldsymbol{\sigma}}) \\ &= -\langle \vec{\boldsymbol{y}}, \vec{\boldsymbol{m}} \rangle + \log \int_{\mathcal{Q}(\eta)} \exp \langle \vec{\boldsymbol{y}}, \vec{\boldsymbol{\sigma}} \rangle \, \mathrm{d}\mu^K(\vec{\boldsymbol{\sigma}}). \end{aligned}$$

and for $0 \le d \le D-1$, let

$$G_d(\vec{\boldsymbol{y}}) = \frac{1}{\zeta_d} \log \mathbb{E} \exp \zeta_d G_{d+1} \left(\vec{\boldsymbol{y}} + \vec{\boldsymbol{\eta}}_{d+1} (\xi'(q_{d+1}) - \xi'(q_d))^{1/2} \right).$$

By properties of Ruelle cascades,

$$\varphi^{\mathrm{Sp}}(0) = \frac{1}{N} \mathbb{E} \, G_0((\vec{\boldsymbol{h}} + \lambda \vec{\boldsymbol{m}}) + \vec{\boldsymbol{\eta}}_0 \xi'(q_0)^{1/2}) + KR(\boldsymbol{h}, \boldsymbol{m}).$$

We will estimate the spherical integral G_D, and through it the functions G_d for $0 \le d \le D-1$, by comparison with a Gaussian integral. This step relies on the following lemma, which is a straightforward extension of [Tal06a, Lemma 3.1]; we defer the proof to the end of this section. For $B \ge 1$, let ν_B denote the measure of $\mathcal{N}(0, \frac{1}{B})$. Let $\chi^2(d)$ denote a χ^2 random variable with d degrees of freedom.

Lemma 5.5.3. *For all $\vec{\boldsymbol{y}} \in (\mathbb{R}^N)^K$,*

$$\exp G_D(\vec{\boldsymbol{y}}) \le \mathbb{P} \left(\chi^2(N) \ge BN \right)^{-K} \exp \left(-\langle \vec{\boldsymbol{y}}, \vec{\boldsymbol{m}} \rangle \right) \int \exp \langle \vec{\boldsymbol{y}}, \vec{\boldsymbol{\rho}} \rangle \, \mathrm{d}\nu_B^{N \times K}(\vec{\boldsymbol{\rho}}).$$

The probability term in this lemma can be controlled by the following standard bound, whose proof we also defer.

Lemma 5.5.4. *If $B \geq 1$ and $N \geq 2$, then*

$$\mathbb{P}(\chi^2(N) \geq BN) \geq \exp(-BN/2).$$

It remains to analyze the terms in Lemma 5.5.3 involving $\vec{\boldsymbol{y}}$. Define further

$$G'_D(\vec{\boldsymbol{y}}) = -\langle \vec{\boldsymbol{y}}, \vec{\boldsymbol{m}} \rangle + \log \int \exp\langle \vec{\boldsymbol{y}}, \vec{\boldsymbol{\rho}} \rangle \, \mathrm{d}\nu_b^K(\vec{\boldsymbol{\rho}}) = \frac{\|\vec{\boldsymbol{y}}\|_2^2}{2B} - \langle \vec{\boldsymbol{y}}, \vec{\boldsymbol{m}} \rangle,$$
$$G'_d(\vec{\boldsymbol{y}}) = \frac{1}{\zeta_d} \log \mathbb{E} \exp \zeta_d G_{d+1} \left(\vec{\boldsymbol{y}} + \vec{\boldsymbol{\eta}}_{d+1} (\xi'(q_{d+1}) - \xi'(q_d))^{1/2} \right) \qquad \text{for } 0 \leq d \leq D-1,$$

Henceforth, suppose $N \geq 2$. Lemmas 5.5.3 and 5.5.4 imply that

$$\varphi^{\mathrm{Sp}}(0) \leq \frac{1}{N} \mathbb{E} \, G'_0((\vec{\boldsymbol{h}} + \lambda \vec{\boldsymbol{m}}) + \vec{\boldsymbol{\eta}}_0 \xi'(q_0)^{1/2}) + KR(\boldsymbol{h}, \boldsymbol{m}) + \frac{1}{2} KB. \tag{5.5.3}$$

Consider a new state space $\mathbb{R}^K$ with elements $\vec{y} = (y(u))_{u \in \mathbb{L}}$ where $y(u) \in \mathbb{R}$, equipped with the natural inner product

$$\langle \vec{y}^1, \vec{y}^2 \rangle = \sum_{u \in \mathbb{L}} \vec{y}^1(u) \vec{y}^2(u)$$

and norm $\|\vec{y}\|_2^2 = \langle \vec{y}, \vec{y} \rangle$. Generate the $\mathbb{R}^K$-valued Gaussians $\vec{\eta}_0 \sim \mathcal{N}(0, M^1)$ and, for $1 \leq d \leq D$, $\vec{\eta}_d \sim \mathcal{N}(0, M^d)$. Recall that $\boldsymbol{h} = (h, \ldots, h)$. Let $\boldsymbol{m} = (m_1, \ldots, m_N)$, and let $\vec{1} \in \mathbb{R}^K$ denote the all-1 vector. For $1 \leq i \leq N$, define the following functions on $\mathbb{R}^K$.

$$\Gamma^i_D(\vec{y}) = \frac{\|\vec{y}\|^2}{2B} - m_i \langle \vec{1}, \vec{y} \rangle,$$
$$\Gamma^i_d(\vec{y}) = \frac{1}{\zeta_d} \log \mathbb{E} \exp \zeta_d \Gamma^i_{d+1} \left(\vec{y} + \vec{\eta}_{d+1} (\xi'(q_{d+1}) - \xi'(q_d))^{1/2} \right) \qquad \text{for } 0 \leq d \leq D-1.$$

By independence of the $1 \leq i \leq N$ coordinates in the G'_d, (5.5.3) implies

$$\varphi^{\mathrm{Sp}}(0) \leq \frac{1}{N} \sum_{i=1}^N \mathbb{E} \, \Gamma^i_0((h + \lambda m_i)\vec{1} + \vec{\eta}_0 \xi'(q_0)^{1/2}) + KR(\boldsymbol{h}, \boldsymbol{m}) + \frac{1}{2} KB. \tag{5.5.4}$$

It remains to compute the Gaussian integrals Γ^i_d. For this, we rely on the following lemma. We defer the proof, which is a standard computation with Gaussian integrals. Let $\mathbb{S}_K$ denote the set of $K \times K$ positive definite matrices, and let $|\cdot|$ denote the matrix determinant.

Lemma 5.5.5. *Suppose $\zeta > 0$ and $\Lambda, \Sigma \in \mathbb{S}_K$ satisfy $\Lambda - \zeta\Sigma \in \mathbb{S}_K$. If $\vec{v} \in \mathbb{R}^K$ and $\vec{\eta} \sim \mathcal{N}(0, \Sigma)$,*

then

$$\frac{1}{\zeta}\log\mathbb{E}\exp\frac{1}{2}\zeta\left[(\vec{y}+\vec{\eta})^\top\Lambda^{-1}(\vec{y}+\vec{\eta})-2\vec{v}^\top(\vec{y}+\vec{\eta})\right]$$
$$=\frac{1}{2}\left[\vec{y}^\top(\Lambda-\zeta\Sigma)^{-1}\vec{y}-2\vec{v}^\top\Lambda(\Lambda-\zeta\Sigma)^{-1}\vec{y}\right]+\frac{1}{2\zeta}\log\frac{|\Lambda|}{|\Lambda-\zeta\Sigma|}+\frac{1}{2}\vec{v}^\top(\zeta\Sigma)(\Lambda-\zeta\Sigma)^{-1}\Lambda\vec{v}.$$

We can compute the expectations in (5.5.4) by applying this lemma recursively. Define

$$\overline{\mathscr{K}}(\xi)=\left\{(B,\zeta)\in\mathbb{R}^+\times\overline{\mathscr{L}}:B>\int_{q_0}^1\xi''(q')\zeta(q')\,\mathrm{d}q'\right\}.$$

Proposition 5.5.6. *Let $\zeta\in\mathcal{M}_{\vec{q}}$, and suppose $(B,\kappa\zeta)\in\overline{\mathscr{K}}(\xi)$. Then, for $B_{\kappa\zeta}$ defined as in* (5.2.3)*,*

$$\mathbb{E}\,\Gamma_0^i((h+\lambda m_i)\vec{1}+\vec{\eta}_0\xi'(q_0)^{1/2})\le\frac{K}{2}\left[\frac{(h+(\lambda-B)m_i)^2+\xi'(q_0)}{B_{\kappa\zeta}(q_0)}+\int_{q_0}^1\frac{\xi''(q)}{B_{\kappa\zeta}(q)}\,\mathrm{d}q-Bm_i^2\right].$$

Proof. Let $\Lambda_D=BI_K$, and for $0\le d\le D-1$, let

$$\Lambda_d=\Lambda_{d+1}-\zeta_d(\xi'(q_{d+1})-\xi'(q_d))M^{d+1}.$$

We will first show that $\Lambda_0,\dots,\Lambda_D\in\mathbb{S}_K$, so that we can apply Lemma 5.5.5. For $q\in[q_0,1]$, we define

$$\Lambda(q)=BI_K-\int_q^1\xi''(q')M(q')\zeta(q')\,\mathrm{d}q'.$$

Note that $\Lambda_d=\Lambda(q_d)$ for all $0\le d\le D$. Since $M(q)\preceq\kappa(q)I_K$ in the Loewner order,

$$\Lambda(q)\succeq\left(B-\int_q^1\xi''(q')\kappa(q')\zeta(q')\,\mathrm{d}q'\right)I_K=B_{\kappa\zeta}(q)I_K.\tag{5.5.5}$$

So, the hypothesis $(B,\kappa\zeta)\in\overline{\mathscr{K}}(\xi)$ implies $\Lambda(q)\in\mathbb{S}_K$ for all $q\in[q_0,1]$. In particular $\Lambda_0,\dots,\Lambda_D\in\mathbb{S}_K$.

Further, define $\vec{v}_D=m_i\vec{1}$, and for $0\le d\le D-1$, define $\vec{v}_d=\Lambda_d^{-1}\Lambda_{d+1}\vec{v}_{d+1}$. This implies that $\vec{v}_d=Bm_i\Lambda_d^{-1}\vec{1}$. We can write Γ_D^i as

$$\Gamma_D^i(\vec{y})=\frac{1}{2}\left(\vec{y}^\top\Lambda_D^{-1}\vec{y}-2\vec{v}_D^\top\vec{y}\right).$$

By a recursive computation with Lemma 5.5.5 (which applies because $\Lambda_0,\dots,\Lambda_D\in\mathbb{S}_K$), we have

for all $0 \le d \le D$ that

$$\begin{aligned}\Gamma_d^i(\vec{y}) &= \frac{1}{2}\left[\vec{y}^\top \Lambda_d^{-1}\vec{y} - 2\vec{v}_d^\top \vec{y} + \sum_{d'=d}^{D-1}\frac{1}{\zeta_{d'}}\log\frac{|\Lambda_{d'+1}|}{|\Lambda_{d'}|} + \sum_{d'=d}^{D-1}\vec{v}_{d'+1}(\Lambda_{d'+1}-\Lambda_{d'})\Lambda_{d'}^{-1}\Lambda_{d'+1}\vec{v}_{d'+1}\right] \\ &= \frac{1}{2}\left[\vec{y}^\top \Lambda_d^{-1}\vec{y} - 2Bm_i\vec{1}^\top\Lambda_d^{-1}\vec{y} + \sum_{d'=d}^{D-1}\frac{1}{\zeta_{d'}}\log\frac{|\Lambda_{d'+1}|}{|\Lambda_{d'}|} + B^2m_i^2\sum_{d'=d}^{D-1}\vec{1}^\top\Lambda_{d'+1}^{-1}(\Lambda_{d'+1}-\Lambda_{d'})\Lambda_{d'}^{-1}\vec{1}\right].\end{aligned}$$

Note that

$$\sum_{d'=d}^{D-1}\vec{1}^\top\Lambda_{d'+1}^{-1}(\Lambda_{d'+1}-\Lambda_{d'})\Lambda_{d'}^{-1}\vec{1} = \sum_{d'=d}^{D-1}\vec{1}^\top(\Lambda_{d'}^{-1}-\Lambda_{d'+1}^{-1})\vec{1} = \vec{1}^\top(\Lambda_d^{-1}-\Lambda_D^{-1})\vec{1} = \vec{1}^\top\Lambda_d^{-1}\vec{1} - \frac{K}{B}.$$

So,

$$\begin{aligned}\Gamma_0^i(\vec{y}) &= \frac{1}{2}\left[\vec{y}^\top\Lambda_0^{-1}\vec{y} - 2Bm_i\vec{1}^\top\Lambda_0^{-1}\vec{y} + B^2m_i^2\vec{1}^\top\Lambda_0^{-1}\vec{1} + \sum_{d=0}^{D-1}\frac{1}{\zeta_d}\log\frac{|\Lambda_{d+1}|}{|\Lambda_d|} - KBm_i^2\right] \\ &= \frac{1}{2}\left[(\vec{y}-Bm_i\vec{1})^\top\Lambda_0^{-1}(\vec{y}-Bm_i\vec{1}) + \sum_{d=0}^{D-1}\frac{1}{\zeta_d}\log\frac{|\Lambda_{d+1}|}{|\Lambda_d|} - KBm_i^2\right]\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{E}\,\Gamma_0^i((h+\lambda m_i)\vec{1} + \vec{\eta}_0\xi'(q_0)^{1/2}) \\ &= \frac{1}{2}\left[(h+(\lambda-B)m_i)^2\mathsf{Tr}(\Lambda_0^{-1}\vec{1}\vec{1}^\top) + \xi'(q_0)\mathsf{Tr}(\Lambda_0^{-1}M^1) + \sum_{d=0}^{D-1}\frac{1}{\zeta_d}\log\frac{|\Lambda_{d+1}|}{|\Lambda_d|} - KBm_i^2\right].\end{aligned}$$

By Jacobi's formula,

$$\frac{\mathrm{d}}{\mathrm{d}q}\log|\Lambda(q)| = \xi''(q)\zeta(q)\mathsf{Tr}(\Lambda(q)^{-1}M(q)),$$

so

$$\frac{1}{\zeta_d}\log\frac{|\Lambda_{d+1}|}{|\Lambda_d|} = \int_{q_d}^{q_{d+1}}\xi''(q)\mathsf{Tr}(\Lambda(q)^{-1}M(q))\,\mathrm{d}q.$$

Therefore,

$$\begin{aligned}&\mathbb{E}\,\Gamma_0^i((h+\lambda m_i)\vec{1} + \vec{\eta}_0\xi'(q_0)^{1/2}) \\ &= \frac{1}{2}\left[(h+(\lambda-B)m_i)^2\mathsf{Tr}(\Lambda(q_0)^{-1}\vec{1}\vec{1}^\top) + \xi'(q_0)\mathsf{Tr}(\Lambda(q_0)^{-1}M(q_0)) + \int_{q_0}^1\mathsf{Tr}(\Lambda(q)^{-1}M(q))\,\mathrm{d}q - KBm_i^2\right].\end{aligned}$$

Finally, for each $q \in [q_0, 1)$, (5.5.5) implies $\Lambda(q)^{-1} \preceq \frac{I_K}{B_{\kappa\zeta}(q)}$, so

$$\mathsf{Tr}(\Lambda(q)^{-1}M(q)) \le \mathsf{Tr}\left(\frac{M(q)}{B_{\kappa\zeta}(q)}\right) = \frac{K}{B_{\kappa\zeta}(q)},$$

and similarly $\mathsf{Tr}(\Lambda(q_0)^{-1}\vec{1}\vec{1}^\top) \le \frac{K}{B_{\kappa\zeta}(q_0)}$. This implies the result. □

Proposition 5.5.6 and (5.5.4) readily imply the following bound on $F_N^{\mathrm{Sp}}(\mathcal{Q}(\eta))$.

Proposition 5.5.7. *Let $B \ge 1$, $N \ge 2$, and $\lambda \in \mathbb{R}$. Let $\zeta \in \mathcal{M}_{\vec{q}}$, and suppose $(B, \kappa\zeta) \in \overline{\mathscr{K}}(\xi)$. Then,*

$$F_N^{\mathrm{Sp}}(\mathcal{Q}(\eta)) \le \frac{K}{2}\left[\frac{\|\boldsymbol{h} + (\lambda - B)\boldsymbol{m}\|_N^2 + \xi'(q_0)}{B_{\kappa\zeta}(q_0)} + 2R(\boldsymbol{h}, \boldsymbol{m}) + \int_{q_0}^1 \left(\frac{\xi''(q)}{B_{\kappa\zeta}(q)} + B_{\kappa\zeta}(q)\right)\,\mathrm{d}q\right] + 3K^2\xi''(1)\eta + K|\lambda|\eta.$$

Proof. By averaging Proposition 5.5.6 over $1 \le i \le N$, we get

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}\,\Gamma_0^i((h + \lambda m_i)\vec{1} + \vec{\eta}_0\xi'(q_0)^{1/2}) \le \frac{K}{2}\left[\frac{\|\boldsymbol{h} + (\lambda - B)\boldsymbol{m}\|_N^2 + \xi'(q_0)}{B_{\kappa\zeta}(q_0)} + \int_{q_0}^1 \frac{\xi''(q)}{B_{\kappa\zeta}(q)}\,\mathrm{d}q - Bq_0\right]$$

where we used that $\|\boldsymbol{m}\|_N^2 = q_0$. Equation (5.5.4) implies that

$$\varphi^{\mathrm{Sp}}(0) \le \frac{K}{2}\left[\frac{\|\boldsymbol{h} + (\lambda - B)\boldsymbol{m}\|_N^2 + \xi'(q_0)}{B_{\kappa\zeta}(q_0)} + 2R(\boldsymbol{h}, \boldsymbol{m}) + \int_{q_0}^1 \frac{\xi''(q)}{B_{\kappa\zeta}(q)}\,\mathrm{d}q + (1 - q_0)B\right].$$

By Proposition 5.4.1, this implies

$$F_N^{\mathrm{Sp}}(\mathcal{Q}(\eta)) \le \frac{K}{2}\left[\frac{\|\boldsymbol{h} + (\lambda - B)\boldsymbol{m}\|_N^2 + \xi'(q_0)}{B_{\kappa\zeta}(q_0)} + 2R(\boldsymbol{h}, \boldsymbol{m}) + \int_{q_0}^1 \frac{\xi''(q)}{B_{\kappa\zeta}(q)}\,\mathrm{d}q + (1 - q_0)B - \int_{q_0}^1 (q - q_0)\xi''(q)\kappa(q)\zeta(q)\,\mathrm{d}q\right] + 3K^2\xi''(1)\eta + K|\lambda|\eta$$

By integration by parts,

$$-\int_{q_0}^1 (q - q_0)\xi''(q)\kappa(q)\zeta(q)\,\mathrm{d}q = (q - q_0)\int_q^1 \xi''(q')\kappa(q')\zeta(q')\,\mathrm{d}q'\Big|_{q=q_0}^1 - \int_{q_0}^1\int_q^1 \xi''(q')\kappa(q')\zeta(q')\,\mathrm{d}q'\,\mathrm{d}q = \int_{q_0}^1 B_{\kappa\zeta}(q)\,\mathrm{d}q - (1 - q_0)B,$$

which yields the result. □

The next lemma upper bounds our estimates for $F_N^{\mathrm{Sp}}(\mathcal{Q}(\eta))$ in terms of the Parisi functional uniformly in $\boldsymbol{m}$.

Lemma 5.5.8. *Let $q_0 \in [0, 1]$. For $(B, \zeta) \in \mathscr{K}(\xi)$, $\boldsymbol{h} = (h, \ldots, h)$, $\|\boldsymbol{m}\|_N^2 = q_0$, there exists*

$\lambda \in [0, B]$ *such that*

$$\frac{1}{2}\left[\frac{\|\boldsymbol{h} + (\lambda - B)\boldsymbol{m}\|_N^2 + \xi'(q_0)}{B_\zeta(q_0)} + 2R(\boldsymbol{h}, \boldsymbol{m}) + \int_{q_0}^1 \left(\frac{\xi''(q)}{B_\zeta(q)} + B_\zeta(q)\right)\,\mathrm{d}q\right] \le \mathsf{P}^{\mathrm{Sp}}(\zeta).$$

Proof. We take $\lambda = \int_0^1 \xi''(q)\zeta(q)\,\mathrm{d}q$. The condition $(B, \zeta) \in \mathscr{K}(\xi)$ implies that $\lambda \in [0, B]$. Note that $\lambda - B = -B_\zeta(0)$. It suffices to prove that

$$\frac{\|\boldsymbol{h} - B_\zeta(0)\boldsymbol{m}\|_N^2 + \xi'(q_0)}{B_\zeta(q_0)} + 2R(\boldsymbol{h}, \boldsymbol{m}) \le \frac{\|\boldsymbol{h}\|_N^2}{B_\zeta(0)} + \int_0^{q_0} \left(\frac{\xi''(q)}{B_\zeta(q)} + B_\zeta(q)\right)\,\mathrm{d}q.$$

Note that

$$\frac{\xi'(q_0)}{B_\zeta(q_0)} = \int_0^{q_0} \frac{\xi''(q)}{B_\zeta(q_0)}\,\mathrm{d}q \le \int_0^{q_0} \frac{\xi''(q)}{B_\zeta(q)}\,\mathrm{d}q \qquad \text{and} \qquad q_0 B_\zeta(0) \le \int_0^{q_0} B_\zeta(q)\,\mathrm{d}q.$$

So, it suffices to prove that

$$\frac{\|\boldsymbol{h} - B_\zeta(0)\boldsymbol{m}\|_N^2}{B_\zeta(q_0)} + 2R(\boldsymbol{h}, \boldsymbol{m}) \le \frac{\|\boldsymbol{h}\|_N^2}{B_\zeta(0)} + q_0 B_\zeta(0).$$

This rearranges to (using that $\|\boldsymbol{m}\|_N^2 = q_0$)

$$0 \le \left(\frac{1}{B_\zeta(0)} - \frac{1}{B_\zeta(q_0)}\right)\left(\|\boldsymbol{h}\|_N^2 - 2B_\zeta(0)R(\boldsymbol{h}, \boldsymbol{m}) + B_\zeta(0)^2\|\boldsymbol{m}\|_N^2\right),$$

which follows from Cauchy-Schwarz. □

We are now ready to prove Proposition 5.5.2.

Proof of Proposition 5.5.2. Recall that the restriction of $\underline{\zeta} + \kappa\zeta \in \mathscr{L}$ on $[q_0, 1)$ is $\kappa\zeta$. Because $(B, \underline{\zeta} + \kappa\zeta) \in \mathscr{K}(\xi)$, we have $(B, \kappa\zeta) \in \overline{\mathscr{K}}(\xi)$, and so Proposition 5.5.7 applies. Combining this with Lemma 5.5.8 applied on $(B, \underline{\zeta} + \kappa\zeta)$ gives the result. □

5.5.2 From Free Energy to Ground State Energy

Next, we will prove Proposition 5.5.1 by taking Proposition 5.5.2 to low temperature. We introduce the following temperature-scaled free energy. For $\beta > 0$ and $\eta \in (0, 1)$, let

$$F_N^{\mathrm{Sp}}(\beta, \mathcal{Q}(\eta)) = \frac{1}{N}\log \mathbb{E}\int_{\mathcal{Q}(\eta)} \exp \beta \mathcal{H}_N(\vec{\boldsymbol{\sigma}})\,\mathrm{d}\mu^K(\vec{\boldsymbol{\sigma}}).$$

This free energy can be upper bounded by the following application of Proposition 5.5.2.

Corollary 5.5.9. *Let $\zeta \in \mathcal{M}_{\vec{q}}$ and $\underline{\zeta} \in \underline{\mathscr{L}}$ be arbitrary. Let $\beta > 0$ and suppose $(B, \underline{\zeta} + \beta\kappa\zeta) \in \mathscr{K}(\xi)$, $B \geq \beta^{-1}$, and $N \geq 2$. Then,*

$$\frac{1}{\beta} F_N^{\mathrm{Sp}}(\beta, \mathcal{Q}(\eta)) \leq K \mathsf{P}^{\mathrm{Sp}}(B, \underline{\zeta} + \beta\kappa\zeta) + 3K^2\xi''(1)\beta\eta + KB\eta.$$

Proof. The hypothesis $(B, \underline{\zeta} + \beta\kappa\zeta) \in \mathscr{K}(\xi)$ implies $(\beta B, \beta^{-1}\underline{\zeta} + \kappa\zeta) \in \mathscr{K}(\beta^2\xi)$. The hypothesis $B \geq \beta^{-1}$ implies $\beta B \geq 1$. By Proposition 5.5.2 with parameters $(\beta^2\xi, \beta h)$ (corresponding to the Hamiltonian βH_N), ζ, βB, and $\beta^{-1}\underline{\zeta}$,

$$F_N^{\mathrm{Sp}}(\beta, \mathcal{Q}(\eta)) \leq K \mathsf{P}^{\mathrm{Sp}}_{\beta^2\xi, \beta h}(\beta B, \beta^{-1}\underline{\zeta} + \kappa\zeta) + 3K^2\xi''(1)\beta^2\eta + KB\beta\eta. \tag{5.5.6}$$

We can verify that

$$\mathsf{P}^{\mathrm{Sp}}_{\beta^2\xi, \beta h}(\beta B, \beta^{-1}\underline{\zeta} + \kappa\zeta) = \beta \mathsf{P}^{\mathrm{Sp}}_{\xi, h}(B, \underline{\zeta} + \beta\kappa\zeta).$$

So, dividing (5.5.6) by β gives the result. □

The following lemma relates the ground state energy $\mathsf{OPT}_N^{\mathrm{Sp}}(\mathcal{Q}(\eta))$ to this free energy at large inverse temperature β. We defer the proof, which is a relatively standard approximation argument.

Lemma 5.5.10. *There exists a constant C depending only on ξ, h such that for all $\beta > 0$, $\eta \in (0, \frac{1}{2})$, and $N \geq C \log \max(K, 2)$,*

$$\mathsf{OPT}_N^{\mathrm{Sp}}(\mathcal{Q}(\eta)) \leq \frac{1}{\beta} F_N^{\mathrm{Sp}}(\beta, \mathcal{Q}(2\eta)) + CK\left(\eta + \frac{\log \frac{1}{\eta}}{\beta} + \frac{1}{\sqrt{N}}\right).$$

Proof of Proposition 5.5.1. Let C be large enough that Lemma 5.5.10 is satisfied and $C \log 2 \geq 2$. For all $N \geq C \log \max(K, 2)$, Corollary 5.5.9 (with 2η in place of η) and Lemma 5.5.10 imply that

$$\mathsf{OPT}_N^{\mathrm{Sp}}(\eta) \leq K \mathsf{P}^{\mathrm{Sp}}(B, \underline{\zeta} + \beta\kappa\zeta) + 6K^2\xi''(1)\beta\eta + 2KB\eta + CK\left(\eta + \frac{\log \frac{1}{\eta}}{\beta} + \frac{1}{\sqrt{N}}\right).$$

By applying the estimate $K \leq K^2$ and absorbing constants depending on only ξ, h into C, we deduce

$$\mathsf{OPT}_N^{\mathrm{Sp}}(\eta) \leq K \mathsf{P}^{\mathrm{Sp}}(B, \underline{\zeta} + \beta\kappa\zeta) + CK^2\left(\beta\eta + B\eta + \eta + \frac{\log \frac{1}{\eta}}{\beta} + \frac{1}{\sqrt{N}}\right).$$

Finally, because $B \geq \beta^{-1}$, we have $\beta + B \geq \beta + \beta^{-1} \geq 2$, so by increasing the constant C we may drop the term η from the sum. □

5.5.3 Proof of the Main Upper Bound

We now complete the proof of Proposition 5.3.2. We will set the parameters of Proposition 5.5.1 such that $(B, \underline{\zeta} + \beta\kappa\zeta)$ approximates the minimizer of P^{Sp} in $\mathscr{L}$ and the error term is small.

For $\zeta \in \mathscr{L}$ and $\delta, x \in [0,1)$, we define a perturbation $\zeta_{\delta,x} \in \mathscr{L}$ of ζ by

$$\zeta_{\delta,x}(q) = \begin{cases} \zeta(x+\delta) & q \in [x, x+\delta), \\ \zeta(q) & \text{otherwise.} \end{cases}$$

Note that $\zeta_{0,x} = \zeta$.

We now set several constants depending only on ξ, h, ε. Let C be the constant given by Proposition 5.5.1. By continuity of the Parisi functional P^{Sp} on $\mathscr{K}(\xi)$, we may pick $(B^*, \zeta^*) \in \mathscr{K}(\xi)$ and a small constant $\Delta \in (0,1)$ such that the following properties hold.

(a) ζ^* is positive-valued, right-continuous, and piecewise constant with finitely many jump discontinuities $0 < x_1 < \cdots < x_r < 1$.

(b) For all $\delta \in [0, \Delta]$ and $x \in [0,1)$, $(B^*, \zeta^*_{\delta,x}) \in \mathscr{K}(\xi)$ and

$$\mathsf{P}^{\mathrm{Sp}}(B^*, \zeta^*_{\delta,x}) \le \mathsf{ALG} + \frac{\varepsilon}{2}. \tag{5.5.7}$$

The perturbations $\zeta^*_{\delta,x}$ will be used in the following way. Given $q_0 \in [0,1]$, we will apply Proposition 5.5.1 with $\underline{\zeta} + \beta\kappa\zeta = \zeta^*_{(1-q_0)\Delta,q_0}$. In particular, we will construct β, $\kappa = \kappa^{\vec{k},\vec{p},\vec{q}}$ and $\zeta \in \mathcal{M}_{\vec{q}}$ such that $\beta\kappa\zeta = \zeta^*_{(1-q_0)\Delta,q_0}$ on $[q_0, 1)$. Because ζ is increasing, we must construct a κ that decreases rapidly enough to make this equality hold. In the below proof, the fact that $\zeta^*_{(1-q_0)\Delta,q_0}$ does not have any discontinuities in $[q_0, q_0 + (1-q_0)\Delta]$ implies that $q_1 > q_0 + (1-q_0)\Delta$, which implies that $p_1 > \Delta$ for any χ-aligned $\vec{p}, \vec{q}$. This allows us to construct a suitable κ while keeping $K = \prod_{d=1}^{D} k_d$ bounded by a constant.

Proof of Proposition 5.3.2, spherical case. We first set the constants K_0, η_0, N_0. For $x \in (0,1]$, let $\zeta^*(x^-) = \lim_{y \to x^-} \zeta^*(y)$. Let

$$K_0 = \prod_{j=1}^{r} \left(\left\lfloor \frac{\zeta^*(x_j)}{\Delta\zeta^*(x_j^-)} \right\rfloor + 1 \right).$$

This is well-defined because ζ^* is positive-valued. Let $\eta_0 \in (0, \frac{1}{2})$ satisfy the inequalities

$$CK_0\left(B^*\eta_0 + \eta_0^{1/2} + \eta_0^{1/2}\log\frac{1}{\eta_0}\right) \le \frac{\varepsilon}{4}, \tag{5.5.8}$$

$$\eta_0 \le (B^*)^2, \tag{5.5.9}$$

$$\eta_0 < \zeta^*(1^-)^{-2}. \tag{5.5.10}$$

Finally, let N_0 satisfy $N_0 \ge C\log\max(K_0, 2)$ and

$$\frac{CK_0}{\sqrt{N_0}} \le \frac{\varepsilon}{4}. \tag{5.5.11}$$

We emphasize that K_0, η_0, N_0 depend only on ξ, h, ε.

In the below analysis, we always set $\eta = \eta_0$ (this clearly satisfies $\eta \ge \eta_0$) and $\beta = \eta_0^{-1/2}$.

We are given a correlation function $\chi : [0,1] \to [0,1]$ and a point $\boldsymbol{m} \in \mathbb{R}^N$ with $\|\boldsymbol{m}\|_N^2 = \chi(0)$. We set $q_0 = \chi(0)$; we will set the rest of $\vec{q}$ below. We will construct $D, \vec{k}, \vec{p}, \vec{q}, \zeta$ such that on $[q_0, 1)$,

$$\beta\kappa^{\vec{k},\vec{p},\vec{q}}\zeta = \zeta^*_{(1-q_0)\Delta,q_0}. \tag{5.5.12}$$

Let

$$S = \{x_1, \dots, x_r\} \cap (q_0 + (1-q_0)\Delta, 1).$$

Set $D - 1 = |S|$. Set $\vec{q}$ such that $(q_1, \dots, q_{D-1})$ is the set S in increasing order and $q_D = 1$.

By Proposition 5.3.1(ii), χ is either strictly increasing or constant. If χ is strictly increasing, set $\vec{p} = (p_0, \dots, p_D)$ by $p_d = \chi^{-1}(q_d)$ for all $q_d \le \chi(1)$ and $p_d = 1$ for all $q_d > \chi(1)$. If χ is constant, its unique value is $q_0 = \chi(0)$; set $p_0 = 0$ and $p_d = 1$ for all $1 \le d \le D$. In either case, $\vec{p}, \vec{q}$ are clearly χ-aligned. Moreover, we always have $p_1 > \Delta$: if χ is increasing, this follows from $q_1 > q_0 + (1-q_0)\Delta$ and Proposition 5.3.1(iii), while if χ is constant this is obvious.

Set $k_1 = 1$, and for $1 \le d \le D-1$, set

$$k_{d+1} = \left\lfloor \frac{\zeta^*(q_d^-)}{\Delta\zeta^*(q_d)} \right\rfloor + 1.$$

Because $q_1, \dots, q_{D-1}$ are a subset of $x_1, \dots, x_r$, we indeed have $K = \prod_{d=1}^D k_d \le K_0$.

This constructs $D, \vec{k}, \vec{p}, \vec{q}, \eta$, which defines $\mathcal{H}_N^{\vec{k},\vec{p}}$, $\mathcal{Q}(\eta) = \mathcal{Q}^{\mathrm{Sp}}(Q^{\vec{k},\vec{q}}, \boldsymbol{m}, \eta)$, and $\kappa^{\vec{k},\vec{p},\vec{q}}$. Finally, we construct the sequence $(\zeta_{-1}, \zeta_0, \dots, \zeta_D)$ satisfying

$$0 = \zeta_{-1} < \zeta_0 < \cdots < \zeta_D = 1 \tag{5.5.13}$$

such that the $\zeta \in \mathcal{M}_{\vec{q}}$ defined by (5.4.3) satisfies (5.5.12) on $[q_0, 1)$. In particular, we define ζ_d for

$0 \le d \le D-1$ by

$$\zeta_d = \frac{\zeta^*_{(1-q_0)\Delta,q_0}(q_d)}{\beta\kappa^{\vec{k},\vec{p},\vec{q}}(q_d)}.$$

For this choice of ζ_d, (5.5.12) holds at $q_0, q_1, \dots, q_{d-1}$ by inspection. Because ζ, $\kappa^{\vec{k},\vec{p},\vec{q}}$ and $\zeta^*_{(1-q_0)\Delta,q_0}$ are all piecewise constant and right-continuous on $[q_0, 1)$ with jump discontinuities only at $q_1, \dots, q_{D-1}$, (5.5.12) holds on $[q_0, 1)$. It remains to verify that this choice of ζ_d satisfies the increasing condition (5.5.13). Because $\zeta^*_{(1-q_0)\Delta,q_0}$ is positive-valued, $\zeta_0 > \zeta_{-1} = 0$. At each $1 \le d \le D-1$, we have

$$\frac{\zeta_d}{\zeta_{d-1}} = \frac{\zeta^*_{(1-q_0)\Delta,q_0}(q_d)}{\zeta^*_{(1-q_0)\Delta,q_0}(q_{d-1})} \cdot \frac{\kappa^{\vec{k},\vec{p},\vec{q}}(q_{d-1})}{\kappa^{\vec{k},\vec{p},\vec{q}}(q_d)}$$

By (5.4.2),

$$\kappa^{\vec{k},\vec{p},\vec{q}}(q_d) \le \sum_{j=d+1}^{D-1} \left[(k_{j+1}-1) \prod_{\ell=j+2}^{D} k_\ell \right] + 1 = \prod_{\ell=d+2}^{D} k_\ell,$$

where we upper bounded all the p_d by 1. So,

$$\frac{\kappa^{\vec{k},\vec{p},\vec{q}}(q_{d-1})}{\kappa^{\vec{k},\vec{p},\vec{q}}(q_d)} = 1 + \frac{(k_{d+1}-1)\prod_{\ell=d+2}^{D} k_\ell}{\kappa^{\vec{k},\vec{p},\vec{q}}(q_d)} p_d \ge 1 + (k_{d+1}-1)p_d \ge k_{d+1}p_d \ge k_{d+1}\Delta.$$

Here we used that $p_d \ge p_1 \ge \Delta$. Further noting that $\zeta^*_{(1-q_0)\Delta,q_0}(q_{d-1}) = \zeta^*(q_d^-)$, we have

$$\frac{\zeta_d}{\zeta_{d-1}} \ge \frac{\Delta\zeta^*_{(1-q_0)\Delta,q_0}(q_d)}{\zeta^*_{(1-q_0)\Delta,q_0}(q_{d-1})} \cdot k_{d+1} = \frac{\Delta\zeta^*(q_d)}{\zeta^*(q_d^-)} \cdot k_{d+1} > 1$$

by definition of k_{d+1}. Thus $\zeta_d > \zeta_{d-1}$ for $1 \le d \le D-1$. Finally, because $\kappa^{\vec{k},\vec{p},\vec{q}}(q_{D-1}) = 1$,

$$\zeta_{D-1} = \frac{\zeta^*_{(1-q_0)\Delta,q_0}(q_{D-1})}{\beta} = \eta_0^{1/2}\zeta^*(1^-) < 1 = \zeta_D,$$

using (5.5.10). Thus the ζ we constructed satisfies (5.5.12) and (5.5.13).

Define $\underline{\zeta} \in \underline{\mathscr{L}}$ by $\underline{\zeta} = \zeta^*$ on $[0, q_0)$. Thus, as elements of $\mathscr{L}$,

$$\underline{\zeta} + \beta\kappa^{\vec{k},\vec{p},\vec{q}}\zeta = \zeta^*_{(1-q_0)\Delta,q_0}.$$

By construction, $(B^*, \zeta^*_{(1-q_0)\Delta,q_0}) \in \mathscr{K}(\xi)$, and (5.5.9) implies $B^* \ge \beta^{-1}$. By Proposition 5.5.1,

$$\frac{1}{N}\mathbb{E}\max_{\vec{\boldsymbol{\sigma}} \in \mathcal{Q}(\eta)} \mathcal{H}_N(\vec{\boldsymbol{\sigma}}) \le K\mathsf{P}^{\mathrm{Sp}}(B^*, \zeta^*_{(1-q_0)\Delta,q_0}) + CK^2\left(B^*\eta + \eta^{1/2} + \eta^{1/2}\log\frac{1}{\eta} + \frac{1}{\sqrt{N}}\right).$$

By (5.5.7),

$$K\mathsf{P}^{\mathrm{Sp}}(B^*, \zeta^*_{(1-q_0)\delta,q_0}) \le K\left(\mathsf{ALG} + \frac{\varepsilon}{2}\right).$$

By (5.5.8),

$$CK^2\left(B^*\eta + \eta^{1/2} + \eta^{1/2}\log\frac{1}{\eta}\right) \le \frac{K\varepsilon}{4}.$$

Finally, by (5.5.11),

$$\frac{CK^2}{\sqrt{N}} \le \frac{K\varepsilon}{4}.$$

Combining the last four inequalities gives the result. □

5.5.4 Deferred Proofs

Here we give the proofs of Lemmas 5.5.3, 5.5.4, 5.5.5, and 5.5.10, which are all relatively standard. We recall the following lemma, due to Talagrand, from which Lemma 5.5.3 readily follows.

Lemma 5.5.11 ([Tal06a, Lemma 3.1])**.** *For all $\boldsymbol{y} \in \mathbb{R}^N$, the following inequality holds.*

$$\int_{S_N} \exp\langle \boldsymbol{y}, \boldsymbol{\sigma}\rangle \, \mathrm{d}\mu(\boldsymbol{\sigma}) \le \mathbb{P}\left(\chi^2(N) \ge BN\right)^{-1} \int \exp\langle \boldsymbol{y}, \boldsymbol{\rho}\rangle \, \mathrm{d}\nu_B^N(\boldsymbol{\rho}).$$

Proof of Lemma 5.5.3. Using $\mathcal{Q}(\eta) \subseteq S_N^K$ and Lemma 5.5.11, we get

$$\begin{aligned}
\exp G_D(\vec{\boldsymbol{y}}) &= \exp(-\langle \vec{\boldsymbol{y}}, \vec{\boldsymbol{m}}\rangle) \int_{\mathcal{Q}(\eta)} \exp\langle \vec{\boldsymbol{y}}, \vec{\boldsymbol{\sigma}}\rangle \, \mathrm{d}\mu^K(\vec{\boldsymbol{\sigma}}) \\
&\le \exp(-\langle \vec{\boldsymbol{y}}, \vec{\boldsymbol{m}}\rangle) \int_{S_N^K} \exp\langle \vec{\boldsymbol{y}}, \vec{\boldsymbol{\sigma}}\rangle \, \mathrm{d}\mu^K(\vec{\boldsymbol{\sigma}}) \\
&= \exp(-\langle \vec{\boldsymbol{y}}, \vec{\boldsymbol{m}}\rangle) \prod_{u\in\mathbb{L}} \int_{S_N} \exp\langle \boldsymbol{y}(u), \boldsymbol{\sigma}(u)\rangle \, \mathrm{d}\mu(\vec{\boldsymbol{\sigma}}(u)) \\
&\le \mathbb{P}\left(\chi^2(N) \ge BN\right)^{-K} \exp(-\langle \vec{\boldsymbol{y}}, \vec{\boldsymbol{m}}\rangle) \prod_{u\in\mathbb{L}} \int \exp\langle \boldsymbol{y}(u), \boldsymbol{\rho}(u)\rangle \, \mathrm{d}\nu_B^N(\vec{\boldsymbol{\rho}}(u)) \\
&= \mathbb{P}\left(\chi^2(N) \ge BN\right)^{-K} \exp(-\langle \vec{\boldsymbol{y}}, \vec{\boldsymbol{m}}\rangle) \int \exp\langle \vec{\boldsymbol{y}}, \vec{\boldsymbol{\rho}}\rangle \, \mathrm{d}\nu_B^{N\times K}(\vec{\boldsymbol{\rho}}).
\end{aligned}$$

□

Proof of Lemma 5.5.4. Using the probability density of $\chi^2(N)$, we compute:

$$\begin{aligned}
\mathbb{P}(\chi^2(N) \geq BN) &= \int_{BN}^{\infty} \frac{x^{N/2-1}e^{-x/2}}{2^{N/2}\Gamma\left(\frac{N}{2}\right)}\,\mathrm{d}x \\
&= \frac{(N/2)^{N/2}}{\Gamma\left(\frac{N}{2}\right)} \int_{B}^{\infty} y^{N/2-1}e^{-Ny/2}\,\mathrm{d}y \\
&\geq \frac{(N/2)^{N/2}}{\Gamma\left(\frac{N}{2}\right)} \int_{B}^{\infty} e^{-Ny/2}\,\mathrm{d}y \\
&= \frac{(N/2)^{N/2-1}}{\Gamma\left(\frac{N}{2}\right)} e^{-BN/2} \\
&\geq e^{-BN/2},
\end{aligned}$$

where the last step uses that $(N/2)^{N/2-1} \geq \Gamma\left(\frac{N}{2}\right)$ for $N \geq 2$. □

Proof of Lemma 5.5.5. By a straightforward computation,

$$\begin{aligned}
&\mathbb{E}\exp\frac{1}{2}\zeta\left[(\vec{y}+\vec{\eta})^\top\Lambda^{-1}(\vec{y}+\vec{\eta}) - 2\vec{v}^\top(\vec{y}+\vec{\eta})\right] \\
&= |\Sigma|^{-1/2}(2\pi)^{-K/2}\int \exp\left[-\frac{1}{2}\left(\vec{x}^\top\Sigma^{-1}\vec{x} - \zeta(\vec{y}+\vec{x})^\top\Lambda^{-1}(\vec{y}+\vec{x}) + 2\zeta\vec{v}^\top(\vec{y}+\vec{x})\right)\right]\,\mathrm{d}\vec{x} \\
&= |\Sigma|^{-1/2}(2\pi)^{-K/2}\int \exp\left[-\frac{1}{2}\left(\vec{x}^\top\left(\Sigma^{-1}-\zeta\Lambda^{-1}\right)\vec{x} - 2\zeta(\Lambda^{-1}\vec{y}-\vec{v})^\top\vec{x} - \zeta\vec{y}^\top\Lambda^{-1}\vec{y} + 2\zeta\vec{v}^\top\vec{y}\right)\right]\,\mathrm{d}\vec{x} \\
&= |\Sigma|^{-1/2}|\Sigma^{-1}-\zeta\Lambda^{-1}|^{-1/2}\exp\frac{1}{2}\left(\zeta^2(\Lambda^{-1}\vec{y}-\vec{v})^\top\left(\Sigma^{-1}-\zeta\Lambda^{-1}\right)^{-1}(\Lambda^{-1}\vec{y}-\vec{v}) + \zeta\vec{y}^\top\Lambda^{-1}\vec{y} - 2\zeta\vec{v}^\top\vec{y}\right) \\
&= \frac{|\Lambda|^{1/2}}{|\Lambda-\zeta\Sigma|^{1/2}}\exp\frac{\zeta}{2}\left(\vec{y}^\top(\Lambda-\zeta\Sigma)^{-1}\vec{y} - 2\vec{v}^\top\Lambda(\Lambda-\zeta\Sigma)^{-1}\vec{y} + \vec{v}^\top(\zeta\Sigma)(\Lambda-\zeta\Sigma)^{-1}\Lambda\vec{v}\right).
\end{aligned}$$

Taking logarithms and dividing by ζ yields the result. □

Proof of Lemma 5.5.10. Define the random variable

$$\vec{\boldsymbol{\sigma}}^* = \mathrm{argmax}_{\vec{\boldsymbol{\sigma}}\in\mathcal{Q}(\eta)}\mathcal{H}_N(\vec{\boldsymbol{\sigma}}),$$

where we break ties arbitrarily. For $\delta > 0$, define

$$\mathcal{B}(\vec{\boldsymbol{\sigma}}^*, \delta) = \left\{\vec{\boldsymbol{\sigma}} \in S_N^K : \|\boldsymbol{\sigma}(u) - \boldsymbol{\sigma}^*(u)\|_N \leq \delta \text{ for all } u \in \mathbb{L}\right\}.$$

If $\vec{\boldsymbol{\sigma}} \in \mathcal{B}(\vec{\boldsymbol{\sigma}}^*, \eta/3)$, then for each $u \in \mathbb{L}$ we can write $\boldsymbol{\sigma}(u) = \boldsymbol{\sigma}^*(u) + \delta(u)\boldsymbol{\rho}(u)$, where $\boldsymbol{\rho}(u) \in S_N$ and $0 \leq \delta(u) \leq \eta/3$. Then, for all $u \in \mathbb{L}$,

$$|R(\boldsymbol{\sigma}(u), \boldsymbol{m}) - q_0| \leq |R(\boldsymbol{\sigma}^*(u), \boldsymbol{m}) - q_0| + \delta(u)|R(\boldsymbol{\rho}(u), \boldsymbol{m})| \leq \eta + \eta/3 \leq 2\eta,$$

and for all $u, v \in \mathbb{L}$,

$$
\begin{aligned}
&|R(\boldsymbol{\sigma}(u), \boldsymbol{\sigma}(v)) - q_{u\wedge v}| \\
&\le |R(\boldsymbol{\sigma}^*(u), \boldsymbol{\sigma}^*(v)) - q_{u\wedge v}| + \delta(u)|R(\boldsymbol{\sigma}^*(u), \boldsymbol{\rho}(v))| + \delta(v)|R(\boldsymbol{\sigma}^*(v), \boldsymbol{\rho}(u))| + \delta(u)\delta(v)|R(\boldsymbol{\rho}(u), \boldsymbol{\rho}(u))| \\
&\le \eta + \eta/3 + \eta/3 + \eta/3 = 2\eta.
\end{aligned}
$$

So, $\mathcal{B}(\vec{\boldsymbol{\sigma}}^*, \eta/3) \subseteq \mathcal{Q}(2\eta)$.

Let constants c, C_1 be given by Proposition 5.2.3. By this proposition, the event

$$
S = \left\{ \sup_{u\in\mathbb{L}} \sup_{\boldsymbol{\sigma}\in S_N} \left\| \nabla H_N^{(u)}(\boldsymbol{\sigma}) \right\|_N \le C_1 \right\}
$$

has probability $\mathbb{P}(S) \ge 1 - Ke^{-cN}$. Here we use the fact that for $\boldsymbol{v} \in \mathbb{R}^N$, $\|\boldsymbol{v}\|_N = \|\boldsymbol{v}\|_{\mathrm{op}}$. On S,

$$
\mathcal{H}_N(\vec{\boldsymbol{\sigma}}) \ge \mathcal{H}_N(\vec{\boldsymbol{\sigma}}^*) - \frac{C_1 NK\eta}{3}
$$

for all $\vec{\boldsymbol{\sigma}} \in \mathcal{B}(\vec{\boldsymbol{\sigma}}^*, \eta/3)$. So,

$$
\begin{aligned}
F_N^{\mathrm{Sp}}(\beta, \mathcal{Q}(2\eta)) &= \frac{1}{N} \log \mathbb{E} \int_{\mathcal{Q}(2\eta)} \exp \beta \mathcal{H}_N(\vec{\boldsymbol{\sigma}}) \, \mathrm{d}\mu^K(\vec{\boldsymbol{\sigma}}) \\
&\ge \frac{1}{N} \log \mathbb{E}\, \mathbb{I}(S) \int_{\mathcal{B}(\vec{\boldsymbol{\sigma}}^*, \eta/3)} \exp \beta \mathcal{H}_N(\vec{\boldsymbol{\sigma}}) \, \mathrm{d}\mu^K(\vec{\boldsymbol{\sigma}}) \\
&\ge \frac{1}{N} \log \mathbb{E}\, \mathbb{I}(S) \int_{\mathcal{B}(\vec{\boldsymbol{\sigma}}^*, \eta/3)} \exp \beta \left(\mathcal{H}_N(\vec{\boldsymbol{\sigma}}^*) - \frac{C_1 NK\eta}{3} \right) \, \mathrm{d}\mu^K(\vec{\boldsymbol{\sigma}}) \\
&\ge \frac{1}{N} \log \mathbb{E}\, \mathbb{I}(S) \exp \beta \mathcal{H}_N(\vec{\boldsymbol{\sigma}}^*) - \frac{\beta C_1 K\eta}{3} + \frac{1}{N} \log \mu^K(\mathcal{B}(\vec{\boldsymbol{\sigma}}^*, \eta/3)) \\
&= \beta \mathsf{OPT}_N^{\mathrm{Sp}}(\mathcal{Q}(\eta)) - \frac{\beta C_1 K\eta}{3} + \frac{1}{N} \log \mu^K(\mathcal{B}(\vec{\boldsymbol{\sigma}}^*, \eta/3)) \\
&\quad + \frac{1}{N} \log \mathbb{E}\, \mathbb{I}(S) \exp \beta \left(\mathcal{H}_N(\vec{\boldsymbol{\sigma}}^*) - \mathbb{E}\, \mathcal{H}_N(\vec{\boldsymbol{\sigma}}^*) \right).
\end{aligned}
$$

The set $\mathcal{B}(\vec{\boldsymbol{\sigma}}^*, \eta/3)$ is the product of K spherical caps in S_N. By elementary properties of the spherical measure, there exists a large C such that $\mu^K(\mathcal{B}(\vec{\boldsymbol{\sigma}}^*, \eta/3)) \le \eta^{CNK}$, and so

$$
\frac{1}{N} \log \mu^K(\mathcal{B}(\vec{\boldsymbol{\sigma}}^*, \eta/3)) \ge -CK \log \frac{1}{\eta}.
$$

By Proposition 5.3.8,

$$
\mathbb{P}\left(\mathcal{H}_N(\vec{\boldsymbol{\sigma}}^*) - \mathbb{E}\, \mathcal{H}_N(\vec{\boldsymbol{\sigma}}^*) \le -K\sqrt{4 \log 2 \cdot \xi(1) N} \right) \le \frac{1}{2}.
$$

By a union bound, the complement of this event and S simultaneously hold with probability at least

$\frac{1}{2} - Ke^{-cN}$. Thus,

$$\frac{1}{N}\log \mathbb{E}\,\mathbb{I}(S)\exp\beta\left(\mathcal{H}_N(\vec{\boldsymbol{\sigma}}^*) - \mathbb{E}\,\mathcal{H}_N(\vec{\boldsymbol{\sigma}}^*)\right) \geq -\beta K\sqrt{\frac{4\log 2\cdot\xi(1)}{N}} + \frac{1}{N}\log\left(\frac{1}{2} - Ke^{-cN}\right).$$

Putting this all together, we can choose a large C dependent only on ξ, h such that

$$F_N^{\mathrm{Sp}}(\beta, \mathcal{Q}(2\eta)) \geq \beta\mathsf{OPT}_N^{\mathrm{Sp}}(\mathcal{Q}(\eta)) - CK\beta\eta - CK\log\frac{1}{\eta} - \frac{CK\beta}{\sqrt{N}} - \frac{1}{N}\log\frac{1}{\frac{1}{2} - Ke^{-cN}}.$$

By choosing C large enough, we can ensure that if $N \geq C\log\max(K,2)$, then $Ke^{-cN} \leq \frac{1}{4}$. Then, we may absorb the last term into the term $CK\log\frac{1}{\eta}$. Rearranging yields the result. □

5.6 Overlap-Constrained Upper Bound on the Ising Grand Hamiltonian

In this section we upper-bound $\varphi^{\mathrm{Is}}(0)$. We take the reference measure μ to be the counting measure so that integrals over $\mathcal{Q}^{\mathrm{Is}}(\eta)$ become sums.

We define $(Z_0, \dots, Z_D)$ similarly to G_d of the previous section, but as a sum over all of $(\boldsymbol{\Sigma}_N)^K$ directly. As before, define $\vec{\boldsymbol{\eta}}_0, \dots \vec{\boldsymbol{\eta}}_D$ to be independent Gaussians as in (5.5.1) and (5.5.2). For $\vec{\boldsymbol{y}} \in (\mathbb{R}^K)^N$, define

$$\begin{aligned}
Z_D(\vec{\boldsymbol{y}}) &= \log \sum_{\vec{\boldsymbol{\sigma}}\in(\boldsymbol{\Sigma}_N)^K} \exp\sum_{u\in\mathbb{L}} \langle \boldsymbol{h} + \lambda\boldsymbol{m} + \boldsymbol{y}(u), \pi(\boldsymbol{\sigma}(u))\rangle \\
&= \log\prod_{i=1}^N\prod_{u\in\mathbb{L}}\left(2\cosh\left(h + \lambda m_i + \boldsymbol{y}(u)_i\right)\exp\left(-m_i(h+\lambda m_i + \boldsymbol{y}(u)_i)\right)\right) \\
&= \sum_{i=1}^N\sum_{u\in\mathbb{L}}\Big(\log\left(2\cosh\left(h+\lambda m_i + \boldsymbol{y}(u)_i\right)\right) - m_i(h+\lambda m_i + \boldsymbol{y}(u)_i)\Big).
\end{aligned}$$

Given the sequence $0 = \zeta_{-1} < \zeta_0 < \zeta_1 < \cdots < \zeta_L = 1$, recursively set

$$Z_d(\vec{\boldsymbol{y}}) = \frac{1}{\zeta_d}\,\mathbb{E}\,\zeta_d Z_{d+1}\left(\vec{\boldsymbol{y}} + \vec{\boldsymbol{\eta}}_{d+1}(\xi'(q_{d+1}) - \xi'(q_d))^{1/2}\right).$$

Then $Z_0 \equiv Z_0(0)$ is a deterministic function of $\boldsymbol{m}$ and h.

Proposition 5.6.1. *For any $\boldsymbol{m} \in [-1,1]^N$,*

$$\varphi^{\mathrm{Is}}(0) \leq \frac{1}{N}Z_0 + KR(\boldsymbol{h}, \boldsymbol{m}).$$

Proof. Recall from (5.4.8) that

$$\varphi^{\mathrm{Is}}(0) = KR(\boldsymbol{h},\boldsymbol{m}) + \frac{1}{N}\log \mathbb{E} \sum_{\alpha\in\mathbb{N}^D} \nu_\alpha \sum_{\vec{\boldsymbol{\sigma}}\in\mathcal{Q}^{\mathrm{Is}}(\eta)} \exp\left(\sum_{u\in\mathbb{L}} \langle \boldsymbol{h}+\lambda\boldsymbol{m}, \pi(\boldsymbol{\sigma}(u))\rangle + \sum_{u\in\mathbb{L}}\sum_{i=1}^{N} g^{(u)}_{\xi',i}(\alpha)\pi(\boldsymbol{\sigma}(u))_i \right).$$

Summing over all of $(\boldsymbol{\Sigma}_N)^K$ gives the upper bound

$$\varphi^{\mathrm{Is}}(0) \le KR(\boldsymbol{h},\boldsymbol{m}) + \frac{1}{N}\log \mathbb{E} \sum_{\alpha\in\mathbb{N}^D} \nu_\alpha \sum_{\vec{\boldsymbol{\sigma}}\in(\boldsymbol{\Sigma}_N)^K} \exp\left(\sum_{u\in\mathbb{L}} \langle \boldsymbol{h}+\lambda\boldsymbol{m}, \pi(\boldsymbol{\sigma}(u))\rangle + \sum_{u\in\mathbb{L}}\sum_{i=1}^{N} g^{(u)}_{\xi',i}(\alpha)\pi(\boldsymbol{\sigma}(u))_i \right).$$

Similarly to previous sections or as in [Pan13b, Theorem 2.9], properties of Ruelle cascades imply that the right hand side above equals

$$KR(\boldsymbol{h},\boldsymbol{m}) + \frac{1}{N} Z_0$$

because the coordinates $i \in [N]$ now decouple. □

5.6.1 Properties of Parisi PDEs

Here we review properties of Parisi PDEs. We begin with the 1-dimensional case for general $\zeta \in \mathscr{L}$ and consider the PDE

$$\partial_t \Phi_\zeta(t,x) + \frac{1}{2}\xi''(t)\left(\partial_{xx}\Phi_\zeta(t,x) + \zeta(t)(\partial_x\Phi_\zeta(t,x))^2\right) = 0 \tag{5.6.1}$$
$$\Phi_\zeta(1,x) = f_0(x).$$

For $\beta > 0$ we will consider the initial conditions $f_0(x) = \log(\cosh(\beta x)/\beta) - ax$ for $a = m_i \in [-1,1]$ which leads to solution $\Phi^\beta_{a,\zeta}$. When not specified, we take $\beta = 1$ and $a = 0$, so for instance $\Phi_{a,\zeta} = \Phi^1_{a,\zeta}$ and $\Phi^\beta_\zeta = \Phi^\beta_{0,\zeta}$. We also allow the $\beta = \infty$ case $\Phi^\infty_{a,\zeta}$ corresponding to $f_0(x) = |x| - ax$. Note that (5.1.4) corresponds to the case $(a,\beta) = (0,\infty)$. Regularity properties for solutions to (5.6.1) were derived in several works such as [JT16, Che17] for $\zeta \in \mathscr{U}$. For $\zeta \in \mathscr{L}$ they are given in Chapter B. The following result follows from Proposition B.1.1 part (b) and Lemma B.1.4.

Proposition 5.6.2. *For $\zeta \in \mathscr{L}$ and $(a,\beta) \in [-1,1]\times(0,\infty]$, the function $\Phi^\beta_{a,\zeta}$ is continuous on $[0,1]\times\mathbb{R}$ and 2-Lipschitz in x. Moreover both*

$$\partial_{xx}\Phi^\beta_{a,\zeta}(t,x) \quad \text{and} \quad \partial_t\Phi^\beta_{a,\zeta}(t,x)$$

are uniformly bounded on $(t,x) \in [0,1-\varepsilon]\times\mathbb{R}$ for any $\varepsilon > 0$. Finally $\Phi^\beta_{a,\zeta}(t,x)$ is convex in x.

The following result is shown in Lemma B.1.5.

Proposition 5.6.3. *For $\zeta \in \mathscr{L}$ the SDE*

$$\mathrm{d}X_t = \xi''(t)\zeta(t)\partial_x \Phi^{\beta}_{a,\zeta}(t, X_t)\mathrm{d}t + \sqrt{\xi''(t)}\mathrm{d}B_t, \quad X_0 = X_0 \tag{5.6.2}$$

has strong and pathwise unique solution.

Finally the next result follows from Proposition B.1.1 part (c).

Proposition 5.6.4. *For $\zeta_1, \zeta_2 \in \mathscr{L}$, and $\beta \in (0, \infty]$,*

$$|\Phi^{\beta}_{\zeta_1} - \Phi^{\beta}_{\zeta_2}| \le \int_0^1 \xi''(t)|\zeta_1(t) - \zeta_2(t)|dt.$$

The Multi-Dimensional Parisi PDE

Here we define the Parisi PDE on $\mathbb{R}^K$. For simplicity we restrict attention to finitely supported $\zeta \in \mathcal{M}_{\vec{q}}$. We construct $\Phi^{\mathbb{L}}$ via the Hopf-Cole transformation and verify that it solves a version of (5.6.1).

Recall the definition of $M^d = M^{\vec{k},\vec{p},d} \in \mathbb{R}^{K\times K}$ given by

$$M^{\vec{k},\vec{p},d}_{u^1,u^2} = \mathbb{I}\{u^1 \wedge u^2 \ge d\} p_{u^1 \wedge u^2}.$$

As before, $M(t) = M^d$ for $t \in [q_{d-1}, q_d)$.

For an atomic measure $\zeta \in \mathcal{M}_{\vec{q}}$ consider the function $\Phi^{\mathbb{L}}_{\zeta}(t, \vec{x}) : [0,1] \times \mathbb{R}^K \to \mathbb{R}$ defined as as follows. The $t = 1$ boundary condition is

$$\Phi^{\mathbb{L}}_{a,\zeta}(1, \vec{x}) = \sum_{u\in\mathbb{L}} \log\left(2\cosh x(u)\right) - ax(u).$$

For $t \in [q_0, 1)$, $\Phi^{\mathbb{L}}_{a,\zeta}$ is defined recursively by

$$\Phi^{\mathbb{L}}_{a,\zeta}(t,x) = \frac{1}{\zeta(t)} \log \mathbb{E} \exp\left(\zeta(t)\Phi^{\mathbb{L}}_{a,\zeta}(q_{d+1}, \vec{x} + \vec{\eta}_{d+1} \cdot (\xi'(q_{d+1}) - \xi'(t))^{1/2})\right), \quad t \in [q_d, q_{d+1})$$

where $\vec{\eta}_0 \sim \mathcal{N}(0, M^1)$ and $\vec{\eta}_d \sim \mathcal{N}(0, M^d)$ for $1 \le d \le D$ are independent Gaussian vectors in $\mathbb{R}^K$. For $t \in [0, q_0)$, we extend the definition of ζ so that $\zeta(t) = 0$ and define

$$\Phi^{\mathbb{L}}_{a,\zeta}(t,x) = \mathbb{E}\, \Phi^{\mathbb{L}}_{a,\zeta}(q_0, \vec{x} + \vec{\eta}_0 \cdot (\xi'(q_0) - \xi'(t))^{1/2}).$$

Proposition 5.6.5. *For any $\zeta \in \mathcal{M}_{\vec{q}}$,*

$$Z_0 = \frac{1}{N}\sum_{i=1}^{N} \Phi^{\mathbb{L}}_{m_i,\zeta}(0,(h+\lambda m_i)\vec{1}).$$

Proof. This follows from Lemma 5.6.6 since the recursive definition of $\Phi^{\mathbb{L}}_{a,\zeta}(t,x)$ restricted to times $t \in \{q_d\}_{d\in[D]}$ is exactly that of Z_0 up to an spatial shift of $(h+\lambda m_i)\vec{1}$. □

We defer the proof of the next lemma, which is a standard computation.

Lemma 5.6.6. *The function $\Phi^{\mathbb{L}}_{a,\zeta}$ is smooth on each time interval $[q_d, q_{d+1}] \times \mathbb{R}^K$. Moreover it is continuous and solves the K-dimensional Parisi PDE*

$$\partial_t \Phi^{\mathbb{L}}_{a,\zeta}(t,\vec{x}) = -\frac{\xi''(t)}{2}\left(\langle M(t), \nabla^2 \Phi^{\mathbb{L}}_{a,\zeta}\rangle + \zeta(t)\langle M(t), (\nabla \Phi^{\mathbb{L}}_{a,\zeta})^{\otimes 2}\rangle\right). \tag{5.6.3}$$

Finally $|\partial_{x(u)}\Phi^{\mathbb{L}}_{a,\zeta}(t,\vec{x})| \le 1+|a|$ holds for all $(t,\vec{x},u) \in [0,1]\times\mathbb{R}^K\times\mathbb{L}$.

Auffinger-Chen Representation

As shown by [AC15] the Parisi PDE admits a stochastic control formulation. We now recall such representations in the cases of interest starting with the 1-dimensional case. For $0 \le t_1 \le t_2 \le 1$ let $\mathcal{D}[t_1,t_2]$ be the space of processes $v \in C([t_1,t_2];\mathbb{R})$ with $\sup_{t_1\le r\le t_2}|v_r| \le 2$ which are progressively measurable with respect to filtration supporting a standard Brownian motion B_t. Define the functional

$$\mathcal{X}^{t_1,t_2}_{a,\zeta}(x,v) = \mathbb{E}\left[\mathcal{Y}^{t_1,t_2}_{a,\zeta}(x,v) - \mathcal{Z}^{t_1,t_2}_{a,\zeta}(v)\right]$$

where

$$\mathcal{Y}^{t_1,t_2}_{a,\zeta}(x,v) \equiv \Phi^{\beta}_{a,\zeta}\left(t_2, x + \int_{t_1}^{t_2}\zeta(r)\xi''(r)v_r\mathrm{d}r + \int_{t_1}^{t_2}\sqrt{\xi''(r)}\mathrm{d}B_r\right),$$
$$\mathcal{Z}^{t_1,t_2}_{a,\zeta}(v) \equiv \frac{1}{2}\int_{t_1}^{t_2}\zeta(r)\xi''(r)v_r^2\mathrm{d}r.$$

Note that since $|v_r| \le 2$ is uniformly bounded and $||\xi''\cdot\zeta||_1 < \infty$ there are no continuity issues near $t=1$. The next proposition, whose standard proof we defer, relates $\Phi^{\beta}_{a,\zeta}$ to stochastic control.

Proposition 5.6.7. *For any $\zeta \in \mathscr{L}$, $[t_1,t_2] \subseteq [0,1]$, $a \in [-1,1]$ and $\beta \in (0,\infty]$, the function $\Phi^{\beta}_{a,\zeta}$ satisfies*

$$\Phi^{\beta}_{a,\zeta}(t_1,x) = \sup_{v\in\mathcal{D}[t_1,t_2]} \mathcal{X}^{t_1,t_2}_{a,\zeta}(x,v). \tag{5.6.4}$$

Moreover the maximum in (5.6.4) *is achieved by*

$$v_r = \partial_x \Phi^{\beta}_{a,\zeta}(r, X_r)$$

where X_r *solves the SDE* (5.6.2) *with initial condition* $X_{t_1} = x$.

The corresponding stochastic control formulation in $\mathbb{R}^K$ is as follows. For $0 \le t_1 \le t_2 \le 1$ let $\mathcal{D}^{\mathbb{L}}[t_1, t_2]$ be the space of processes $\vec{v} \in C([t_1, t_2]; \mathbb{R}^K)$ with $\sup_{t_1 \le r \le t_2} |\vec{v}_r|_\infty \le 2$ which are progressively measurable with respect to a filtration supporting an $\mathbb{R}^K$ valued Brownian motion $\vec{B}_r = (B^u_r)_{u \in \mathbb{L}}$. Define the functional

$$\mathcal{X}^{\mathbb{L},t_1,t_2}_{a,\zeta}(\vec{x}, \vec{v}) \equiv \mathbb{E}\left[\mathcal{Y}^{\mathbb{L},t_1,t_2}_{a,\zeta}(\vec{x}, \vec{v}) - \mathcal{Z}^{\mathbb{L},t_1,t_2}_{a,\zeta}(\vec{v})\right]$$

where

$$\mathcal{Y}^{\mathbb{L},t_1,t_2}_{a,\zeta}(\vec{x}, \vec{v}) \equiv \Phi^{\mathbb{L}}_{a,\zeta}\left(t_2, \vec{x} + \int_{t_1}^{t_2} \zeta(r)\xi''(r)M(r)\vec{v}_r \mathrm{d}r + \int_{t_1}^{t_2} \sqrt{\xi''(r)M(r)}\mathrm{d}\vec{B}_r\right),$$
$$\mathcal{Z}^{\mathbb{L},t_1,t_2}_{a,\zeta}(\vec{v}) \equiv \frac{1}{2}\int_{t_1}^{t_2} \zeta(r)\xi''(r)\langle M(r), \vec{v}_r^{\otimes 2}\rangle \mathrm{d}r.$$

In the multi-dimensional case we restrict attention to finitely supported $\zeta \in \mathcal{M}_{\vec{q}}$ to avoid the by-now routine process of extending regularity properties of $\Phi^{\mathbb{L}}_{\zeta}$ to general ζ. The proof is again deferred.

Proposition 5.6.8. *For any* $\zeta \in \mathcal{M}_{\vec{q}}$, $[t_1, t_2] \subseteq [0, 1]$ *and* $a \in [-1, 1]$, *the function* $\Phi^{\mathbb{L}}_{a,\zeta}$ *satisfies*

$$\Phi^{\mathbb{L}}_{a,\zeta}(t_1, \vec{x}) = \sup_{\vec{v} \in \mathcal{D}^{\mathbb{L}}[t_1,t_2]} \mathcal{X}^{\mathbb{L},t_1,t_2}_{a,\zeta}(\vec{x}, \vec{v}). \tag{5.6.5}$$

Moreover (5.6.5) *is maximized by* $\vec{v}_s = \nabla \Phi^{\mathbb{L}}_{a,\zeta}(s, \vec{X}_s)$ *where the* $\mathbb{R}^K$*-valued process* $\vec{X}_s$ *solves*

$$\vec{X}_s = \vec{x} + \int_{t_1}^{s} \zeta(r)\xi''(r)M(r)\nabla \Phi^{\mathbb{L}}_{a,\zeta}(r, \vec{X}_r)\mathrm{d}r + \int_{t_1}^{s} \sqrt{\xi''(r)M(r)}\mathrm{d}\vec{B}_r, \qquad s \in [t_1, t_2].$$

5.6.2 Relations Among Parisi PDEs

Following [CPS22, Section 8] we relate $\Phi_{a,\zeta}$ to Φ_ζ. Note that we always consider times $t \in [0, 1]$ with endpoint conditions at $t = 1$, while [CPS22] defines the boundary condition for $\Phi_{a,\zeta}$ at time $t = 1 - q_0$, see e.g. Equation (3.25) therein.

Proposition 5.6.9. *For any $a \in [-1,1]$ and $\zeta \in \mathscr{L}$, with $y = x - a\int_0^1 \xi''(t)\zeta(t)dt$,*

$$\Phi_\zeta(0,y) - ay = \Phi_{a,\zeta}(0,x) + \frac{a^2}{2}\int_0^1 \xi''(t)\zeta(t)dt.$$

Proof. By setting $y = x - a\int_0^1 \xi''(s)\zeta(s)\mathrm{d}s$, it suffices to show that for all $t \in [0,1]$,

$$\Phi_{a,\zeta}(t,x) = \Phi_\zeta\left(t, x - a\int_t^1 \xi''(s)\zeta(s)\mathrm{d}s\right) - ax + \frac{a^2}{2}\int_t^1 \xi''(s)\zeta(s)\mathrm{d}s.$$

(In particular the desired result is obtained by setting $t = 0$.) It suffices to show this for ζ continuous. Set

$$f(t,x) \equiv \Phi_\zeta\left(t, x - a\int_t^1 \xi''(s)\zeta(s)\mathrm{d}s\right) - ax + \frac{a^2}{2}\int_t^1 \xi''(s)\zeta(s)\mathrm{d}s$$

and define

$$b(t,x) \equiv x - a\int_t^1 \xi''(s)\zeta(s)\mathrm{d}s.$$

Then we compute

$$\partial_t f(t,x) = \partial_t\Phi_\zeta(t,b(t,x)) + a\xi''(t)\zeta(t)\partial_x\Phi_\zeta(t,b(t,x)) - \frac{a^2}{2}\xi''(t)\zeta(t)$$

and

$$\begin{aligned}\partial_x f(t,x) &= \partial_x\Phi_\zeta(t,b(t,x)) - a,\\ \partial_{xx} f(t,x) &= \partial_{xx}\Phi_\zeta(t,b(t,x)).\end{aligned}$$

It follows that

$$\partial_t f(t,x) = -\frac{\xi''(t)}{2}\left(\partial_{xx} f(t,x) + \zeta(t)\left(\partial_x f(t,x)\right)^2\right).$$

Note that at time 1, $f(1,x) = \log(2\cosh(x)) - ax = \Phi_{a,\zeta}(1,x)$. Uniqueness of solutions to the Parisi PDE as in [JT16, Lemma 13] completes the proof.

□

Lemma 5.6.10. *For any $\zeta, \gamma \in \mathscr{L}$ and any $(t,x,\beta) \in [0,1]\times\mathbb{R}\times(0,\infty]$,*

$$\Phi^\beta_{a,\zeta}(t,x) \le \Phi^\beta_{a,\zeta+\gamma}(t,x).$$

Proof. We use the Auffinger-Chen representation (5.6.4) for $\Phi^\beta_{a,\zeta}$ and $\Phi^\beta_{a,\zeta+\gamma}$. For any control v, consider the modified control

$$w_t \equiv \frac{\zeta(t)v_t}{\zeta(t)+\gamma(t)}.$$

It is not difficult to see that

$$\mathcal{Y}^{t,1}_{a,\zeta}(x,v) = \mathcal{Y}^{t,1}_{a,\zeta+\gamma}(x,w)$$

since the resulting SDE is the same, while

$$\mathcal{Z}^{t,1}_{a,\zeta}(v) \geq \mathcal{Z}^{t,1}_{a,\zeta+\gamma}(w).$$

Therefore

$$\mathcal{X}^{t,1}_{a,\zeta}(x,v) \leq \mathcal{X}^{t,1}_{a,\zeta+\gamma}(x,w)$$

Since v was arbitrary, we are done by Proposition 5.6.7. □

Define $\overline{\zeta} = \zeta|_{[q_0,1]}$ and $\underline{\zeta} = \zeta|_{[0,q_0]}$ when $\zeta \in \mathscr{L}$ and $q_0 \in [0,1]$ are given. The next lemma is analogous to Lemma 5.5.8 and will be used to connect our estimates for $\varphi(0)$ to the Parisi functional uniformly in $\boldsymbol{m}$.

Lemma 5.6.11. *For $\zeta \in \mathscr{L}$, with $\lambda = \int_0^1 \xi''(t)\zeta(t)dt$,*

$$\frac{1}{N}\sum_{i=1}^N \Phi^\infty_{m_i,\overline{\zeta}}(0, h+\lambda m_i) - \frac{1}{2}\int_{q_0}^1 (t-q_0)\overline{\zeta}(t)\xi''(t)dt + R(\boldsymbol{h},\boldsymbol{m}) \leq \mathsf{P}^{\mathrm{Is}}_{\xi,h}(\zeta).$$

Proof. Define the constants

$$\begin{aligned}
I &= \int_0^1 t\xi''(t)\zeta(t)dt, \quad J = \lambda = \int_0^1 \xi''(t)\zeta(t)dt,\\
\overline{I} &= \int_{q_0}^1 t\xi''(t)\overline{\zeta}(t)dt, \quad \overline{J} = \int_{q_0}^1 \xi''(t)\overline{\zeta}(t)dt,\\
\underline{I} &= \int_0^{q_0} t\xi''(t)\underline{\zeta}(t)dt, \quad \underline{J} = \int_0^{q_0} \xi''(t)\underline{\zeta}(t)dt.
\end{aligned}$$

Then $I = \overline{I} + \underline{I}$ and $J = \overline{J} + \underline{J}$ and $q_0\underline{J} \geq \underline{I}$. Recalling that $\mathsf{P}^{\mathrm{Is}}_{\xi,h}(\zeta) = \Phi^\infty_\zeta(0,h) - \frac{I}{2}$, we estimate

$$\begin{aligned}
\mathsf{P}^{\mathrm{Is}}_{\xi,h}(\zeta) &= \Phi^\infty_\zeta(0,h) - \tfrac{I}{2} && \\
&= \tfrac{1}{N}\textstyle\sum_{i=1}^N \left(\Phi^\infty_{m_i,\zeta}(0,h+\lambda m_i) + \frac{m_i^2 J}{2}\right) - \frac{I}{2} + R(\boldsymbol{h},\boldsymbol{m}) && \text{Prop 5.6.2}\\
&= \tfrac{1}{N}\textstyle\sum_{i=1}^N \Phi^\infty_{m_i,\zeta}(0,h+\lambda m_i) + \frac{q_0 J}{2} - \frac{I}{2} + R(\boldsymbol{h},\boldsymbol{m}) && \|\boldsymbol{m}\|_N^2 = q_0\\
&= \tfrac{1}{N}\textstyle\sum_{i=1}^N \Phi^\infty_{m_i,\zeta}(0,h+\lambda m_i) + \frac{q_0\overline{J}}{2} - \frac{\overline{I}}{2} + \frac{q_0\underline{J}-\underline{I}}{2} + R(\boldsymbol{h},\boldsymbol{m}) && \zeta = \overline{\zeta}+\underline{\zeta}\\
&\geq \tfrac{1}{N}\textstyle\sum_{i=1}^N \Phi^\infty_{m_i,\zeta}(0,h+\lambda m_i) + \frac{q_0\overline{J}}{2} - \frac{\overline{I}}{2} + R(\boldsymbol{h},\boldsymbol{m}) && q_0\underline{J} \geq \underline{I}\\
&\geq \tfrac{1}{N}\textstyle\sum_{i=1}^N \Phi^\infty_{m_i,\overline{\zeta}}(0,h+\lambda m_i) + \frac{q_0\overline{J}}{2} - \frac{\overline{I}}{2} + R(\boldsymbol{h},\boldsymbol{m}). && \text{Lem 5.6.10}
\end{aligned}$$

This is exactly what we wanted to show.

□

The next crucial lemma upper-bounds $\Phi^{\mathbb{L}}_{a,\zeta}$ using the 1-dimensional function $\Phi_{a,\kappa\zeta}$. As in the spherical case, multiplying by κ will allow us to pass from increasing $\zeta \in \mathcal{M}_{\vec{q}}$ to arbitrary functions in $\mathscr{L}$.

Lemma 5.6.12. *For any $\zeta \in \mathcal{M}_{\vec{q}}$, $\vec{x} \in \mathbb{R}^K$, $a \in [-1,1]$ and $t \in [0,1]$,*

$$\Phi^{\mathbb{L}}_{a,\zeta}(t,\vec{x}) \le \sum_{u\in\mathbb{L}} \Phi_{a,\kappa\zeta}(t, x(u)). \tag{5.6.6}$$

Proof. Define

$$\widetilde{\mathcal{Z}}^{\mathbb{L},t_1,t_2}_{a,\zeta}(\vec{v}) \equiv \frac{1}{2}\int_{t_1}^{t_2} \zeta(r)\xi''(r)\kappa(r)^{-1}\langle M(r)^2, \vec{v}_r^{\otimes 2}\rangle \mathrm{d}r.$$

Since $M(r) \preceq \kappa(r) I_K$, in the Loewner order, it follows that

$$\kappa(r)^{-1} M(r)^2 \preceq M(r).$$

Hence $\widetilde{\mathcal{Z}}^{\mathbb{L},t_1,t_2}_{a,\zeta}(\vec{v}) \le \mathcal{Z}^{\mathbb{L},t_1,t_2}_{a,\zeta}(\vec{v})$ for any $\vec{x}$ and $\vec{v}$. Setting

$$\widetilde{\mathcal{X}}^{\mathbb{L},t_1,t_2}_{a,\zeta}(\vec{x},\vec{v}) \equiv \mathcal{Y}^{\mathbb{L},t_1,t_2}_{a,\zeta}(\vec{x},\vec{v}) - \widetilde{\mathcal{Z}}^{\mathbb{L},t_1,t_2}_{a,\zeta}(\vec{v}),$$

it follows that

$$\mathcal{X}^{t,1}_{a,\zeta}(\vec{x},\vec{v}) \le \widetilde{\mathcal{X}}^{t,1}_{a,\zeta}(\vec{x},\vec{v})$$

always holds. Next for any $\vec{v} \in \mathcal{D}^{\mathbb{L}}[t,1]$ and $r \in [t,1]$, define $\vec{V}_r = \frac{M(r)\vec{v}_r}{\kappa(r)} \in \mathbb{R}^K$. Then

$$\langle M(r)^2, \vec{v}_r^{\otimes 2}\rangle = \|M(r)\vec{v}_r\|_2^2 = \kappa(r)^2 \left\|\vec{V}_r\right\|^2$$

and so (including the relevant Brownian motions as arguments in a slight abuse of notation),

$$\widetilde{\mathcal{Z}}^{\mathbb{L},t,1}_{a,\zeta}(\vec{v},\vec{B}) = \sum_{u\in\mathbb{L}} \mathcal{Z}^{t,1}_{a,\kappa\zeta}(V(u), B(u)).$$

Moreover since $\kappa(r)\vec{V}_r(u) = M(r)\vec{v}_r(u)$,

$$\widetilde{\mathcal{Y}}^{\mathbb{L},t,1}_{a,\zeta}(\vec{x},\vec{v},\vec{B}) = \sum_{u\in\mathbb{L}} \mathcal{Y}^{t,1}_{a,\kappa\zeta}(x(u), V(u), B(u)).$$

Since each coordinate $B_r(u)$ of $\vec{B}_r$ has the marginal law of a 1-dimensional Brownian motion,

$$\widetilde{\mathcal{X}}^{\mathbb{L},t,1}_{a,\zeta}(\vec{x},\vec{v}) = \sum_{u\in\mathbb{L}} \mathcal{X}^{t,1}_{a,\kappa\zeta}(x(u), V(u)).$$

Therefore we obtain

$$\begin{aligned}\mathcal{X}^{t,1}_{a,\zeta}(\vec{x},\vec{v}) &\le \widetilde{\mathcal{X}}^{t,1}_{a,\zeta}(\vec{x},\vec{v})\\ &= \sum_{u\in\mathbb{L}} \mathcal{X}^{t,1}_{a,\kappa\zeta}(x(u),V(u))\\ &\le \sum_{u\in\mathbb{L}} \Phi_{a,\kappa\zeta}(t,x(u)).\end{aligned}$$

Since $\vec{v}\in\mathcal{D}^{\mathbb{L},t,1}$ was arbitrary this concludes the proof. □

5.6.3 Zero Temperature Limit

We now apply the above results with $(\beta^2\xi,\beta\boldsymbol{h},\beta\lambda)$ in place of $(\xi,\boldsymbol{h},\lambda)$, which corresponds to scaling $\mathcal{H}_N$ to $\beta\mathcal{H}_N$. We accordingly define $\Phi_{\beta^2\xi,\zeta}$ and $\varphi^{\mathrm{Is}}_{\beta^2\xi}(0)$ by making this substitution in their definitions. It is not hard to derive the scaling relation

$$\Phi_{\beta^2\xi,\zeta}(t,\beta x) = \beta\cdot\Phi^{\beta}_{\beta\zeta}(t,x),\quad (t,x)\in[0,1]\times\mathbb{R} \tag{5.6.7}$$

for any $\beta\in(0,\infty)$ and $\zeta\in\mathscr{L}$.

We will also use the following simple estimate to pass to the zero temperature limit.

Proposition 5.6.13. $\sup_\zeta\left|\Phi^{\beta}_{\zeta}(t,x)-\Phi^{\infty}_{\zeta}(t,x)\right|\le\frac{\log 2}{\beta}$.

Proof. Recall that $\Phi^{\beta}_{\zeta}(1,x)$ is convex and 1-Lipschitz while $\Phi^{\infty}_{\zeta}(1,x)=|x|\le\Phi^{\beta}_{\zeta}(1,x)$. It follows that

$$\begin{aligned}\sup_{x\in\mathbb{R}}\left|\Phi^{\beta}_{\zeta}(1,x)-\Phi^{\infty}_{\zeta}(1,x)\right| &= \left|\Phi^{\beta}_{\zeta}(1,0)-\Phi^{\infty}_{\zeta}(1,0)\right|\\ &= \frac{\log 2}{\beta}.\end{aligned}$$

Hence

$$\left|\mathcal{X}^{0,1}_{\zeta,\beta}(v,x)-\mathcal{X}^{0,1}_{\zeta,\infty}(v,x)\right|\le\frac{\log 2}{\beta}$$

holds for any control v, since the only difference is from the boundary value at time $t=1$ in $\mathcal{Y}$. Proposition 5.6.7 now implies the desired result. □

Below, recall the definition $\overline{\zeta}=\zeta|_{[q_0,1]}$.

Lemma 5.6.14. *Let $(\vec{p},\vec{q},\vec{k})$ be as in Section 5.4, and fix $\beta>0$ and $\zeta\in\mathscr{L}$ such that $\overline{\zeta}\in\mathcal{M}_{\vec{q}}$. With*

$$\lambda=\int_0^1\xi''(t)\kappa(t)\zeta(t)\mathrm{d}t$$

we have

$$F_N^{\mathrm{Is}}(\beta, \mathcal{Q}(\eta)) \le \beta K \mathsf{P}^{\mathrm{Is}}(\beta\kappa\zeta) + 3\beta^2 K^2 \xi''(1)\eta + K\beta\lambda\eta + K\log 2.$$

Proof. Applying Proposition 5.4.1 with $\bar{\zeta}$ and $(\beta^2\xi, \beta\boldsymbol{h}, \beta\lambda)$ in place of $(\xi, \boldsymbol{h}, \lambda)$ in the first line,

$$\begin{aligned}
F_N^{\mathrm{Is}}(\beta,\mathcal{Q}(\eta)) - 3\beta^2K^2\xi''(1)\eta - K\beta\lambda\eta &\le \varphi^{\mathrm{Is}}_{\beta^2\xi}(0) - \tfrac{\beta^2K}{2}\textstyle\int_{q_0}^1 (q-q_0)\xi''(q)\kappa(q)\zeta(q)\,\mathrm{d}q && \text{Props 5.6.1,5.6.5}\\
&\le \tfrac{1}{N}\textstyle\sum_{i=1}^N \Phi^{\mathbb{L}}_{\beta^2\xi,m_i,\bar{\zeta}}(q_0, \beta h + \beta\lambda m_i) - \tfrac{\beta^2K}{2}\int_{q_0}^1 (q-q_0)\xi''(q)\kappa(q)\zeta(q)\,\mathrm{d}q + \beta K R(\boldsymbol{h},\boldsymbol{m}) && \text{Lem 5.6.12}\\
&\le \tfrac{K}{N}\textstyle\sum_{i=1}^N \Phi_{\beta^2\xi,m_i,\kappa\bar{\zeta}}(q_0, \beta h + \beta\lambda m_i) - \tfrac{\beta^2K}{2}\int_{q_0}^1 (q-q_0)\xi''(q)\kappa(q)\zeta(q)\,\mathrm{d}q + \beta K R(\boldsymbol{h},\boldsymbol{m}) && (5.6.7)\\
&= \tfrac{\beta K}{N}\textstyle\sum_{i=1}^N \Phi^{\beta}_{m_i,\beta\kappa\bar{\zeta}}(q_0, h + \lambda m_i) - \tfrac{\beta^2K}{2}\int_{q_0}^1 (q-q_0)\xi''(q)\kappa(q)\zeta(q)\,\mathrm{d}q + \beta K R(\boldsymbol{h},\boldsymbol{m}) && \text{Prop 5.6.13}\\
&= \tfrac{\beta K}{N}\textstyle\sum_{i=1}^N \Phi^{\infty}_{m_i,\beta\kappa\bar{\zeta}}(q_0, h + \lambda m_i) - \tfrac{\beta K}{2}\int_{q_0}^1 (q-q_0)\xi''(q)\beta\kappa(q)\zeta(q)\,\mathrm{d}q\\
&\quad + \beta K R(\boldsymbol{h},\boldsymbol{m}) + K\log 2 && \text{Lem 5.6.11}\\
&\le \beta K \mathsf{P}^{\mathrm{Is}}(\beta\kappa\zeta) + K\log 2.
\end{aligned}$$

Here terms modified from the previous line are in red text.

□

All that remains is to approximate an arbitrary $\zeta_* \in \mathscr{L}$ by $\beta\kappa\zeta$ on $[q_0, 1)$ for $\zeta \in \mathcal{M}_{\vec{q}}$ and choose parameters appropriately. We do this now.

Proof of Proposition 5.3.2, Ising case. First choose $\zeta_* = \zeta_*(\xi, h, \varepsilon) \in \mathscr{L}$ such that

$$\mathsf{P}^{\mathrm{Is}}(\zeta_*) \le \inf_{\zeta\in\mathscr{L}} \mathsf{P}^{\mathrm{Is}}(\zeta) + \frac{\varepsilon}{10} = \mathsf{ALG}^{\mathrm{Is}} + \frac{\varepsilon}{10}. \tag{5.6.8}$$

Since $\int_0^1 \xi''(t)\zeta_*(t)\mathrm{d}t < \infty$, the monotone convergence theorem guarantees

$$\lim_{\beta\to\infty} \int_0^1 \xi''(t)\cdot|\min(\zeta(t),\beta) - \zeta(t)|\,\mathrm{d}t = 0.$$

Define $\zeta_\beta(t) = \min(\zeta(t), \beta)$. Therefore there exists

$$\beta = \beta(\zeta_*, \xi, \varepsilon) = \beta(\xi, h, \varepsilon) \ge \frac{20\log 2}{\varepsilon} \tag{5.6.9}$$

sufficiently large so that (recall Proposition 5.6.4)

$$\mathsf{P}^{\mathrm{Is}}(\zeta_\beta) - \mathsf{P}^{\mathrm{Is}}(\zeta_*) \le 2\int_0^1 \xi''(t)\cdot|\zeta_\beta(t) - \zeta(t)|\,\mathrm{d}t \le \frac{\varepsilon}{10}. \tag{5.6.10}$$

For $\delta > 0$, let $q_0^\delta = q_0$ and $q_\delta^{d+1} = \min(q_d^\delta + \delta, 1)$. This determines D which satisfies $q_{D-1} < q_D = 1$. Since $\zeta_\beta \in \mathscr{L}$ is bounded and has bounded variation, there exists $\delta = \delta(\xi, \zeta_\beta, \varepsilon) = \delta(\xi, \boldsymbol{h}, \varepsilon) > 0$

such that the function

$$\zeta_{\beta,\delta}(t) = \begin{cases} \zeta_\beta(t), & t \in [0, q_0) \\ \max\left(\delta, \zeta_\beta(q_j^\delta)\right), & t \in [q_j^\delta, q_{j+1}^\delta), j \geq 0 \end{cases}$$

satisfies

$$\mathsf{P}^{\mathrm{Is}}(\zeta_\beta) - \mathsf{P}^{\mathrm{Is}}(\zeta_{\beta,\delta}) \leq 2\int_0^1 \xi''(t)\,|\zeta_\beta(t) - \zeta_{\beta,\delta}(t)|\,\mathrm{d}t \leq \frac{\varepsilon}{10}. \tag{5.6.11}$$

(Note in particular that δ does not depend on q_0.) Observe that $\zeta_{\beta,\delta}(t) \in [\delta, \beta]$ holds for all $t \in [0,1]$. Next define

$$k_1 = k_2 = \cdots = k_D = k_* \equiv \left\lceil \frac{\beta}{\delta^2} \right\rceil.$$

This leads to $p_d^\delta = \chi^{-1}(q_d^\delta)$ with $\delta \leq p_1^\delta \leq p_D^\delta = 1$ and hence $\kappa(t) = \kappa_d$ for $t \in [q_d^\delta, q_{d+1}^\delta)$, where

$$\delta k_*^{D-d} \leq \kappa_d \leq k_*^{D-d}.$$

Next define

$$\widehat{\zeta}_{\beta,\delta}(t) \equiv \frac{\zeta_{\beta,\delta}(t)}{\beta\kappa(t)}, \quad t \in [q_0, 1]$$

so that $\beta\kappa\widehat{\zeta}_{\beta,\delta} = \zeta_{\beta,\delta}$. Note that

$$\sup_{t\in[0,1]} \widehat{\zeta}_{\beta,\delta}(t) \leq \frac{\sup_{s\in[0,1]} \zeta_{\beta,\delta}(s)}{\beta} \leq 1.$$

Additionally $\widehat{\zeta}_{\beta,\delta}$ is nondecreasing since if $q_d \leq t_d < q_{d+1}^\delta \leq t_{d+1} \leq q_{d+2}^\delta$, then

$$\begin{aligned} \frac{\widehat{\zeta}_{\beta,\delta}(t_d)}{\widehat{\zeta}_{\beta,\delta}(t_{d+1})} &= \frac{\zeta_{\beta,\delta}(t_d)}{\zeta_{\beta,\delta}(t_d)} \cdot \frac{\kappa_{d+1}}{\kappa_d} \\ &\leq \frac{\beta}{\delta^2 k_*} \leq 1 \end{aligned}$$

by definition of k_*. Set

$$\lambda = \int_0^1 \xi''(t)\kappa(t)\widehat{\zeta}_{\beta,\delta}(t)\mathrm{d}t$$

and

$$\eta = \frac{\varepsilon}{30\beta K\xi''(1) + 10\lambda}. \tag{5.6.12}$$

We now show that using $\widehat{\zeta}_{\beta,\delta}$ in the interpolation implies Proposition 5.3.2. Take $\vec{p}, \vec{q}, \vec{k}, D, \beta, \eta$ as

above.

$$\begin{aligned}
\tfrac{1}{N}\,\mathbb{E}\left[\max_{\vec{\boldsymbol{\sigma}}\in\mathcal{Q}^{\mathrm{Is}}(\eta)}\mathcal{H}_N^{\vec{k},\vec{p}}(\vec{\boldsymbol{\sigma}})\right] &\le F_N^{\mathrm{Is}}(\beta,\mathcal{Q}(\eta))/\beta \\
&\le K\mathsf{P}^{\mathrm{Is}}(\beta\kappa\widehat{\zeta}_{\beta,\delta}) + 3\beta K^2\xi''(1)\eta + K\lambda\eta + \tfrac{K\log 2}{\beta} && \big\downarrow \textit{Lem 5.6.14} \\
&\le K\cdot\mathsf{P}^{\mathrm{Is}}(\zeta_{\beta,\delta}) + \tfrac{2K\varepsilon}{10} && \big\downarrow (5.6.9),\ (5.6.12) \\
&\le K\cdot\mathsf{P}^{\mathrm{Is}}(\zeta_{\beta}) + \tfrac{3K\varepsilon}{10} && \big\downarrow (5.6.11) \\
&\le K\cdot\mathsf{P}^{\mathrm{Is}}(\zeta_{*}) + \tfrac{4K\varepsilon}{10} && \big\downarrow (5.6.10) \\
&\le K\cdot\mathsf{ALG}^{\mathrm{Is}} + \tfrac{5K\varepsilon}{10}. && \big\downarrow (5.6.8)
\end{aligned}$$

Moreover the values D, η and K above are bounded depending only on ξ, h and ε. Indeed $D \le \delta^{-1}+1$, η is bounded as in (5.6.12), and $K = \prod_{d=1}^{D} k_i = k_*^D = \left\lceil \frac{\beta}{\delta^2}\right\rceil^D$. Meanwhile β as defined in (5.6.9) also depends only on ξ, h, ε. This concludes the proof.

□

5.6.4 Deferred Proofs

Here we give the missing proofs for this section, which are all relatively standard.

Proof of Lemma 5.6.6. We assume $\zeta(t) > 0$ as the $\zeta(t) = 0$ case is clear. We consider only the case $t \in [q_{D-1}, 1)$ as the remaining cases are identical by induction. Let $\vec{y} = \vec{y}(t) \in \mathbb{R}^K$ be the Gaussian random vector

$$\vec{y} = \vec{\eta}_D(\xi''(1) - \xi''(t))^{1/2}.$$

Below A always denotes

$$A(\vec{x}+\vec{y}) = \Phi_{a,\zeta}^{\mathbb{L}}(1, \vec{x}+\vec{y})$$

and for convenience we set $m = \zeta_D = \zeta(t)$ for $t \in [q_{D-1}, 1)$. First note that since $|\partial_{x(u)}\Phi_{a,\zeta}^{\mathbb{L}}(1,\vec{x})| \le 1 + |a|$ holds, there are no issues of convergence in any of the expectations even though $\vec{y}$ has unbounded support.

By differentiating in the endpoint value $\vec{x}+\vec{y}$ before taking expectation in $\vec{y}$ it follows that

$$\nabla\Phi_{a,\zeta}^{\mathbb{L}} = \frac{\mathbb{E}[\nabla A e^{mA}]}{\mathbb{E}[e^{mA}]}.$$

This immediately implies that $|\partial_{x(u)}\Phi_{a,\zeta}^{\mathbb{L}}(t,\vec{x})| \le 1 + |a|$. Similarly one has

$$\partial_{x_i x_j}\Phi_{a,\zeta}^{\mathbb{L}} = \frac{\mathbb{E}\left[\partial_{x_i x_j}A + m(\partial_{x_i}A)(\partial_{x_j}A)e^{mA}\right]}{\mathbb{E}[e^{mA}]} - m\left(\frac{\mathbb{E}[\partial_{x_i}e^{mA}]}{\mathbb{E}[e^{mA}]}\right)\left(\frac{\mathbb{E}[\partial_{x_j}e^{mA}]}{\mathbb{E}[e^{mA}]}\right).$$

Combining, we compute

$$\langle T, \nabla^2 \Phi^{\mathbb{L}}_{a,\zeta} \rangle + m\langle T, (\nabla L)^{\otimes 2}\rangle = \frac{1}{\mathbb{E}[e^{mA}]}\,\mathbb{E}[\big(\langle T, \nabla^2 A\rangle + m\langle T, (\nabla A)^{\otimes 2}\rangle\big)\, e^{mA}]$$

Next, note that the time-derivative of the covariance of $\vec{y}(t)$ is $M(t)$. Since $M(t)$ is positive semidefinite we can couple together $(\vec{y}(t))_{t\in[q_{L-1},1]}$ via

$$\vec{y}(t) = \int_t^1 \sqrt{\xi''(r)M(r)}\mathrm{d}\vec{B}_r$$

where $\vec{B}_r$ is a standard Brownian motion in $\mathbb{R}^K$. Applying Ito's formula backward in time now implies

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}\, e^{mA(\vec{x},\vec{y}(t))} = -\frac{1}{2}m\,\mathbb{E}\left[\big(\langle T, \nabla^2 A\rangle + m\langle M(t), (\nabla A)^{\otimes 2}\rangle\big)\, e^{mA}\right].$$

Therefore we conclude

$$\begin{aligned}\partial_t \Phi^{\mathbb{L}}_{a,\zeta} &= -\frac{\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}\, e^{mA(\vec{x},\vec{y}(t))}}{m\,\mathbb{E}\, e^{mA(\vec{x},\vec{y}(t))}} \\ &= -\frac{1}{2}\langle T, \nabla^2 \Phi^{\mathbb{L}}_{a,\zeta}\rangle + m\langle T, (\nabla \Phi^{\mathbb{L}}_{a,\zeta})^{\otimes 2}\rangle.\end{aligned}$$

□

Proof of Proposition 5.6.7. Set

$$W_s = x + \int_{t_1}^s \zeta(r)\xi''(r)v_r\mathrm{d}r + \int_{t_1}^s \sqrt{\xi''(r)}\mathrm{d}B_r$$

and

$$V_s \equiv \Phi^{\beta}_{a,\zeta}(s, W_s) - \frac{1}{2}\int_{t_1}^s \zeta(r)\xi''(r)v_r^2\mathrm{d}r.$$

Ito's formula gives

$$\mathrm{d}V_t = \left(\partial_t \Phi^{\beta}_{a,\zeta}(t, W_t) + \zeta(t)\xi''(t)v_t\partial_x\Phi^{\beta}_{a,\zeta}(t, W_t) + \frac{1}{2}\xi''(t)\partial_{xx}\Phi^{\beta}_{a,\zeta}(t, W_t) - \frac{1}{2}\zeta(t)\xi''(t)v_t^2\right)\mathrm{d}t + Y_t\mathrm{d}B_t.$$

Here Y_t is irrelevant and (5.6.1) lets us rewrite the finite variation part of $\mathrm{d}V_t$ as

$$\begin{aligned}\partial_t \Phi^{\beta}_{a,\zeta}(t, X_t) + \zeta(t)\xi''(t)v_t\partial_x\Phi^{\beta}_{a,\zeta}(t, W_t) + \frac{1}{2}\xi''(t)\partial_{xx}\Phi^{\beta}_{a,\zeta}(t, W_t) - \frac{1}{2}\zeta(t)\xi''(t)v_t^2 \\ = -\frac{1}{2}\zeta(t)\xi''(t)\left(v_t - \partial_x\Phi^{\beta}_{a,\zeta}(t, W_t)\right)^2 \\ \leq 0.\end{aligned}$$

We conclude that

$$\Phi_\zeta^\beta(t_1, x) \geq \mathcal{X}_\zeta^{t_1,t_2}(x, v)$$

with equality when $v_r = \partial_x \Phi_\zeta^\beta(r, W_r)$ holds for all $r \in [t_1, t_2]$. By uniqueness of solutions for SDEs with Lipschitz coefficients, this implies $W_r = X_r$.

□

Proof of Proposition 5.6.8. The proof is similar to the 1-dimensional case. First, the SDE defining $\vec{X}_t$ has strong and pathwise unique solutions since $\nabla \Phi_{a,\zeta}^{\mathbb{L}}(t, \vec{x})$ is uniformly bounded and Lipschitz in $\vec{x}$. Set

$$\vec{W}_s = \vec{x} + \int_{t_1}^s \zeta(r)\xi''(r)M(r)\vec{v}_r \mathrm{d}r + \int_{t_1}^s \sqrt{\xi''(r)M(r)}\mathrm{d}\vec{B}(r)$$

and

$$V_s^{\mathbb{L}} \equiv \Phi_{a,\zeta}^{\mathbb{L}}\left(s, \vec{X}_s\right) - \frac{1}{2}\int_{t_1}^s \zeta(r)\xi''(r)\langle M(r), \vec{v}_r^{\otimes 2}\rangle \mathrm{d}r.$$

By Ito's formula,

$$\mathrm{d}V_t^{\mathbb{L}} = \left(\partial_t \Phi_\zeta^{\mathbb{L}}(t, \vec{W}_t) + \zeta(t)\xi''(t)\vec{v}_t \partial_x \Phi_\zeta^{\mathbb{L}}(t, \vec{W}_t) + \frac{1}{2}\xi''(t)\partial_{xx}\Phi_\zeta^{\mathbb{L}}(t, \vec{X}_t) - \frac{\xi''(t)}{2}\langle M(t), \vec{v}_t^{\otimes 2}\rangle\right)\mathrm{d}t + Y_t^{\mathbb{L}}\mathrm{d}B_t.$$

Here $Y_t^{\mathbb{L}}$ is again irrelevant. By (5.6.3) the finite variation part of $\mathrm{d}V_t^{\mathbb{L}}$ is

$$\begin{aligned}\partial_t \Phi_{a,\zeta}^{\mathbb{L}}(t, \vec{W}_t) + \left\langle M(t), \vec{v}_t \otimes \nabla \Phi_{a,\zeta}^{\mathbb{L}}(t, \vec{W}_t)\right\rangle + \frac{1}{2}\xi''(t)\partial_{xx}\Phi_{a,\zeta}^{\mathbb{L}}(t, \vec{W}_t) - \frac{\xi''(t)}{2}\langle M(t), \vec{v}_t^{\otimes 2}\rangle \\ = -\frac{1}{2}\left\langle M(t), \left(\vec{v}_t - \nabla \Phi_{a,\zeta}^{\mathbb{L}}(t, \vec{W}_t)\right)^{\otimes 2}\right\rangle \\ \leq 0.\end{aligned}$$

We conclude that

$$\Phi_{a,\zeta}^{\mathbb{L}}(t_1, \vec{x}) \geq \mathcal{X}_{a,\zeta}^{\mathbb{L},t_1,t_2}(\vec{x}, \vec{v})$$

with equality when

$$\vec{v}_r = \nabla \Phi_{a,\zeta}^{\mathbb{L}}(r, \vec{W}_r)$$

holds for all $r \in [t_1, t_2]$. Again, uniqueness of solutions to SDEs with Lipschitz coefficients implies $\vec{W}_r = \vec{X}_r$.

□

5.7 Necessity of Full Branching Trees

In this section we show, roughly speaking, that it is necessary to use a full branching tree to obtain our results within the overlap gap framework. We restrict for convenience to the setting of spherical models with null external field $h = 0$ and set $\mathsf{ALG}_\xi^{\mathrm{Sp}} = \mathsf{ALG}_{\xi,0}^{\mathrm{Sp}} = \int_0^1 \xi''(t)^{1/2}\,\mathrm{d}t$ (recall Proposition 5.2.2) and $\mathsf{OPT}_\xi^{\mathrm{Sp}} = \mathsf{OPT}_{\xi,0}^{\mathrm{Sp}}$.

A consequence of Theorem 27, proved near the end of this section, can be expressed informally as follows for any ξ with $\mathsf{ALG}_\xi^{\mathrm{Sp}} < \mathsf{OPT}_\xi^{\mathrm{Sp}}$. Recall the canonical bijection between finite ultrametric spaces and edge-weighted rooted trees (or see Subsection 5.7.2 for a reminder). For all finite ultrametric spaces X of diameter at most $\sqrt{2}$ whose corresponding rooted tree does not contain a subdivision of a full binary subtree of depth D, with probability at least $1 - e^{-\Omega(N)}$ the following holds. There exists an isometric (up to the scaling factor $\sqrt{N}$) embedding $\iota : X \to S_N$ such that

$$H_N(\iota(x)) \geq (\mathsf{ALG}_\xi^{\mathrm{Sp}} + \varepsilon_{\xi,D})N, \quad \forall x \in X.$$

Here $\varepsilon_{\xi,D} > 0$ is a constant depending only on ξ and D, and in particular is independent of the size of the ultrametric X. In other words, to rule out algorithms achieving better than $\mathsf{ALG}_\xi^{\mathrm{Sp}} + \varepsilon$ using forbidden ultrametrics, as $\varepsilon \to 0$ it is necessary to take $D \to \infty$, in effect using the full power of Proposition 5.3.2.

The full statement of Theorem 27 shows that in fact a super-constant amount of branching must occur at all "depths" in $[0, 1]$ where $\xi''(t)^{-1/2}$ is strictly convex. We also show in Theorem 28 that there exists an embedding ι as above with large average energy

$$\frac{1}{|X|}\sum_{x \in X} H_N(\iota(x)) \geq (\mathsf{ALG}_\xi^{\mathrm{Sp}} + \varepsilon_{\xi,D})N$$

unless "almost all of" X branches a super-constant amount at "almost all such depths". Note that this average energy is what the Guerra-Talagrand interpolation actually allows one to upper bound. Throughout this section we always consider just a single Hamiltonian H_N. This corresponds to the case $\vec{p} \approx (1, 1, \ldots, 1)$, i.e. a correlation function $\chi(p)$ which sharply increases near $p = 1$ such as $\chi(p) = p^{100}$.

Our plan to prove Theorem 27 is as follows. If $\mathsf{ALG}_\xi^{\mathrm{Sp}} < \mathsf{OPT}_\xi^{\mathrm{Sp}}$, there exists an interval $[a, b] \subseteq [0, 1]$ on which $(\xi'')^{-1/2}$ is strictly convex. Let $\mathbb{T}$ be the finite rooted tree with leaf set corresponding to the ultrametric space X. Let $\varepsilon > 0$ be a small constant depending only on ξ and D. We use the algorithm of [Sub21] to find embeddings of ancestor points $\iota(x_a)$ for each $x \in X$ of norm $\|\iota(x_a)\|_2 = \sqrt{aN}$ which satisfy

$$H_N(\iota(x_a)) \geq \left(\int_0^a \xi''(t)^{1/2}\,\mathrm{d}t - \varepsilon\right) N.$$

Next we embed the depth $[a, b]$ parts of $\mathbb{T}$ so that the resulting depth b ancestor embeddings $\iota(x_b)$ satisfy

$$H_N(\iota(x_b)) \geq \left(\int_0^b \xi''(t)^{1/2} \, \mathrm{d}t + 2\varepsilon \right) N.$$

In other words, from radius $\sqrt{aN}$ to $\sqrt{bN}$, the embedded points' energy grows by $\int_a^b \xi''(t)^{1/2} \, \mathrm{d}t + 3\varepsilon$, which exceeds the maximum possible growth of an overlap concentrated algorithm by a small constant 3ε. This is the main step of our procedure, and it succeeds whenever the portion of $\mathbb{T}$ at depths in $[a, b]$ does not contain a full binary tree of depth D. The proof uses induction on D, and the $D = 1$ case is described in Figure 5.2. We remark that our proof is essentially constructive assuming access to an oracle to find many orthogonal near-maximizers of H_N on arbitrary bands as guaranteed by Lemma 5.7.4.

Finally we again use the algorithm of [Sub21] to define embeddings of the leaves $\iota(x) \in S_N$ for $x \in X$ with

$$H_N(\iota(x)) \geq \left(\int_0^1 \xi''(t)^{1/2} \, \mathrm{d}t + \varepsilon \right) N.$$

We remark that in previous multi-OGP arguments, ultrametricity of the forbidden configuration does not explicitly enter. However in these arguments, it is always *possible* that the structure of replicas identified is an ultrametric. Specifically, in a "star" multi-OGP [RV17a, GS17, GK21a] all the replicas are pairwise equidistant. For the "ladder" OGP implementations of [Wei22, BH21], the forbidden structure is defined by applying some stopping rule to choose a finite number of solutions from a "stably evolving" sequence of algorithmic outputs. In both settings it is possible that the resulting configuration is a star ultrametric with all pairwise nonzero distances equal. However, the rooted tree corresponding to such an ultrametric does not contain even a full binary tree of depth $D = 2$. Therefore Theorem 27 strongly suggests that existing OGP arguments are incapable of ruling out Lipschitz $\mathcal{A}$ from achieving energies down to the algorithmic threshold $\mathsf{ALG}_\xi^{\mathrm{Sp}}$.

5.7.1 Preparation

For given ξ and $t \in [0, 1]$, define

$$\mathsf{ALG}_\xi^{\mathrm{Sp}}(t) = \int_0^t \xi''(s)^{1/2} \mathrm{d}s$$

so that $\mathsf{ALG}_\xi^{\mathrm{Sp}}(1) = \mathsf{ALG}_\xi^{\mathrm{Sp}}$. Define also

$$\mathsf{ALG}_\xi^{\mathrm{Sp}}([a, b]) = \mathsf{ALG}_\xi^{\mathrm{Sp}}(b) - \mathsf{ALG}_\xi^{\mathrm{Sp}}(a).$$

Define

$$\xi_a(t) = \xi(t) - \xi(a) - (t - a)\xi'(a).$$

Note that $\xi_a(a) = \xi_a'(a) = 0$, and $\xi_a''(t) = \xi''(t)$ for all t. Define the rescaled mixture function

$$\xi_{[a,b]}(t) = \xi_a\left(a + (b-a)t\right).$$

We derive

$$\mathsf{ALG}^{\mathrm{Sp}}_{\xi_{[a,b]}} = \int_0^1 \sqrt{\xi''_{[a,b]}(t)}\mathrm{d}t = \int_a^b \sqrt{\xi_a''(s)}\mathrm{d}s = \int_a^b \sqrt{\xi''(s)}\mathrm{d}s = \mathsf{ALG}^{\mathrm{Sp}}_{\xi}([a,b]).$$

Correspondingly, define

$$\mathsf{OPT}^{\mathrm{Sp}}_{\xi}([a,b]) = \mathsf{OPT}^{\mathrm{Sp}}_{\xi_{[a,b]}}.$$

Proposition 5.7.1. *Suppose $\frac{\mathrm{d}^2}{\mathrm{d}t^2}(\xi''(t)^{-1/2}) > 0$ for $t \in [a,b] \subseteq [0,1]$. Then*

$$\mathsf{OPT}^{\mathrm{Sp}}_{\xi}([a,b]) > \mathsf{ALG}_{\xi}([a,b]). \tag{5.7.1}$$

Proof. The result follows from Proposition 5.2.2 applied to $\xi_{[a,b]}$. □

The next proposition follows from the work [Sub18] and ensures the existence of many approximately orthogonal replicas which each approximately achieve the ground state energy in spherical spin glasses without external field. In Lemma 5.7.4 we make several simple modifications to this result, for instance requiring that the replicas be exactly orthogonal.

Proposition 5.7.2. *Suppose $\frac{\mathrm{d}^2}{\mathrm{d}t^2}(\xi''(t)^{-1/2}) > 0$ for $t \in [0,1]$. Then for any $C, \varepsilon > 0$ and $k \in \mathbb{N}$, for $N \geq N_0 = N_0(\xi, C, \varepsilon, k)$, with probability at least $1 - e^{-CN}$ either $H_N \notin K_N$ (recall Proposition 5.2.3) or the following holds. There exist k points $\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_k \in S_N$ with*

$$|R(\boldsymbol{\sigma}_i, \boldsymbol{\sigma}_j)| \leq \varepsilon, \quad 1 \leq i < j \leq k$$

and

$$H_N(\boldsymbol{\sigma}_i) \geq N(\mathsf{OPT}^{\mathrm{Sp}}_{\xi} - \varepsilon), \quad i \in [k].$$

Proof. With the absence of external field, it follows from [Sub18, Lemma 42] that 0 is multi-samplable. Let $\mathcal{Q}_k(\varepsilon) \subseteq B(\boldsymbol{m}, \varepsilon)^k \cap S_N^k$ denote the set of $\vec{\boldsymbol{\sigma}}$ with $|R(\boldsymbol{\sigma}_i, \boldsymbol{\sigma}_j)| \leq \varepsilon$ for $i \neq j$. Let μ be the uniform measure on S_N. Define

$$F_{N,\beta} = \frac{1}{\beta N} \log \int_{S_N} \exp \beta H_N(\boldsymbol{\sigma}) \, \mathrm{d}\mu(\boldsymbol{\sigma})$$

to be the quenched free energy of H_N on S_N at inverse temperature β and

$$\widetilde{F}_{N,\beta}(\boldsymbol{m}) = \widetilde{F}_{N,\beta}(\boldsymbol{m}, k_N, \varepsilon) \equiv \frac{1}{\beta N k_N} \log \int_{\mathcal{Q}_{k_N}(\varepsilon)} \exp \beta \sum_{i=1}^{k_N} H_N(\boldsymbol{\sigma}_i) \, \mathrm{d}\mu^k(\vec{\boldsymbol{\sigma}}).$$

Here k_N grows to ∞ with N at a suitably slow rate. By [Sub18, Proposition 1 and Theorem 3][4] it follows that for $N \geq N_0$ sufficiently large,

$$\mathbb{P}\left[\mathbb{E}\,\widetilde{F}_{N,\beta}(\boldsymbol{m}) - \mathbb{E}\,F_{N,\beta} \geq -\varepsilon\right] \geq 1 - e^{-CN}.$$

Therefore there exists some $\vec{\boldsymbol{\sigma}} \in \mathcal{Q}_{k_N}(\varepsilon)$ satisfying

$$\sum_{i=1}^{k_N} H_N(\boldsymbol{\sigma}_i) \geq N k_N(\mathsf{OPT}_\xi^{\mathrm{Sp}} - \varepsilon - o_\beta(1)).$$

Here $o_\beta(1)$ is a value tending to 0 as $\beta \to \infty$, uniformly in everything else. Assuming $H_N \in K_N$, the values $\frac{1}{N}|H_N(\boldsymbol{\sigma}_i)|$ are uniformly bounded by a constant C_1 (because $H_N(0) = 0$). It follows by Markov's inequality that at least $k_N\left(\frac{\varepsilon}{10C_1} - \varepsilon - o_\beta(1)\right)$ of the $\boldsymbol{\sigma}_i$ satisfy $H_N(\boldsymbol{\sigma}_i) \geq N(\mathsf{OPT}_\xi^{\mathrm{Sp}} - \frac{\varepsilon}{2} - o_\beta(1))$. Since $k_N \to \infty$, eventually

$$k \leq \left\lfloor k_N\left(\frac{\varepsilon}{10C_1} - \varepsilon - o_\beta(1)\right)\right\rfloor$$

for suitably large β, which completes the proof. □

For fixed $\boldsymbol{m}$, define the first-order Taylor expansion

$$\overline{H}_N^{\boldsymbol{m}}(\boldsymbol{\sigma}) = H_N(\boldsymbol{m}) + \langle \nabla H_N(\boldsymbol{m}), \boldsymbol{\sigma} - \boldsymbol{m}\rangle.$$

of H_N and write

$$H_N = \overline{H}_N^{\boldsymbol{m}} + \widehat{H}_N^{\boldsymbol{m}}.$$

For $0 \leq a \leq b \leq 1$ with $\boldsymbol{m} \in \sqrt{a} \cdot S_N$, define $B(\boldsymbol{m}, 0, b) = B(\boldsymbol{m}, 0) \cap \sqrt{b} \cdot S_N$ and its convex hull $B(\boldsymbol{m}, 0, [a, b])$.

Lemma 5.7.3. *For any fixed $\boldsymbol{m}$, the law of $\widehat{H}_N^{\boldsymbol{m}}$ restricted to $B(\boldsymbol{m}, 0)$ is a Gaussian process with covariance*

$$\mathbb{E}[\widehat{H}_N^{\boldsymbol{m}}(\boldsymbol{\sigma}^1)\widehat{H}_N^{\boldsymbol{m}}(\boldsymbol{\sigma}^2)] = N\xi_a(R(\boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2)). \tag{5.7.2}$$

Moreover the restrictions of $\widehat{H}_N^{\boldsymbol{m}}$ and $\overline{H}_N^{\boldsymbol{m}}$ to $B(\boldsymbol{m}, 0)$ are independent.

Proof. Note that for all $\boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2 \in B(\boldsymbol{m}, 0)$,

$$R(\boldsymbol{\sigma}^1 - \boldsymbol{m}, \boldsymbol{\sigma}^2 - \boldsymbol{m}) = R(\boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2) - a.$$

[4] In the statement of [Sub18, Theorem 3], there are values δ_N, ρ_N which also shrink with N. We are taking $\varepsilon = \rho_N$ a small constant and ignoring the constraint from δ, so our value of $\widetilde{F}_{N,\beta}(\boldsymbol{m})$ is larger than that of [Sub18]. Therefore the lower bound on $\widetilde{F}_{N,\beta}(\boldsymbol{m})$ we use is somewhat weaker than in the results cited.

Since $\xi_a(t)$ has all derivatives non-negative for $t \geq a$, we may sample a centered Gaussian process $\widetilde{H}_N$ on $B(\boldsymbol{m}, 0, [a, 1])$ with covariance given by

$$\begin{aligned}\mathbb{E}[\widetilde{H}_N(\boldsymbol{\sigma}^1)\widetilde{H}_N(\boldsymbol{\sigma}^2)] &= N\xi_a(R(\boldsymbol{\sigma}^1 - \boldsymbol{m}, \boldsymbol{\sigma}^2 - \boldsymbol{m}) + a) \\ &= N\xi_a(R(\boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2)).\end{aligned}$$

Next, generate the independent centered Gaussian process $\underline{H}_N$ by

$$\mathbb{E}[\underline{H}_N(\boldsymbol{\sigma}^1)\underline{H}_N(\boldsymbol{\sigma}^2)] = N\left(\xi(a) + \xi'(a)\left(R(\boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2) - a\right)\right).$$

It follows by adding covariances (with $x = R(\boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2)$ in the definition of ξ_a) that

$$\widetilde{H}_N + \underline{H}_N \stackrel{d}{=} H_N$$

when restricted to $B(\boldsymbol{m}, 0)$. Since $\xi_a(a) = \xi_a'(a) = 0$, it follows that $\widetilde{H}_N(\boldsymbol{m}) = 0$ and $\nabla\widetilde{H}_N(\boldsymbol{m}) = 0$ hold almost surely. Therefore $\underline{H}_N = \overline{H}_N^{\boldsymbol{m}}$ is the first-order Taylor expansion of H_N around $\boldsymbol{m}$, and then also $\widetilde{H}_N = \widehat{H}_N^{\boldsymbol{m}}$. Moreover $\widetilde{H}_N$ and $\underline{H}_N$ are independent by construction. This concludes the proof. □

In the following Lemma 5.7.4, we refine Proposition 5.7.2 in several simple but convenient ways. In particular, Lemma 5.7.3 implies the same result uniformly over all bands $B(\boldsymbol{m}, 0, b)$; it also guarantees exact orthogonality. Lemma 5.7.4 will serve as a useful tool for embedding more complicated ultrametric trees. Roughly speaking, it gives a way to gain on the embedding algorithm of [Sub21] (stated later as Proposition 5.7.10).

Lemma 5.7.4. *Suppose* $\frac{\mathrm{d}^2}{\mathrm{d}t^2}(\xi''(t)^{-1/2}) > 0$ *for* $t \in [a, b] \subseteq [0, 1]$. *Then there exists* $\varepsilon > 0$ *depending only on* ξ, a, b *such that for any* k, *for* $N \geq N_0(\xi, a, b, k)$ *sufficiently large and some* $c = c(\xi, a, b, k)$, *with probability* $1 - e^{-cN}$ *the following holds.*

For any $\boldsymbol{m}$ *with* $||\boldsymbol{m}||_N^2 = a \leq 1$ *and any linear subspace* $W \subseteq \mathbb{R}^N$ *with* $\dim(W) \geq N - k$, *there exist* k *points* $\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_k \in W + \boldsymbol{m}$ *such that*

$$R(\boldsymbol{\sigma}_i - \boldsymbol{m}, \boldsymbol{\sigma}_j - \boldsymbol{m}) = (b - a) \cdot \mathbb{I}(i = j) \quad \forall i, j \in [k] \tag{5.7.3}$$

and

$$H_N(\boldsymbol{\sigma}_i) \geq H_N(\boldsymbol{m}) + N(\mathsf{ALG}_\xi^{\mathrm{Sp}}([a, b]) + \varepsilon) \quad \forall i \in [k]. \tag{5.7.4}$$

Proof. Consider a (non-random) $\eta\sqrt{N}$-net $\mathcal{N}_\eta$ on $\sqrt{a} \cdot S_N$ of size at most $(10/\eta)^N$. For any $\boldsymbol{m} \in$

$\sqrt{a} \cdot S_N$, the Hamiltonian $\widehat{H}_N^{\boldsymbol{m}}(\boldsymbol{\sigma})$ restricted to $B(\boldsymbol{m}, 0, b)$ has covariance

$$\begin{aligned}\mathbb{E}\, \widehat{H}_N^{\boldsymbol{m}}(\boldsymbol{\sigma}_1)\widehat{H}_N^{\boldsymbol{m}}(\boldsymbol{\sigma}_2) &= N\xi_a(R(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)) \\ &= N\xi_{[a,b]}\left(R\left(\frac{\boldsymbol{\sigma}_1 - \boldsymbol{m}}{\sqrt{b-a}}, \frac{\boldsymbol{\sigma}_2 - \boldsymbol{m}}{\sqrt{b-a}}\right)\right).\end{aligned}$$

Since

$$||\boldsymbol{\sigma} - \boldsymbol{m}||_2 = \sqrt{N(b-a)}$$

for $\boldsymbol{\sigma} \in B(\boldsymbol{m}, 0, b)$, we conclude that $\widehat{H}_N^{\boldsymbol{m}}$ is exactly an $N-1$ dimensional spin glass with mixture $\xi_{[a,b]}$ on $B(\boldsymbol{m}, 0, b)$ up to rescaling the input.

Fix a large constant M, and choose ε sufficiently small depending on M. We apply Proposition 5.7.2 to $\widehat{H}_N^{\boldsymbol{m}}$ with mixture $\xi_{[a,b]}(t)$ based on the observation just above. Recall that the constant C in Proposition 5.7.2 can be arbitrarily large. It follows by a union bound that with probability $1 - e^{-C_1 N}$, for all $\boldsymbol{n} \in \mathcal{N}_\eta$ there exist $\widetilde{\boldsymbol{\sigma}}_1(\boldsymbol{n}), \dots, \widetilde{\boldsymbol{\sigma}}_M(\boldsymbol{n})$ satisfying

$$|R(\widetilde{\boldsymbol{\sigma}}_i(\boldsymbol{n}) - \boldsymbol{n}, \widetilde{\boldsymbol{\sigma}}_j(\boldsymbol{n}) - \boldsymbol{n}) - (b-a) \cdot \mathbb{I}(i=j)| \le \varepsilon \quad \forall 1 \le i < j \le M$$

and

$$\widehat{H}_N(\widetilde{\boldsymbol{\sigma}}_i(\boldsymbol{n})) \ge N(\mathsf{OPT}_\xi^{\mathrm{Sp}}([a,b]) - \varepsilon) \quad \forall i \in [M]. \tag{5.7.5}$$

For any $\boldsymbol{m} \in \sqrt{a} \cdot S_N$, there exists by definition $\boldsymbol{n} \in \mathcal{N}_\eta$ with $||\boldsymbol{m} - \boldsymbol{n}|| \le \eta\sqrt{N}$. Then with $\widetilde{\boldsymbol{\sigma}}_i = \widetilde{\boldsymbol{\sigma}}_i(\boldsymbol{n})$ as above,

$$|R(\widetilde{\boldsymbol{\sigma}}_i - \boldsymbol{m}, \widetilde{\boldsymbol{\sigma}}_j - \boldsymbol{m}) - (b-a) \cdot \mathbb{I}(i=j)| \le \varepsilon_1 \quad \forall 1 \le i < j \le M$$

for some $\varepsilon_1 = o_{\varepsilon,\eta}(1)$ tending to 0 as $\varepsilon, \eta \to 0$. Define the linear subspace $\widetilde{W} \subseteq W$ by

$$\widetilde{W} = W \cap \boldsymbol{m}^\perp \cap (\nabla H_N)^\perp$$

where $(\cdot)^\perp$ denotes orthogonal complement. Let $P_{\widetilde{W}^\perp}$ be the orthogonal projection matrix onto $\widetilde{W}^\perp$. It is easy to see that

$$\left|\left|\sum_{i=1}^M (\widetilde{\boldsymbol{\sigma}}_i - \boldsymbol{m})^{\otimes 2}\right|\right|_2^2 \le (1 + M\varepsilon)N \le 2N$$

for ε sufficiently small. Then

$$\begin{aligned}\sum_{i=1}^M ||P_{\widetilde{W}^\perp}(\widetilde{\boldsymbol{\sigma}}_i)||_2^2 &= \left\langle P_{\widetilde{W}^\perp}, \sum_{i=1}^M (\widetilde{\boldsymbol{\sigma}}_i - \boldsymbol{m})^{\otimes 2} \right\rangle \\ &\le ||P_{\widetilde{W}^\perp}||_2^2 \cdot \left|\left|\sum_{i=1}^M (\boldsymbol{\sigma}_i - \boldsymbol{m})^{\otimes 2}\right|\right|_2^2 \\ &\le 2(k+2)N.\end{aligned}$$

By the pigeonhole principle, at most $M-k$ values $i \in [M]$ can satisfy

$$||P_{\widetilde{W}^\perp}(\boldsymbol{\sigma}_i - \boldsymbol{m})||_2^2 \ge \frac{2(k+2)N}{M-k}.$$

It follows that there exist a subset $\widetilde{\boldsymbol{\sigma}}_{i_1}, \dots, \widetilde{\boldsymbol{\sigma}}_{i_k}$ with

$$||P_{\widetilde{W}^\perp}(\widetilde{\boldsymbol{\sigma}}_{i_j} - \boldsymbol{m})||_2^2 \le \eta N, \quad j \in [k]$$

where $\eta \le \frac{2(k+2)}{M-k}$ is arbitrarily small (by choosing M large depending on k). Defining $\boldsymbol{\sigma}'_{i_1}, \dots, \boldsymbol{\sigma}'_{i_k}$ by

$$\boldsymbol{\sigma}'_{i_j} - \boldsymbol{m} = P_{\widetilde{W}^\perp}(\widetilde{\boldsymbol{\sigma}}_{i_j} - \boldsymbol{m}),$$

we have

$$\boldsymbol{\sigma}'_{i_1}, \dots, \boldsymbol{\sigma}'_{i_k} \in \boldsymbol{m} + \widetilde{W}$$

satisfying

$$|R(\boldsymbol{\sigma}'_{i_j} - \boldsymbol{m}, \boldsymbol{\sigma}_{i_\ell} - \boldsymbol{m}) - (b-a) \cdot \mathbb{I}(j=\ell)| \le \varepsilon_2, \quad j, \ell \in [k]$$

and

$$||\boldsymbol{\sigma}'_{i_j} - \widetilde{\boldsymbol{\sigma}}_{i_j}||_2^2 \le \eta N, \quad j \in [k].$$

Here $\varepsilon_2 = o_{\varepsilon_1,\eta}(1)$ tends to 0 as ε_1 and η tend to 0. Using Gram-Schmidt orthonormalization inside the affine subspace $B(\boldsymbol{m}, 0)$, for $\varepsilon_3 = o_{\varepsilon_2}(1)$ we may find $\hat{\boldsymbol{\sigma}}_1, \dots, \hat{\boldsymbol{\sigma}}_k \in B(\boldsymbol{m}, [a,b]) \cap W$ satisfying

$$R(\hat{\boldsymbol{\sigma}}_i - \boldsymbol{m}, \hat{\boldsymbol{\sigma}}_j - \boldsymbol{m}) = (b-a) \cdot \mathbb{I}(i=j) \quad \forall 1 \le i < j \le k$$

and

$$||\hat{\boldsymbol{\sigma}}_j - \boldsymbol{\sigma}'_{i_j}||_2^2 \le \varepsilon_3 N \quad \forall j \in [k].$$

Assuming H_N is $C_1\sqrt{N}$-Lipschitz with respect to the $\|\cdot\|_2$ norm (recall Proposition 5.2.3), this implies based on (5.7.5) that for some $\varepsilon_4 = o_{\varepsilon_3}(C_1+1)$,

$$\widehat{H}_N(\hat{\boldsymbol{\sigma}}_j) \ge N(\mathsf{OPT}_\xi^{\mathrm{Sp}} - \varepsilon_4) \quad \forall i \in [k]$$

and

$$||\hat{\boldsymbol{\sigma}}_j - \widetilde{\boldsymbol{\sigma}}_{i_j}||_2^2 \leq 2(\varepsilon_3 + \eta)N \quad \forall j \in [k].$$

Recalling Proposition 5.7.1, this completes the proof. □

5.7.2 Trees and Ultrametrics

We recall the well known connection between trees and ultrametric spaces. Here and throughout given a rooted tree $\mathbb{T}$ with root $r(\mathbb{T})$ we denote by $\mathrm{pa}(v)$ the parent of $v \in V(\mathbb{T})\backslash\{r(\mathbb{T})\}$

Definition 5.7.5. *[BD98a] A dated, rooted tree $\mathbb{T}$ with range $[a, b] \subseteq [0, 1]$ is a finite tree rooted at $r(\mathbb{T}) \in V(\mathbb{T})$ together with a height function*

$$|\cdot| : V(\mathbb{T}) \to [a, b]$$

satisfying the following properties.

- $|r(\mathbb{T})| = a$.
- $|v| = b$ *for all leaves* $v \in L(\mathbb{T})$.
- $|\mathrm{pa}(v)| < |v|$ *for all* $v \in V(\mathbb{T})\backslash\{r(\mathbb{T})\}$.

We say $\mathbb{T}$ is reduced if no $v \in V(\mathbb{T})$ except possibly $r(\mathbb{T})$ has exactly 1 child.

In a rooted tree, let $u \wedge v \in V(\mathbb{T})$ denote the least common ancestor of vertices u and v. To any dated rooted tree $\mathbb{T}$, we associate a metric $d_T : V(\mathbb{T}) \times V(\mathbb{T}) \to [0, \sqrt{2}]$ characterized by

$$|u \wedge v| = \frac{|u| - d_T(u, v)^2 + |v|}{2}, \qquad u, v \in V(\mathbb{T}). \tag{5.7.6}$$

When $u, v \in L(\mathbb{T})$ are leaves and $\mathbb{T}$ has range $[a, b]$, this becomes

$$|u \wedge v| = b - \frac{d_T(u, v)^2}{2}, \qquad u, v \in L(\mathbb{T}). \tag{5.7.7}$$

Crucially, observe that for $u, v \in L(\mathbb{T})$, the value $d_T(u, v)$ is a strictly decreasing function of $|u \wedge v|$. Therefore d_T defines an ultrametric on $L(\mathbb{T})$, or in fact the set of vertices at any fixed height. The specific decreasing bijection between $|u \wedge v| \in [0, 1]$ and $d_T(u, v) \in [0, \sqrt{2}]$ for $u, v \in L(\mathbb{T})$ can in general be arbitrary; the one above is suited for embeddings into Euclidean space since

$$R(\boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2) = \frac{R(\boldsymbol{\sigma}^1, \boldsymbol{\sigma}^1) - ||\boldsymbol{\sigma}^1 - \boldsymbol{\sigma}^2||_N^2 + R(\boldsymbol{\sigma}^2, \boldsymbol{\sigma}^2)}{2}, \qquad \boldsymbol{\sigma}^1, \boldsymbol{\sigma}^2 \in \mathbb{R}^N. \tag{5.7.8}$$

The following type of result is well known and seems to be folklore.

Proposition 5.7.6. *[RTV86, Section 6],[BD98a] For any finite set X,* (5.7.6) *defines a bijection between the following two isomorphism classes.*

1. *Dated, rooted reduced trees with range* $[0,1]$ *and leaf set* X.
2. *Ultrametric structures on* X *with diameter at most* $\sqrt{2}$.

Any dated, rooted tree can be naturally reduced by removing vertices with a single child and connecting their parent and child. Hence we will consider general dated, rooted trees to give ourselves more flexibility. We are interested in embeddings of the leaves $L(\mathbb{T})$ into level sets $\{\boldsymbol{\sigma} \in \mathbb{R}^N : H_N(\boldsymbol{\sigma}) \geq (\mathsf{ALG} + \varepsilon)N\}$ which are isometries up to the scaling factor $\sqrt{N}$. It will be convenient to embed the entire vertex set $V(\mathbb{T})$.

Definition 5.7.7. *A* Euclidean embedding *of a dated, rooted tree $\mathbb{T}$ to is a function $\iota : V(\mathbb{T}) \to \mathbb{R}^N$ satisfying*

$$R(\iota(u), \iota(v)) = |u \wedge v| \qquad \forall u, v \in V(\mathbb{T}).$$

or equivalently (by (5.7.8)*),*

$$||\iota(u) - \iota(v)||_N = d_T(u, v) \qquad \forall u, v \in V(\mathbb{T}).$$

The next lemma gives an alternate characterization of Euclidean embeddings. .

Lemma 5.7.8. *$\iota : V(\mathbb{T}) \to \mathbb{R}^N$ is a Euclidean embedding if and only if the following properties hold. Below we implicitly define $|\mathrm{pa}(r(\mathbb{T}))| = 0$ and $\iota(\mathrm{pa}(r(\mathbb{T})) = 0$.*

1. *$\iota(r(\mathbb{T})) = a$.*
2. *For all $v \in V(\mathbb{T})$,*
$$||\iota(v) - \iota(\mathrm{pa}(v))||_N = |v| - |\mathrm{pa}(v)|.$$
3. *For all distinct $u, v \in V(\mathbb{T})$,*
$$R(\iota(u) - \iota(\mathrm{pa}(u)), \iota(v) - \iota(\mathrm{pa}(v))) = 0.$$

Proof. First if ι is a Euclidean embedding, then clearly the first property holds. The second holds because

$$\begin{aligned} ||\iota(v) - \iota(\mathrm{pa}(v))||_N^2 &= R(\iota(v) - \iota(\mathrm{pa}(v)), \iota(v) - \iota(\mathrm{pa}(v))) \\ &= |v \wedge v| - |v \wedge \mathrm{pa}(v)| - |\mathrm{pa}(v) \wedge v| + |\mathrm{pa}(v) \wedge \mathrm{pa}(v)| \\ &= |v| - |\mathrm{pa}(v)|. \end{aligned}$$

For the third property, we compute

$$R(\iota(u) - \iota(\mathrm{pa}(u)), \iota(v) - \iota(\mathrm{pa}(v))) = |u \wedge v| - |v \wedge \mathrm{pa}(v)| - |\mathrm{pa}(u) \wedge v| + |\mathrm{pa}(u) \wedge \mathrm{pa}(v)|.$$

Since $u \neq v$, without loss of generality suppose $v \neq (u \wedge v)$. Then v is an ancestor of neither u nor $\mathrm{pa}(u)$. The third property then follows because

$$u \wedge v = u \wedge \mathrm{pa}(v) \qquad \text{and} \qquad \mathrm{pa}(u) \wedge v = \mathrm{pa}(u) \wedge \mathrm{pa}(v).$$

In the other direction, let us assume the three properties hold and show ι is a Euclidean embedding. Consider vertices $u = u_n$ and $v = v_m$ with ancestor paths

$$(r(\mathbb{T}) = u_0, u_1, \ldots, u_{n-1}), \qquad (r(\mathbb{T}) = v_0, v_1, \ldots, v_{m-1}).$$

Suppose that $u \wedge v = u_d = v_d$, so that $u_j = v_j$ if and only if $j \leq d$. Then we expand

$$\begin{aligned} R(\iota(u), \iota(v)) &= a + \sum_{\substack{1 \leq i \leq n \\ 1 \leq j \leq m}} R\big(\iota(u_i) - \iota(u_{i-1}), \iota(v_j) - \iota(v_{j-1})\big) \\ &= a + \sum_{1 \leq i \leq d} R\big(\iota(u_i) - \iota(u_{i-1}), \iota(u_i) - \iota(u_{i-1})\big) \\ &= a + \sum_{1 \leq i \leq d} |u_i| - |u_{i-1}| \\ &= a + |u_d| - |r(\mathbb{T})| \\ &= |u \wedge v|. \end{aligned}$$

□

Next, define the depth D rooted binary tree $\mathbb{T}_D^2$ on vertex set

$$V(\mathbb{T}_D^2) = \emptyset \cup [2] \cup [2]^2 \cup \cdots \cup [2]^D.$$

As usual, a vertex $v \in [2]^j$ is the parent of $u \in [2]^{j+1}$ if and only if v is an initial substring of u. We say the rooted tree $\mathbb{T}$ *contains* $\mathbb{T}_D^2$ if there exists an ancestry-preserving injection

$$\phi : V(\mathbb{T}_D^2) \to V(\mathbb{T})$$

(which need not preserve the root). Define the branching depth $D(\mathbb{T})$ to be the largest D such that $\mathbb{T}$ contains $\mathbb{T}_D^2$. For $v \in V(\mathbb{T})$, define $D(v) = D(\mathbb{T}_v)$ where $\mathbb{T}_v \subseteq \mathbb{T}$ is the subtree rooted at v.

Proposition 5.7.9. *For any rooted tree $\mathbb{T}$, the set*

$$V_D = \{v \in V(\mathbb{T}) : D(v) = D(\mathbb{T})\} \tag{5.7.9}$$

is a simple path graph with one endpoint at $r(\mathbb{T})$.

Proof. Let $D = D(\mathbb{T})$. Clearly V_D is closed under ancestry and contains $r(\mathbb{T})$. Suppose for sake of contradiction that V_D is not a path with $r(\mathbb{T})$ as an endpoint. Then V_D contains vertices v and w neither of which is an ancestor of the other. But if the disjoint subtrees rooted at v and w each contain $\mathbb{T}^2_D$, then $\mathbb{T}$ contains $\mathbb{T}^2_{D+1}$, a contradiction. □

We will use the following slight generalization of the main result of [Sub21]. It can be seen as the "default" embedding procedure which ensures energy $\mathsf{ALG}^{\mathrm{Sp}}_\xi$ at the leaves, while Lemma 5.7.4 gives a tool to improve over this embedding on intervals $[a, b]$ where $\xi''(t)^{-1/2}$ is convex.

Proposition 5.7.10. *For any $\varepsilon > 0$ and $k \in \mathbb{Z}^+$, there exist c and N_0 depending on ξ, ε, k such that with probability $1 - e^{-cN}$ the following holds for all $N \geq N_0$.*

For any $\boldsymbol{m}$ with $||\boldsymbol{m}||_N^2 = q \leq 1$, any dated, rooted tree $\mathbb{T}$ of order $|V(\mathbb{T})| \leq k$ with range $[q, 1]$, and any linear subspace $W \subseteq \mathbb{R}^N$ with $\dim(W) \geq N - k$, there is an embedding ι of X into $W + \boldsymbol{m}$ such that

$$||\iota(u) - \iota(v)||_N = d(u, v) \quad \forall u, v \in V(\mathbb{T}) \tag{5.7.10}$$

and

$$H_N(\iota(x)) \geq H_N(\boldsymbol{m}) + N \cdot (\mathsf{ALG}^{\mathrm{Sp}}_\xi(\|\iota(u)\|_N^2 - \mathsf{ALG}^{\mathrm{Sp}}_\xi(\|\boldsymbol{m}\|_N^2) - \varepsilon) \quad \forall v \in V(\mathbb{T}). \tag{5.7.11}$$

Proof. The proof is essentially contained in [Sub21, Theorem 4 and Remark 6]. The restriction to $W + \boldsymbol{m}$ has no effect on the proof whenever $k \leq o(N)$. Indeed, a GOE matrix has $\Omega_\varepsilon(N)$ eigenvalues at least $2 - \varepsilon$ with probability $1 - e^{-\Omega_\varepsilon(N^2)}$. This property implies existence of an eigenvalue at least $2 - \varepsilon$ when a GOE matrix is restricted to any subspace of dimension at least $N - \Omega_\varepsilon(N)$ by the Courant-Fisher theorem. Repeating the proof of [Sub21, Theorem 4] with this minor modification establishes the result. □

The following simple lemma is a slightly more general statement of Proposition 5.7.10. It will be used to extend partial embeddings of ancestor-closed subsets of $V(\mathbb{T})$ to all of $V(\mathbb{T})$.

Lemma 5.7.11. *For any $\varepsilon > 0$ and finite dated rooted tree $\mathbb{T}$, there exist c and N_0 depending on ξ, ε, T such that with probability $1 - e^{-cN}$ the following holds for all $N \geq N_0$.*

For any ancestor-closed subset $U \subseteq V(\mathbb{T})$, let $\iota_U : U \to \mathbb{R}^N$ be a Euclidean embedding. Then there is a Euclidean embedding $\iota : V(\mathbb{T}) \to \mathbb{R}^N$ extending ι_U such that for any $v \in V(\mathbb{T})$ with lowest

U-ancestor $u \in U$,

$$H_N(\iota(v)) \geq H_N(\iota(u)) + N \cdot (\mathsf{ALG}_\xi^{\mathrm{Sp}}(|v|) - \mathsf{ALG}_\xi^{\mathrm{Sp}}(|u|) - \varepsilon). \tag{5.7.12}$$

Proof. The result follows by repeated application of Proposition 5.7.10. Indeed, $V(\mathbb{T})\backslash U$ consists of a disjoint union of subtrees $\mathbb{T}_i$ rooted at vertices $u_1, \dots, u_k$ in U. For each $j \in [k]$, given a Euclidean embedding ι_U^{j-1} of

$$U_{j-1} = U \cup \left(\bigcup_{1 \leq i \leq j-1} \mathbb{T}_i \right),$$

we extend it to $\mathbb{T}_j$ as follows. Let

$$W_j = \mathrm{span}(\iota(u)_{u \in U_{j-1}})^\perp$$

be the orthogonal complement of the span of the already-embedded vertices. Then applying Proposition 5.7.10 with $W = W_j$ and $\boldsymbol{m} = \iota(u_j)$, we obtain a Euclidean embedding of $\mathbb{T}_j$ into $W_j + \iota(u_j)$, which joins with ι_U^{j-1} to form an embedding ι_U^j on U_j. It follows from Lemma 5.7.8 that ι_U^j is again a Euclidean embedding of U_j. Moreover (5.7.11) ensures that (5.7.12) is satisfied for each $v \in \mathbb{T}_j$. Repeating this inductively for each $j \in [k]$ completes the proof. □

5.7.3 Proof of Theorems 27 and 28

We now show that full binary trees are necessary for our results, in the sense that trees $\mathbb{T}$ not containing $\mathbb{T}_D^2$ fail as obstructions to energy $(\mathsf{ALG}_\xi^{\mathrm{Sp}} + \varepsilon_{\xi,D})N$ for some $\varepsilon_{\xi,D} > 0$ independent of $|V(\mathbb{T})|$. The main arguments are devoted to proving Lemma 5.7.12, which implies the two main theorems. Lemma 5.7.12 is proved by induction on D, and a representative case for $D = 1$ is depicted in Figure 5.2.

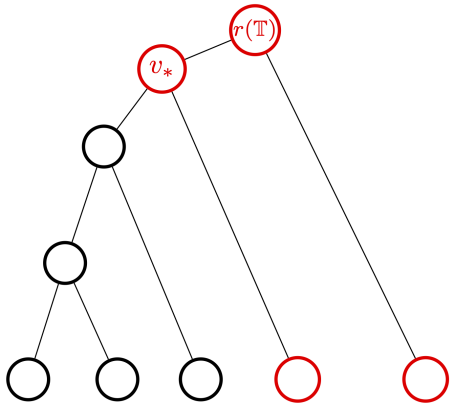


Figure 5.2: A stylized instance of Lemma 5.7.12 in the case $D = 1$ and $[a, b] = [0, 1]$ is displayed. By definition of branching depth, when $D = 1$ the non-leaves of $\mathbb{T}$ consist of a single path. We choose a vertex v_* along this path with small depth $|v_*| = a_2$, and embed v_* to have energy at least $(\mathsf{ALG}(a_2) + 2\varepsilon)N$ using Lemma 5.7.4. The leaves with parent on the segment connecting v_* to $r(\mathbb{T})$ (shown in red) can be embedded one at a time using Lemma 5.7.4. The remaining subtree under v_* is embedded all at once using Proposition 5.7.10. This results in a Euclidean embedding $\iota : V(\mathbb{T}) \to \mathbb{R}^N$ satisfying $H_N(\iota(v)) \geq (\mathsf{ALG} + \varepsilon)N$ for all $v \in L(\mathbb{T})$. For $D > 1$, we repeat this idea recursively.

In the proofs below, we will repeatedly use the principle that $\mathbb{T}$ can be subdivided by placing additional vertices on the edges of $\mathbb{T}$. This only makes constructing an Euclidean embedding more difficult. In particular, we may assume that all leaves have an ancestor of any given height. We sometimes make this explicit by considering the subgraph $\mathbb{T}_{[a,a']}$ of a tree $\mathbb{T}$ with range $[a, b]$, for $a < a' < b$. Precisely, $\mathbb{T}_{[a,a']}$ is the subgraph of vertices with heights in $[a, a']$, where we implicitly assume via subdivision that each leaf in $L(\mathbb{T})$ has ancestors at heights exactly a and a'. Note that unless $a = 0$, $\mathbb{T}_{[a,a']}$ is not in general a tree but is a disjoint union of dated rooted trees each with range $[a, a']$. We similarly define $\mathbb{T}_{[a]}$ to consist of the subset of $V(\mathbb{T})$ at heights exactly a, which without loss of generality contains exactly one ancestor of each leaf of $\mathbb{T}$.

Lemma 5.7.12. *Fix a mixture ξ, and suppose $\frac{\mathrm{d}^2}{\mathrm{d}t^2}(\xi''(t)^{-1/2}) > 0$ for $t \in [a, b] \subseteq [0, 1]$. Fix $D \in \mathbb{N}$ and sufficiently small constants $c, \varepsilon > 0$ depending only on ξ, a, b and D. Then for any $a_1 \in \left[a, \frac{a+b}{2}\right]$, any $k \in \mathbb{N}$, and any dated rooted tree $\mathbb{T}$ with range $[a, b]$, with probability $1-O(e^{-cN})$ over the random choice of H_N, the following holds.*

For any $\boldsymbol{m} \in \sqrt{a_1} \cdot S_N$ and any linear subspace $W \subseteq \mathbb{R}^N$ with $\dim(W) \geq N - k$, there exists a Euclidean embedding

$$\iota : V(\mathbb{T}) \to W + \boldsymbol{m}$$

with $\iota(r(\mathbb{T})) = \boldsymbol{m}$ such that for all $v \in L(\mathbb{T})$,

$$H_N(\iota(v)) \geq H_N(\boldsymbol{m}) + (\mathsf{ALG}_\xi^{\mathrm{Sp}}(b) - \mathsf{ALG}_\xi^{\mathrm{Sp}}(a_1)) + \varepsilon)N. \tag{5.7.13}$$

Proof. We proceed by induction on D.

Base Case In the base case $D = 0$, the tree $\mathbb{T}$ contains only a single leaf v. It then suffices to find a single point $\boldsymbol{\sigma} \in W + \boldsymbol{m}$ with $\|\boldsymbol{\sigma}\|_N^2 = b$ such that

$$H_N(\boldsymbol{\sigma}) \geq H_N(\boldsymbol{m}) + (\mathsf{ALG}_\xi^{\mathrm{Sp}}(b) - \mathsf{ALG}_\xi^{\mathrm{Sp}}(a') + \varepsilon)N.$$

Indeed such a $\boldsymbol{\sigma}$ exists by Lemma 5.7.4.

Inductive Step For $D \geq 1$, assume the result holds for branching depths up to $D - 1$. Our strategy is to first embed the path V_D (recall (5.7.9)), and then apply the inductive hypothesis on the remainder of $\mathbb{T}$ to complete the embedding. We will assume in the remainder of the proof that H_N is $C_1\sqrt{N}$-Lipschitz with respect to the $\|\cdot\|_2$ norm for some constant C_1 as in Proposition 5.2.3.

Define $a_2, a_3 \in \left[a_1, \frac{3a+b}{4}\right]$ such that

$$\sqrt{a_3^2 - a_2^2} = \sqrt{a_2^2 - a_1^2} \leq \frac{\varepsilon}{4C_1}.$$

Let $t = \max_{v \in V_D} |v|$ denote the maximum height of any $v \in V_D$. (Note that t is not affected by adding extraneous vertices to $\mathbb{T}$.)

Case 1: $t \leq a_2$ Let $v_t \in V_D$ satisfy $|v_t| = t$. Take

$$\iota : V_D :\to W + \boldsymbol{m}$$

to be an arbitrary Euclidean embedding with codomain $W + \boldsymbol{m}$. Without loss of generality, we may assume that the children of each $v \in V_D$ have height at most a_3. Next, extend ι to a still arbitrary Euclidean embedding on Q_D, which consists of V_D together with all children of vertices in V_D.

For each vertex $v \in Q_D$, the Lipschitz property implies

$$\begin{aligned} H_N(\iota(v)) &\geq H_N(\boldsymbol{m}) - C_1\sqrt{a_3^2 - a_1^2}N \\ &\geq H_N(\boldsymbol{m}) - \varepsilon_1 N. \end{aligned}$$

Observe that any $v \in Q_D \backslash V_D$ satisfies $D(v) \leq D-1$. Because of this, we can now apply the inductive hypothesis to each subtree $\mathbb{T}_v$ rooted at some $v \in Q_D \backslash V_D$ in an arbitrary order over the v's. More

precisely, suppose a Euclidean embedding mapping a subset $U \subseteq V(\mathbb{T})$ to $W + \boldsymbol{m}$ is given, and that U contains no strict descendants of $v \in Q_D \backslash V_D$. Then we know that $|v| \le a_3 \le \frac{3a+b}{4}$. Define the affine subspace

$$W_v = \text{span}(\iota(u)_{u \in U})^{\perp}.$$

Then by the inductive hypothesis (using the same values a, b), there exists ε_2 depending only on $\xi, a, b, D-1$ such that ι extends to a Euclidean embedding

$$\iota : V \cup \mathbb{T}_v \to W + \boldsymbol{m}$$

such that $\iota(u) \in W_v$ for all $u \in \mathbb{T}_v$, and which satisfies

$$H_N(\iota(u)) \ge H_N(\iota(v)) + \left(\mathsf{ALG}_{\xi}^{\text{Sp}}(b) - \mathsf{ALG}_{\xi}^{\text{Sp}}(|v| + \varepsilon_2)\right) N, \qquad \forall u \in L(\mathbb{T}_v).$$

In particular, the above procedure can be repeated for each v, resulting in an embedding ι defined on all of $V(\mathbb{T})$. Finally for any $u \in L(\mathbb{T})$, we must have $u \in L(\mathbb{T}_v)$ for some v as above, and so

$$\begin{aligned} H_N(\iota(u)) &\ge H_N(\iota(v)) + (\mathsf{ALG}_{\xi}^{\text{Sp}}(b) - \mathsf{ALG}_{\xi}^{\text{Sp}}(|v| + \varepsilon_2))N \\ &\ge H_N(\boldsymbol{m}) + (\mathsf{ALG}_{\xi}^{\text{Sp}}(b) - \mathsf{ALG}_{\xi}^{\text{Sp}}(|v|) + \varepsilon_2 - \varepsilon_1)N \\ &\ge H_N(\boldsymbol{m}) + (\mathsf{ALG}_{\xi}^{\text{Sp}}(b) - \mathsf{ALG}_{\xi}^{\text{Sp}}(a_3) + \varepsilon_1)N + \left(\varepsilon_2 - 2\varepsilon_1 - \xi'(1)\sqrt{a_3^2 - a_1^2}\right) N. \end{aligned}$$

Note that

$$2\varepsilon_1 + \xi'(1)\sqrt{a_3^2 - a_1^2} \le \varepsilon_1 \cdot \left(2 + \frac{\xi'(1)}{C_1}\right).$$

Since ε_2 depended only on ξ, a, b, D and ε_1 was chosen sufficiently small depending on the same parameters, we may assume that

$$\varepsilon_2 - 2\varepsilon_1 - \xi'(1)\sqrt{a_3^2 - a_1^2} \ge 0.$$

Choosing $\varepsilon = \varepsilon_1$ finishes Case 1 of the inductive step.

Case 2: $t \ge a_2$ Let $v_* \in V(\mathbb{T})$ denote the unique vertex on V_D at height a_2 – such a v_* exists without loss of generality. Then applying Lemma 5.7.4 on $\mathbb{T}_{[a_1, a_2]}$, it follows that there exists $\boldsymbol{\sigma} \in W + \boldsymbol{m}$ with $||\boldsymbol{\sigma}||_N^2 = a_2$ such that

$$H_N(\boldsymbol{\sigma}) \ge H_N(\boldsymbol{m}) + (\mathsf{ALG}_{\xi}^{\text{Sp}}(a_2) - \mathsf{ALG}_{\xi}^{\text{Sp}}(a_1) + \varepsilon_2)N$$

for some ε_2 depending only on ξ, a, b. Set $\iota(v_*) = \boldsymbol{\sigma}$. Next we apply Proposition 5.7.10 to the subtree $\mathbb{T}_{v_*}$ rooted at v_*, obtaining a Euclidean embedding

$$\iota : V_D \cup \mathbb{T}_{v_*} \to W + \boldsymbol{m}$$

such that

$$H_N(\iota(x)) \geq H_N(\boldsymbol{m}) + (\mathsf{ALG}_\xi^{\mathrm{Sp}}(a_2) - \mathsf{ALG}_\xi^{\mathrm{Sp}}(a_1) + \varepsilon_3)N$$

for $\varepsilon_3 = \varepsilon_2/2$. Extending to ι to the remainder of $V(\mathbb{T})$ proceeds exactly as in Case 1. □

Below we use Lemma 5.7.12 to show that to rule out energies greater than $\mathsf{ALG}^{\mathrm{Sp}}$, $\mathbb{T}$ must have large branching depth when restricted to any height interval on which $\xi''(t)^{-1/2}$ is convex.

Theorem 27. *Fix ξ and suppose $\frac{\mathrm{d}^2}{\mathrm{d}t^2}(\xi''(t)^{-1/2}) > 0$ for all $t \in [a,b] \subseteq [0,1]$. Fix $D \in \mathbb{N}$ and sufficiently small constants $c, \varepsilon > 0$ depending only on ξ, a, b and D. Then for any dated rooted tree $\mathbb{T}$ with range $[0,1]$ such that every connected component of $\mathbb{T}_{[a,b]}$ has branching depth at most D, with probability $1 - O(e^{-cN})$ over the random choice of H_N, there exists a Euclidean embedding*

$$\iota : V(\mathbb{T}) \to \mathbb{R}^N$$

such that for all $v \in L(\mathbb{T})$,

$$H_N(\iota(v)) \geq (\mathsf{ALG}_\xi^{\mathrm{Sp}}(|v|) + \varepsilon)N. \tag{5.7.14}$$

Proof of Theorem 27. We let $\varepsilon > 0$ be sufficiently small throughout the argument. By Proposition 5.7.10, there exists a Euclidean embedding $\iota : \mathbb{T}_{[0,a]} \to \mathbb{R}^N$ such that for all $v \in \mathbb{T}_{[a]}$,

$$H_N(\iota_a(v)) \geq (\mathsf{ALG}_\xi^{\mathrm{Sp}}(a) - \varepsilon)N. \tag{5.7.15}$$

Here as usual we assume without loss of generality that all leaves in $\mathbb{T}$ have an ancestor at height a. Next we extend ι to a Euclidean embedding

$$\iota : \mathbb{T}_{[0,b]} \to \mathbb{R}^N$$

such that for all $v \in V(\mathbb{T}_{[b]})$ with ancestor u at height a,

$$H_N(\iota(v)) \geq H_N(\iota(u)) + (\mathsf{ALG}_\xi^{\mathrm{Sp}}(b) - \mathsf{ALG}_\xi^{\mathrm{Sp}}(a) + 3\varepsilon)N. \tag{5.7.16}$$

In fact the existence of such an extension follows directly from Lemma 5.7.12 for ε sufficiently small. Here as before we repeatedly apply Lemma 5.7.12 to individual subtrees in $\mathbb{T}_{[a,b]}$, using the subspace W in the statement to ensure the orthogonality constraints are satisfied.

Finally extend ι to $\mathbb{T}_{[b,1]}$ using Lemma 5.7.11 on each component. The result is that for any $x \in L(\mathbb{T})$ with ancestor v at height b,

$$H_N(\iota(x)) \geq H_N(\iota(v)) + (\mathsf{ALG}_\xi^{\mathrm{Sp}}(1) - \mathsf{ALG}_\xi^{\mathrm{Sp}}(b) - \varepsilon)N. \tag{5.7.17}$$

Combining (5.7.15), (5.7.16), and (5.7.17) completes the proof. □

In the Guerra-Talagrand interpolation used for our main argument, it was only possible to directly estimate the *average* energy of the replicas instead of the minimum. In the following final result, we show that to force the average of $H_N(v)$ over the leaves $v \in L(\mathbb{T})$ to be small, it is necessary to use a tree which branches a superconstant amount in any height interval $[a, b]$ as above, *on a set of components of* $\mathbb{T}_{[a,b]}$ *ancestral to almost all leaves.*

Let us illustrate the difference between Theorems 27 and 28 by an example. Form $\mathbb{T}$ by starting with a full symmetric tree as in Proposition 5.3.2 and adding many children of the root as additional leaves. Then by construction (recall also Proposition 5.3.8), with probability $1 - e^{-\Omega(N)}$ any Euclidean embedding $\iota : \mathbb{T} \to \mathbb{R}^N$ satisfies

$$\min_{v \in L(\mathbb{T})} H_N(\iota(v)) \leq (\mathsf{ALG} + \varepsilon)N$$

for $\varepsilon > 0$ as in Proposition 5.3.2 arbitrarily small given ξ. However the same is not true for the average energy. Indeed, Theorem 27 with $D = 1$ implies that the additional leaves in $\mathbb{T}$ can be embedded to each have energy at least $(\mathsf{ALG} + 2\varepsilon')N$ where $\varepsilon' > 0$ depends only on ξ. If the additional leaves form a sufficiently large fraction of $L(\mathbb{T})$, then any Euclidean extension ι to all of $\mathbb{T}$ satisfies

$$\frac{1}{|L(\mathbb{T})|} \sum_{v \in L(\mathbb{T})} H_N(\iota(v)) \geq (\mathsf{ALG} + \varepsilon')N$$

assuming $H_N \in K_N$.

Theorem 28. *Fix a mixture ξ and $\delta > 0$, and suppose $\frac{\mathrm{d}^2}{\mathrm{d}t^2}(\xi''(t)^{-1/2}) > 0$ for $t \in [a, b] \subseteq [0, 1]$. Fix $D \in \mathbb{N}$ and sufficiently small constants $c, \varepsilon > 0$ depending only on ξ, a, b, D and δ. Consider any tree $\mathbb{T}$ with range $[0, 1]$ and $|L(\mathbb{T})| = n$ leaves such that for at least δn of the leaves $v \in |L(\mathbb{T})|$, the subtree of $\mathbb{T}_{[a,b]}$ consisting of ancestors of v has branching depth at most D. With probability $1 - O(e^{-cN})$ over the random choice of H_N, there exists a Euclidean embedding*

$$\iota : V(\mathbb{T}) \to \mathbb{R}^N$$

such that

$$\frac{1}{L(\mathbb{T})} \sum_{v \in L(\mathbb{T})} H_N(\iota(v)) \geq (\mathsf{ALG}_\xi^{\mathrm{Sp}} + \varepsilon)N. \tag{5.7.18}$$

Proof. Take $\varepsilon_0 > 0$ sufficiently small. For $v \in L(\mathbb{T})$ and $t \in [0,1]$, let v^t denote the ancestor of v at height v. As before, Proposition 5.7.10 shows that there exists a Euclidean embedding $\iota : \mathbb{T}_{[0,a]} \to \mathbb{R}^N$ such that for all $v^a \in \mathbb{T}_{[a]}$,

$$H_N(\iota_a(v^a)) \geq (\mathsf{ALG}_\xi^{\mathrm{Sp}}(a) - \delta\varepsilon_0)N. \tag{5.7.19}$$

Let $\widetilde{\mathbb{T}}_{[a,b]} \subseteq \mathbb{T}_{[a,b]}$ consist of all connected components in $\mathbb{T}_{[a,b]}$ of branching depth at most D. Next we extend ι to a Euclidean embedding

$$\iota : \mathbb{T}_{[0,a]} \cup \widetilde{\mathbb{T}}_{[a,b]} \to \mathbb{R}^N$$

such that for all $v^b \in L(\widetilde{\mathbb{T}}_{[a,b]})$ with ancestor v^a at height a,

$$H_N(\iota(v^b)) \geq H_N(\iota(v^a)) + (\mathsf{ALG}_\xi^{\mathrm{Sp}}(b) - \mathsf{ALG}_\xi^{\mathrm{Sp}}(a) + 4\varepsilon_0)N. \tag{5.7.20}$$

Lemma 5.7.12 allows ι to extend to the remainder of $V(\mathbb{T}_{[a,b]})$ such that

$$H_N(\iota(v^b)) \geq H_N(\iota(v^a)) + (\mathsf{ALG}_\xi^{\mathrm{Sp}}(b) - \mathsf{ALG}_\xi^{\mathrm{Sp}}(a) - \delta\varepsilon_0)N. \tag{5.7.21}$$

holds for all $v \in V(\mathbb{T}_{[a,b]})$. Since at least $\delta|L(\mathbb{T})|$ leaves v satisfy $v^a \in \widetilde{\mathbb{T}}_{[a,b]}$, (5.7.20) and (5.7.21) imply

$$\frac{1}{|L(\mathbb{T})|}\sum_{v\in L(\mathbb{T})} H_N(\iota(v^a)) - H_N(\iota(v^b)) \geq (\mathsf{ALG}_\xi^{\mathrm{Sp}}(b) - \mathsf{ALG}_\xi^{\mathrm{Sp}}(a) + 3\delta\varepsilon_0)N \tag{5.7.22}$$

As before we finish by extending ι to $\mathbb{T}_{[b,1]}$, using Lemma 5.7.11 one component at a time. Then for any $v \in L(\mathbb{T})$,

$$H_N(\iota(v)) \geq H_N(\iota(v^b)) + (\mathsf{ALG}_\xi^{\mathrm{Sp}}(1) - \mathsf{ALG}_\xi^{\mathrm{Sp}}(b) - \delta\varepsilon_0)N. \tag{5.7.23}$$

By combining (5.7.19), (5.7.22) and (5.7.23), it follows that the total of $H_N(\iota(v))$ over $v \in L(\mathbb{T})$ is

$$\begin{aligned}\sum_{v\in L(\mathbb{T})}[H_N(\iota(v))] &\geq \sum_{v\in L(\mathbb{T})}\Big(H_N(\iota(v)) - H_N(\iota(v^b)) + H_N(\iota(v^b)) - H_N(\iota(v^a)) + H_N(\iota(v^a))\Big)\\ &\geq |L(\mathbb{T})|\cdot\Big(\mathsf{ALG}_\xi^{\mathrm{Sp}}(1) + \delta\varepsilon_0\Big).\end{aligned}$$

Taking $\varepsilon = \delta\varepsilon_0$ completes the proof. □

Part III

Machine Learning

Chapter 6

Chasing Convex Bodies and Functions

6.1 Introduction

Let X be a d-dimensional normed space and $K_1, K_2, \ldots, K_N \subseteq X$ a finite sequence of convex bodies. In the *chasing convex bodies* problem, a player starting at $x_0 = 0 \in X$ learns the sets K_n one at a time, and after observing K_n moves to a point $x_n \in K_n$. The player's cost is the total path length

$$\text{cost}(x_1, \ldots, x_N) = \sum_{n=1}^{N} ||x_n - x_{n-1}||.$$

Denote the smallest cost (in hind-sight) among all such sequences by

$$\text{cost}(K_1, \ldots, K_N) = \min_{(y_n \in K_n)_{n \leq N}} \sum_{n=1}^{N} ||y_n - y_{n-1}||.$$

The player's goal is to ensure that

$$\text{cost}(x_1, \ldots, x_N) \leq \alpha_d \cdot \text{cost}(K_1, \ldots, K_N) \tag{6.1.1}$$

holds for any sequence $K_1, \ldots, K_N$, where α_d is as small as possible and is independent of N. The challenge is that the points $x_n = x_n(K_1, \ldots, K_n)$ must depend only on the sets revealed so far. To encapsulate this requirement we say the player's path must be *online*, as opposed to the optimal *offline* path which can depend on future information. An online algorithm achieving (6.1.1) for some finite α_d is said to be α_d-*competitive*, and the smallest possible α_d among all online algorithms is

the *competitive ratio* of chasing convex bodies.

In the most general sense, the problem of asking a player to choose an online path $x_1, \ldots, x_N$ through a sequence of subsets $S_1, \ldots, S_N$ in a metric space $\mathcal{X}$ is known as *metrical service systems*. These sets are typically called "requests". When arbitrary subsets $S_i \subseteq \mathcal{X}$ can be requested, the competitive ratio possible is $|\mathcal{X}|-1$ in any metric space [MMS90]. One also considers the slightly more general *metrical task systems* problem in which requests are non-negative cost functions $f_n : \mathcal{X} \to \mathbb{R}^+$ rather than sets and the cost takes the form

$$\text{cost}(x_1, \ldots, x_N) = \sum_{n=1}^{N} d_{\mathcal{X}}(x_n, x_{n-1}) + f_n(x_n)$$

where $\sum_{n=1}^{N} d_{\mathcal{X}}(x_n, x_{n-1})$ is called the *movement cost* while $\sum_{n=1}^{N} f_n(x_n)$ is the *service cost*. As in (6.1.1), one aims to ensure

$$\text{cost}(x_1, \ldots, x_N) \leq \alpha \cdot \text{cost}(f_1, \ldots, f_n) = \alpha \cdot \inf_{(y_n \in \mathcal{X})_{n \leq N}} \text{cost}(y_1, \ldots, y_N). \tag{6.1.2}$$

The competitive ratio of metrical task systems is always $2|\mathcal{X}|-1$ [BLS92]. Actually both competitive ratios just stated are for deterministic algorithms; one may also allow external randomness, so that one chooses $x_n = x_n(S_1, \ldots, S_n, \omega)$ for some random variable ω independent of the sets S_i. One then aims for the same guarantee as in (6.1.1), (6.1.2) with the expected cost of the player on the left-hand side, for any fixed sequence $(S_1, \ldots, S_N)$. With randomization the competitive ratio of metrical task or service systems sharply drops and is known to be in the range $\left[\frac{c_1 \log|\mathcal{X}|}{\log\log|\mathcal{X}|}, c_2 (\log|\mathcal{X}|)^2\right]$, and to be $\Theta(\log|\mathcal{X}|)$ in some specific cases [BLMN05, BBM06, FM03, BCLL19]. However this is not the end of the story as a wide range of problems, including chasing convex bodies, result from restricting which subsets are allowed as requests. The literature on such problems is vast and includes scheduling [Gra66], self-organizing lists [ST85], efficient covering [AAA03], safely using machine-learned advice [BB00, KPS18, LV18b, WZ20], and the famous k-server problem [MMS90, Gro91, KP95, BBMN15].

Chasing convex bodies was proposed in [FL93] to study the interaction between convexity and metrical task systems. Of course the general upper bounds above are of no use as $|X| = \infty$, while the lower bounds also do not apply due to the convexity constraint. [FL93] gave an algorithm with finite competitive ratio for the already non-trivial $d = 2$ case and conjectured that the competitive ratio is finite for any $d \in \mathbb{N}$. The best known asymptotic lower bounds come from requesting the faces of a hypercube by taking $K_n = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) \times [-1, 1]^{d-n}$ for $\varepsilon_i \in \{-1, 1\}$ uniformly random and $n \leq d$. This construction implies that the competitive ratio is at least $\sqrt{d}$ in Euclidean space and at least d for $X = \ell^\infty$ - see [BKL$^+$20, Lemma 5.4] for more on lower bounds. Unlike in many competitive analysis problems, randomization is useless for chasing convex bodies and we may freely restrict attention to deterministic algorithms. This is because $\text{cost}(x_1, \ldots, x_N)$ is convex on X^N, and so randomized paths are no better than their (deterministic) pointwise expectations.

Following a lack of progress on the full conjecture, restricted cases such as chasing subspaces were studied, e.g. [ABN$^+$16]. A notable restriction is chasing *nested* convex bodies, where the convex sets $K_1 \supseteq K_2 \supseteq \dots$ are required to decrease. Nested chasing was introduced in [BBE$^+$18] and solved rather comprehensively in [ABC$^+$19] and then [BKL$^+$20]. The latter work gave an algorithm with optimal competitive ratio up to $O(\log d)$ factors for all ℓ^p spaces based on Gaussian-weighted centroids. Moreover it gave a d-competitive memoryless algorithm based on the Steiner point which we discuss later.

Some time after chasing convex bodies was posed, an equivalent problem called *chasing convex functions* emerged. This is a metrical task systems problem in which requests are convex functions $f_n : X \to \mathbb{R}^+$ instead of convex sets. As described above the total cost

$$\text{cost}(x_1, \dots, x_N) = \sum_{n=1}^{N} ||x_n - x_{n-1}|| + f_n(x_n)$$

decomposes as a movement cost plus a service cost. Chasing convex functions subsumes chasing convex bodies by replacing the body K_n with the function $f_n = 2 \cdot d(x, K_n)$. This is because an arbitrary algorithm for the requests f_n is improved by projecting x_n onto K_n - actually the same argument shows more generally that metrical task systems subsumes metrical service systems. Conversely as observed in [BLLS19], convex function chasing in X can be reduced to convex body chasing in $X \oplus \mathbb{R}$ up to a constant factor by alternating requests of the epigraphs $\{(x, y) \in X \times \mathbb{R} : y \geq f_n(x)\}$ with the hyperplane $X \times \{0\}$. As with chasing convex bodies, randomized algorithms are no better than deterministic algorithms since $\text{cost}(x_1, \dots, x_N)$ remains convex on X^N.

Convex function requests allow one to model many practical problems. Indeed chasing convex functions was originally considered as a model for efficient power management in cooling data centers [LWAT13]. In light of this, restricted or modified versions of chasing convex function have also been studied. For example, [BGK$^+$15] determines the exact competitive ratio in 1 dimension, while works such as [CGW18, GLSW19] show dimension-independent competitive ratios for similar problems with further restrictions on the cost functions.

6.1.1 Main Result

In prior joint work with S. Bubeck, Y.T. Lee, and Y. Li [BLLS19] we gave the first algorithm achieving a finite competitive ratio for convex body chasing. Unfortunately this algorithm used an induction on dimension that led to a exponential competitive ratio $2^{O(d)}$. We give an upper bound of d for the competitive ratio of chasing convex bodies in a general normed space, which is tight for ℓ^∞. In Euclidean space, our algorithm has competitive ratio $O(\sqrt{d \log N})$, nearly matching the lower bound $\sqrt{d}$ when the number of requests N is sub-exponential in d. The statement following combines Theorems 31 and 32.

Theorem 29. *In any d-dimensional normed space there is a $d+1$ competitive algorithm for chasing convex functions and a d competitive algorithm for chasing convex bodies. Moreover in Euclidean space this algorithm is $O(\sqrt{d \log N})$-competitive.*

The proof is inspired by our joint work with S. Bubeck, B. Klartag, Y.T. Lee, and Y. Li [BKL$^+$20] on chasing nested convex bodies. It is shown there that moving to the new body's *Steiner point*, a stable center point of any convex body defined in [Ste40], gives total movement at most d starting from the unit ball in d dimensions. (The argument in [BKL$^+$20] is restricted to Euclidean space but the proof works in general as we will explain.) We extend their argument by defining the *functional Steiner point* of a convex function. Our algorithm follows the functional Steiner point of the so-called *work function* which encodes at any time the effective total cost of all requests so far.

We remark that given the form of (6.1.1), chasing convex bodies may be viewed as an online version of a Lipschitz selection problem. In the broadest generality, for some family $\mathcal{S} \subseteq 2^X$ of subsets of a set X, a selector takes sets $S \in \mathcal{S}$ to elements $s \in S$. Of course the relevant comparison for us is when $\mathcal{S}$ consists of all convex bodies in X. Continuity and Lipschitz properties of general selectors have received significant attention [Shv84, Shv02, Kup05, FS18]. Taking the Hausdorff metric on convex sets, the Steiner point is d-Lipschitz in any normed space. Moreover as explained in [PY89, Section 4], it achieves the exact optimal Lipschitz constant (of order $\Theta(\sqrt{d})$) when X is Euclidean due to a beautiful symmetrization argument. We find it appealing that this in some sense optimal Hausdorff-Lipschitz selector also solves an online version of Lipschitz selection.

Concurrently with this work, C.J. Argue, A. Gupta, G. Guruganesh, and Z. Tang obtained similar results for chasing convex bodies in Euclidean space [AGGT21]. Their algorithm is based on Steiner points of level sets of the work function; these turn out to be almost the same as the functional Steiner point as we show in Section 6.6.

6.2 Problem Setup

6.2.1 Notations and Conventions

The variables T, t, s denote real times while N, n denote integer times. $⨍_{x \in S} f(x)\mathrm{d}x$ denotes the average value $\frac{\int_{x \in S} f(x)\mathrm{d}x}{\int_{x \in S} 1\mathrm{d}x}$ of $f(x)$ on the set S. Denote by $B_1 \subseteq X$ the unit ball and $B_1^* \subseteq X^*$ the dual unit ball. The symbol ∂ denotes boundary, and $\langle \cdot, \cdot \rangle$ denotes the natural pairing between X, X^*.

6.2.2 Continuous Time Formulation

Our proof is more natural in continuous time, so we first solve the problem in this setting and then specialize to discrete time. In continuous time chasing convex functions, we receive a locally bounded family of non-negative convex functions $(f_t : X \to \mathbb{R}^+)_{t\in[0,T]}$. We assume that $f_t(x)$ is piece-wise continuous in t with a locally finite set of continuities. The player constructs a bounded variation path (x_t) online, so that x_s depends only on $(f_t)_{t\leq s}$. We will assume f_t and x_t are cadlag (right-continuous with left-limits) in the time variable t. The cost is again the sum of movement and service costs given by

$$\text{cost}((x_t)_{t\in[0,T]}) = \int_0^T f_t(x_t) + ||x_t'||\mathrm{d}t.$$

Here and throughout, the integral of $||x_t'||$ is understood to mean the total variation of the path x_t. As before the goal is to achieve a small competitive ratio against the optimal offline path. Given a sequence $f_1, f_2, \ldots, f_N$ of convex requests, one readily obtains a corresponding continuous-time problem instance by choosing, for each $t \in [0, N]$, the function $f_t = f_n$ for $t \in (n-1, n]$. The next proposition shows that solving this continuous problem suffices to solve the discrete problem.

Proposition 6.2.1. *Any discrete-time instance of chasing convex function has the same offline optimal cost as its continuous-time counterpart. Meanwhile for any continuous-time online algorithm there exists a discrete-time online algorithm achieving both smaller movement and smaller service cost on every sequence of functions $f_1, \ldots, f_N$.*

Proof. It is easy to see that the continuous and discrete time problems have the same offline optimum value. Given a solution for continous-time convex function chasing, suppose the player sees a discrete time request f_n. The player then computes the continuous time path $(x_t)_{t\in(n-1,n]}$ and moves to some x_{t_n} with $t_n \in (n-1, n]$ and

$$f_n(x_{t_n}) \leq \int_{n-1}^{n} f_n(x_t)\mathrm{d}t.$$

The discretized sequence $(x_{t_1}, \ldots, x_{t_N})$ has a smaller movement cost than the continuous path $(x_t)_{t\in[0,T]}$ because the triangle inequality implies

$$\begin{aligned}\sum_{n=1}^{N} ||x_{t_n} - x_{t_{n-1}}|| &\leq \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} ||x_s'||\mathrm{d}s \\ &= \int_0^{t_N} ||x_s'||\mathrm{d}s \\ &\leq \int_0^{N} ||x_s'||\mathrm{d}s.\end{aligned}$$

The discretized path also has smaller service cost by construction, hence the result.

□

6.3 Functional Steiner Point and Work Function

We begin by recalling the definition of the Steiner point in a d-dimensional normed space X. For a convex body $K \subseteq X$ and $v \in X^*$, define

$$f_K(v) = \arg\max_{x \in K} \langle v, x \rangle,$$
$$h_K(v) = \max_{x \in K} \langle v, x \rangle = \langle f_K(v), x \rangle.$$

h_K is commonly known as the *support function* of K. Let μ denote the cone measure on ∂B_1^*, which can be sampled from by choosing a uniformly random $z \in B_1^*$ and normalizing to $\theta = \frac{z}{||z||}$. For $\theta \in \partial B_1^*$ define $n(\theta) \in X$ to be the outward unit normal defined (for μ-almost all θ) by $||n(\theta)|| = 1$ and $\langle n(\theta), \theta \rangle = 1$.

Definition 6.3.1 ([PY89, Chapter 6])**.** *The Steiner point $s(K) \in X$ is*

$$s(K) = \fint_{v \in B_1^*} f_K(v) dv. \tag{6.3.1}$$
$$= d \fint_{\theta \in \partial B_1^*} h_K(\theta) n(\theta) d\mu(\theta). \tag{6.3.2}$$

The equivalence of the two definitions follows from the divergence theorem and the identity $\nabla h_K = f_K$. The factor d comes from the discrepancy in total measure of the ball and the sphere. See [PY89, Chapter 6] for a careful derivation.

Using Definition 6.3.1, the upper bound d for nested chasing in [BKL$^+$20] immediately extends to any normed space. We recall the main result here. It is not phrased as a competitive ratio because some apriori reductions are possible in nested chasing — roughly speaking we stay inside the unit ball B_1 and treat the offline optimum cost as being 1. Note that both (6.3.1) and (6.3.2) are essential in the argument below.

Theorem 30. *[BKL$^+$20, Theorem 2.1] Let $B_1 \supseteq K_1 \supseteq K_2 \supseteq \cdots \supseteq K_N$ be convex bodies in X, with $x_n = s(K_n)$ for each n. Then $x_n \in K_n$ for each n and*

$$\sum_{n=1}^{N} ||x_n - x_{n-1}|| \leq d.$$

Proof. It follows from (6.3.1) that $s(K) \in K$, so it remains to estimate the total movement. For

convenience take $K_0 = B_1$ the unit ball so that $x_0 = (0, 0, \ldots, 0) = s(K_0)$. From $K_n \subseteq K_{n-1}$ it follows that $h_{K_n}(\theta) \leq h_{K_{n-1}}(\theta)$ for each $n \leq N$ and $\theta \in \partial B_1^*$. Combining with (6.3.2) yields:

$$\begin{aligned}
\sum_{n=1}^{N} ||s(K_n) - s(K_{n-1})|| &\leq d \fint_{\theta \in \partial B_1^*} \sum_{n=1}^{N} |h_{K_n}(\theta) - h_{K_{n-1}}(\theta)| \mathrm{d}\mu(\theta) \\
&= d \fint_{\theta \in \partial B_1^*} \sum_{n=1}^{N} h_{K_{n-1}}(\theta) - h_{K_n}(\theta) \mathrm{d}\mu(\theta) \\
&= d \fint_{\theta \in \partial B_1^*} 1 - h_{K_N}(\theta) \mathrm{d}\mu(\theta) \\
&\leq d.
\end{aligned}$$

Here the last inequality follows from $h_{K_N}(\theta) + h_{K_N}(-\theta) \geq 0$.

□

We now extend the definition of Steiner point to convex functions. The idea is to replace the support function by the concave conjugate (also known as the Fenchel-Legendre transform). Recall that for a convex function $W : X \to \mathbb{R}^+$, the concave conjugate $W^* : X^* \to \mathbb{R} \cup \{-\infty\}$ is defined by

$$W^*(v) = \inf_{w \in X} \left(W(w) - \langle v, w \rangle \right) \tag{6.3.3}$$

Let us assume W is not only convex but also 1-Lipschitz, and that $W(w) - ||w||$ is uniformly bounded. We will refer to such a W as an (abstract) work function. Note $W^*(v)$ is finite whenever $||v|| < 1$ by the last assumption, and moreover the infimum in (6.3.3) is attained. We denote this point attaining this infimum by

$$v^* = \arg\min_{w \in X} \left(W(w) - \langle v, w \rangle \right),$$

the conjugate point to v with respect to W. It satisfies $\nabla W(v^*) = v$ and is well-defined for almost every $v \in B_1^*$ by Alexandrov's theorem. Moreover we have $\nabla W^*(v) = -v^*$. Combining this latter relation with the divergence theorem yields another identity, from which the functional Steiner point is defined.

Definition 6.3.2. *Let X be an arbitrary d-dimensional normed space, and $W : X \to \mathbb{R}^+$ a work function as defined above. The functional Steiner point $s(W) \in X$ is:*

$$s(W) = \fint_{v \in B_1^*} v^* dv. \tag{6.3.4}$$

$$= -d \fint_{\theta \in \partial B_1^*} W^*(\theta) n(\theta) d\mu(\theta). \tag{6.3.5}$$

We remark that if a convex body K is identified with the function $f(x) = d(x, K)$, then the

definitions above agree. We call (6.3.1), (6.3.4) the *primal* definitions and (6.3.2), (6.3.5) the *dual* definitions.

6.3.1 The Work Function

The work function is a central object in online algorithms; in general it records the smallest cost required to satisfy an initial sequence of requests while ending in a given state. Work function based algorithms are essentially optimal among deterministic algorithms for general metrical task systems [BLS92] as well as the k-server problem [KP95].

Definition 6.3.3. *Given requests $(f_s)_{s\le t}$, the work function $W_t(x)$ is the offline-optimal cost among paths satisfying $x_t = x$:*

$$W_t(x) = \inf_{\substack{x_s:[0,t]\to X \\ x_t = x}} ||x_0|| + \int_0^t f_s(x_s) + ||x_s'||ds \tag{6.3.6}$$

$$= \inf_{\substack{x_s:[0,t]\to X \\ x_t = x}} cost_t(x_s). \tag{6.3.7}$$

Here we allow $x_s : [0,t] \to X$ to be any path of bounded variation, and as before interpret $\int_0^t ||x_s'||ds$ to mean the total variation of the path. Likewise for a discrete-time request sequence $(f_1, \dots, f_n)$, the work function $W_n(x)$ is defined as above with $f_t = f_n$ for $t \in (n-1, n]$ or more simply by

$$W_n(x) = \min_{x_1,\dots,x_n \in X} ||x - x_n|| + \sum_{n=1}^{N} ||x_n - x_{n-1}|| + f_n(x_n).$$

For a sequence $(K_1, \dots, K_n)$ of convex set requests the work function W_n is defined analogously.

In the case that $f_s(x)$ is piecewise constant in s (which is all we need for the original discrete-time problem), the best offline continuous time strategy clearly coincides with the best offline discrete time strategy. The infimum is attained in (6.3.7) in general because the paths $(x_s)_{s\le t}$ of variation at most C satisfying $x_t = T$ are compact in the usual topology on cadlag functions for any C, and cost_t is lower semicontinuous.

Denote by $W_t^*(\cdot)$ the concave conjugate of W_t, and v_t^* the point with $\nabla W_t(v_t^*) = v$. We record the following proposition summarizing the properties of the work function and its dual.

Proposition 6.3.4. *In either discrete or continuous time, W_t and W_t^* satisfy:*

1. $W_0(x) = ||x||$.
2. $W_0^*(v) = 0$ *whenever* $||v|| \le 1$.

3. *$W_t(x)$ is increasing in t and is convex for all fixed t.*

4. *$W_t^*(x)$ is increasing in t and concave in x.*

5. *$W_t(x)$ is an abstract work function.*

6. *$W_t^*(v)$ is non-negative and finite whenever $||v|| \leq 1$.*

7. *cost$((f_s)_{s\in[0,t]}) = \min_{x\in X} W_t(x)$.*

Proof. It is clear that $W_0(x) = ||x||$, and that $W_t(x)$ is increasing in t. The computation of W_0^* is clear. Convexity of $W_t(\cdot)$ holds by convexity of $\text{cost}_t(\cdot)$ — given paths $x_s^0 : [0,t] \to X$ and $x_s^1 : [0,t] \to X$ the path $x_s^q : [0,t] \to X$ given by

$$x_s^q = qx_s^1 + (1-q)x_s^0$$

satisfies for any $q \in [0,1]$,

$$\text{cost}_t(x_s^q) \leq q \cdot \text{cost}_t(x_s^1) + (1-q) \cdot \text{cost}_t(x_s^0).$$

Convexity of W_t implies that W_t^* is concave by general properties of the Fenchel-Legendre transform. Because W_t is increasing in t, the definition (6.3.3) implies that W_t^* is increasing in t as well. It is easy to see that W_t is 1-Lipschitz; to show

$$W_t(x) \leq W_t(y) + ||x - y||$$

it suffices to take the lowest cost path to y and then move from y to x. Similarly $W_t(x) - ||x||$ is bounded, making W_t an abstract work function. It follows from this that $W_t^*(v)$ is finite when $||v|| \leq 1$.

□

Lemma 6.3.5. *For all t,*

$$\max_{||\theta||\leq 1} W_t^*(\theta) \leq 2 \cdot \min_x W_t(x),$$
$$\fint_{\theta\in\partial B_1^*} W_t^*(\theta)\, d\mu(\theta) \leq \min_x W_t(x),$$
$$\fint_{v\in B_1} W_t^*(v)\, dv \leq \min_x W_t(x).$$

Proof. Set

$$OPT_t = \arg\min_x W_t(x).$$

The definition (6.3.3) of W_t^* implies

$$W_t^*(\theta) \leq W_T(OPT_t) - \theta \cdot OPT_t.$$

Finally

$$\begin{aligned} |W_t(OPT_t)| &= \inf_{\substack{x_s:[0,t]\to X \\ x_t = OPT_t}} ||x_0|| + \int_0^t f_s(x_s) + ||x_s'|| \mathrm{d}s \\ &\geq \inf_{\substack{x_s:[0,t]\to X \\ x_t = OPT_t}} ||x_0|| + \int_0^t ||x_s'|| \mathrm{d}s \\ &\geq |OPT_t| \end{aligned}$$

holds where the triangle inequality was used in the last line. All assertions now follow.

□

We next compute the time derivative of $W_t^*(v)$ for fixed v with $|v| < 1$. The proof, a simple exercise, is left to the appendix.

Lemma 6.3.6. *For any $\varepsilon > 0$ suppose $f_s(x)$ is jointly continuous in (s, x) and convex in x for $(s, x) \in [t, t + \varepsilon) \times X$. Then for almost all v with $||v|| < 1$,*

$$\frac{d}{dt} W_t^*(v) = f_t(v_t^*).$$

6.4 Linear Competitive Ratio

Our algorithm for continuous-time convex function chasing is defined by setting $x_t = s(W_t)$. In its analysis, the primal definition (6.3.4) controls the service cost while the dual definition (6.3.5) controls the movement cost.

Theorem 31. *$x_t = s(W_t)$ is $d + 1$ competitive for continuous-time convex function chasing in any d-dimensional normed space X. In particular:*

1. *The movement cost of x_t is d-competitive:*

$$\int_0^T ||x_t'|| \, dt \leq d \cdot \min_x W_t(x).$$

2. *The service cost of x_t is 1-competitive:*

$$\int_0^T f_t(x_t)\,dt \leq \min_x W_t(x).$$

Proposition 6.2.1 yields an induced algorithm for chasing bodies/functions in discrete time which we call the discrete-time functional Steiner point.

Corollary 6.4.1. *The discrete-time functional Steiner point is $d+1$ competitive for chasing convex functions and d competitive for chasing convex bodies.*

Proof of Corollary 6.4.1. This follows from Proposition 6.2.1 and the fact that chasing convex bodies has 0 service cost.

□

Proof of Theorem 31. We begin with part 1. From the dual definition (6.3.5) of $s(W_t)$ and the fact that W_t^* increases with t from $W_0^* = 0$,

$$\begin{aligned}\int_0^T ||x_t'||\mathrm{d}t &= d \cdot \int_0^T \left|\left| \frac{\mathrm{d}}{\mathrm{d}t} \fint_{\theta\in\partial B_1^*} W_t^*(\theta)\theta \mathrm{d}\mu(\theta) \right|\right| \\ &\leq d \cdot \int_0^T \fint_{\theta\in\partial B_1^*} \left| \frac{\mathrm{d}}{\mathrm{d}t} W_t^*(\theta) \right| \mathrm{d}\mu(\theta) \\ &= d \cdot \fint_{\theta\in\partial B_1^*} W_T^*(\theta)\mathrm{d}\mu(\theta).\end{aligned}$$

Lemma 6.3.5 implies

$$d \cdot \fint_{\theta\in\partial B_1^*} W_T^*(\theta)\mathrm{d}\mu(\theta) \leq d \min_x W_T(x).$$

This completes the proof of part 1 and we turn to part 2. From the primal definition (6.3.4) and convexity of f_t it follows that

$$f_t(s(W_t)) \leq \fint_{v\in B_1^*} f_t(v_t^*)\mathrm{d}v.$$

Integrating in time and using Lemmas 6.3.6 and 6.3.5 yields:

$$\begin{aligned}\int_0^T f_t(s(W_t))\mathrm{d}t \leq & \fint_{v\in B_1}\int_0^T f_t(v_t^*)\mathrm{d}t\mathrm{d}\mu(\theta)\\ =&\fint_{v\in B_1^*}\int_0^T \frac{\mathrm{d}}{\mathrm{d}t}W_t^*(v)\mathrm{d}t\mathrm{d}v\\ =&\fint_{v\in B_1^*} W_T^*(v)-W_0^*(v)\mathrm{d}v\\ =&\fint_{v\in B_1^*} W_T^*(v)\mathrm{d}v\\ \leq &\min_x W_T(x).\end{aligned}$$

□

Remark 6.4.1. In the continuous time setting, only $f_t(x_t)$ and $\nabla f_t(x_t)$ are actually necessary to solve convex function chasing. This is because the player can always lower bound f_t by

$$f_t(x) \geq \tilde{f}_t(x) \equiv \max\left(f_t(x_t) + \langle \nabla f_t(x_t), x - x_t\rangle, 0\right).$$

As $\tilde{f}_t(x_t) = f_t(x_t)$, by simply pretending the requests are $\tilde{f}_t$, any competitive algorithm can be transformed into one which only uses the values $f_t(x_t)$ and $\nabla f_t(x_t)$ and which obeys the same guarantees.

In the discrete time setting, if we are given $f_n(x_{n-1})$ and $\nabla f_n(x_{n-1})$ before choosing x_n, there is another source of error because $f_n(x_n)$ is totally unknown. However this error is easily controlled when the f_n are uniformly Lipschitz. Let $(x_n)_{n\leq N}$ be the discrete-time functional Steiner point sequence for the functions recursively defined by

$$\tilde{f}_n(x) = \max\left(f_n(x_{n-1}) + \langle \nabla f_n(x_{n-1}), x - x_{n-1}\rangle, 0\right)$$

and let W_N be the discrete-time work function. We obtain:

$$\begin{aligned}\sum_{n=1}^N f_n(x_n) + ||x_n - x_{n-1}|| &\leq \sum_{n=1}^N \tilde{f}_n(x_n) + ||x_n - x_{n-1}|| + \left(\sum_{n=1}^N f_n(x_n) - \tilde{f}_n(x_n)\right)\\ &\leq (d+1)\min_x W_N(x) + \left(\sum_{n=1}^N f_n(x_n) - \tilde{f}_n(x_n)\right).\end{aligned}$$

Suppose now that each f_n is L-lipschitz. Then the equality $f_n(x_{n-1}) = \tilde{f}_n(x_{n-1})$ implies

$|f_n(x_n) - \tilde{f}_n(x_n)| \leq 2L||x_n - x_{n-1}||$. Because Theorem 31 and Proposition 6.2.1 imply

$$\sum_{n=1}^{N} ||x_n - x_{n-1}|| \leq d \min_x W_N(x),$$

it follows that the resulting competitive ratio is at most $(2L+1)d+1$. Similar remarks apply to the result of Theorem 32.

6.5 Competitive Ratio $O(\sqrt{d \log N})$ in Euclidean Space

In this section we prove the discrete-time functional Steiner point has competitive ratio $O(\sqrt{d \log N})$ in Euclidean space (whose norm is denoted by $|| \cdot ||_2$). The same technique applies in any normed space given a suitable concentration result, however we restrict to the Euclidean case for convenience. The idea is as follows. Suppose that the average dual work function increase

$$⨍_{\theta \in \partial B_1^*} W_n^*(\theta) - W_{n-1}^*(\theta) \mathrm{d}\mu(\theta)$$

at time-step n is significant. Then by (6.3.5) the movement from $s(W_{n-1}) \to s(W_n)$ is an integral of pushes by different vectors θ. By concentration of measure, these pushes decorrelate unless the total amount of pushing is exponentially small.

Lemma 6.5.1 ([Bal97, Lemma 2.2])**.** *For any $0 \leq \varepsilon < 1$ and $|w| \leq 1$ in Euclidean space, the set*

$$\{\theta \in \partial B_1 : \langle w, \theta \rangle \geq \varepsilon\}$$

occupies at most $e^{-d\varepsilon^2/2}$ fraction of ∂B_1.

Lemma 6.5.2. *Suppose that $|W_n^*(\theta) - W_{n-1}^*(\theta)| \leq C$ for all $\theta \in \partial B_1$, and set*

$$\lambda = ⨍_{v \in B_1} W_n^*(v) - W_{n-1}^*(v) dv.$$

Then the functional Steiner point movement is at most

$$||s(W_n) - s(W_{n-1})||_2 = O\left(\lambda \sqrt{d\left(1 + \log\left(\frac{C}{\lambda}\right)\right)}\right).$$

Proof. Observe that

$$||s(W_n) - s(W_{n-1})||_2 = \max_{||w||_2 = 1} \langle w, s(W_n) - s(W_{n-1}) \rangle.$$

Fixing a unit vector w, we estimate the inner product on the right-hand side. Set

$$g_n(\theta) = W_n^*(\theta) - W_{n-1}^*(\theta) \geq 0,$$
$$I_z = \fint_{\theta \in \partial B_1^*} g_n(\theta) \cdot 1_{\langle w,\theta\rangle \geq z} \mathrm{d}\mu(\theta).$$

Then $g_n(\theta) \in [0, C]$ for all θ and $\fint_{\theta\in\partial B_1^*} g_n \mathrm{d}\mu(\theta) = \lambda$. Consequently by Lemma 6.5.1,

$$I_z \leq \min\left(\lambda, Ce^{-dz^2/2}\right). \tag{6.5.1}$$

We thus find

$$\begin{aligned}
\langle w, s(W_n) - s(W_{n-1})\rangle &= d \fint_{\theta\in\partial B_1^*} g_n(\theta)\langle w,\theta\rangle \mathrm{d}\mu(\theta) \\
&\leq d \fint_{\substack{\theta\in\partial B_1^* \\ \langle w,\theta\rangle \geq 0}} g_n(\theta)\langle w,\theta\rangle \mathrm{d}\mu(\theta) \\
&= d\int_0^1 I_z \mathrm{d}z \\
&\leq d\int_0^1 \min\left(\lambda, Ce^{-dz^2/2}\right) \mathrm{d}z. \qquad (6.5.2)
\end{aligned}$$

Here the second equality is the tail-sum integral formula. To estimate the resulting integral, set

$$A = \sqrt{\frac{2\log(C/\lambda)}{d}}$$

so that $Ce^{-dA^2/2} = \lambda$. We will assume $A \leq 1$; if $A > 1$ then the expression (6.5.2) is at most $d\lambda \leq dA\lambda$ and it suffices to mimic the below without the second term. We estimate

$$\int_0^1 \min\left(\lambda, Ce^{-dz^2/2}\right) \mathrm{d}z = A\lambda + C\int_A^1 e^{-dz^2/2}\mathrm{d}z.$$

and use the simple bounds

$$\int_A^1 e^{-dz^2/2}\mathrm{d}z \leq \int_0^1 e^{-dz^2/2}\mathrm{d}z \leq O(d^{-1/2}),$$
$$\int_A^1 e^{-dz^2/2}\mathrm{d}z \leq e^{-dA^2/2}\int_A^\infty e^{-dA(z-A)}\mathrm{d}z = \frac{e^{-dA^2/2}}{dA}.$$

Combining,

$$\begin{aligned}\langle w, s(W_n) - s(W_{n-1})\rangle &\leq d\int_0^1 \min\left(\lambda, Ce^{-dz^2/2}\right)\mathrm{d}z \\ &\leq dA\lambda + \min\left(C\sqrt{d}, \frac{Ce^{-dA^2/2}}{A}\right) \\ &= O\left(\lambda\sqrt{d\log\left(\frac{C}{\lambda}\right)}\right) + \min\left(C\sqrt{d}, \lambda\sqrt{\frac{d}{2\log(C/\lambda)}}\right).\end{aligned}$$

With $u = \lambda/C \in [0,1]$, the last term is

$$C\sqrt{d}\cdot\min\left(1, \frac{u}{\sqrt{2\log(1/u)}}\right)$$

For $u \leq [0, 1/2]$, we have $\frac{u}{\sqrt{2\log(1/u)}} \leq O(u)$ giving the bound $O(\lambda\sqrt{d})$. For $u \geq 1/2$ we have $C\sqrt{d} \leq 2\lambda\sqrt{d}$. Hence in both cases,

$$\langle w, s(W_n) - s(W_{n-1})\rangle \leq O\left(\lambda\sqrt{d\left(1+\log\left(\frac{C}{\lambda}\right)\right)}\right)$$

as desired.

□

Theorem 32. *The discrete time functional Steiner point algorithm is $O(\sqrt{d\log N})$ competitive for chasing convex functions in Euclidean space.*

Proof. Call $(x_t)_{t\in[0,N]}$ the continuous path and $(x_{t_n})_{n\leq N}$ the discrete path for $t_n \in (n-1, n]$ as in Proposition 6.2.1. Since the service cost for the discrete path is at most that of the continuous path, we only need to establish the $O(\sqrt{d\log N})$ competitive ratio on the movement of the discrete path. By Lemma 6.3.5,

$$\max_{|\theta|\leq 1} W_N^*(\theta) \leq 2\cdot\min_x W_N(x).$$

Set

$$\lambda_n = \int_{\theta\in\partial B_1^*} W_{t_n}^*(\theta) - W_{t_{n-1}}^*(\theta)\mathrm{d}\mu(\theta).$$

Applying Lemma 6.5.2 with $C = 2\cdot\min_x W_N(x)$ to the movement $||x_{t_n} - x_{t_{n-1}}||_2$ at each step yields:

$$\sum_{n=1}^{N} ||x_{t_n} - x_{t_{n-1}}||_2 \leq O(Cd^{1/2})\cdot\sum_{n\leq N}\frac{\lambda_n}{C}\sqrt{1+\log\left(\frac{C}{\lambda_n}\right)}. \tag{6.5.3}$$

Here the values λ_n are all non-negative and sum to $\int_{\theta \in \partial B_1^*} W_N^*(\theta) \mathrm{d}\mu(\theta) \leq C$. Letting $h(u) = u\sqrt{1+\log(1/u)}$, one readily computes that for $u \in (0,1)$,

$$h'(u) = \frac{2\log(1/u)+1}{2(1+\log(1/u))^{1/2}} \geq 0, \qquad h''(u) = \frac{-2\log(1/u)-3}{4u(1+\log(1/u))^{3/2}} \leq 0.$$

Jensen's inequality therefore implies that setting $\lambda_n = \frac{C}{N}$ for all $n \leq N$ in (6.5.3) gives an upper bound. It follows that the movement cost is at most $O(C\sqrt{d\log(N+1)})$. □

6.6 Steiner Points of Level Sets

6.6.1 A Simplification for Chasing Convex Bodies

Here we show that for chasing convex bodies in discrete time, it suffices to simply set $x_n = s(W_n)$ instead of reducing from a continuous-time problem via Proposition 6.2.1. This simplification does not seem possible for chasing convex functions. The movement cost estimates continue to hold with no changes in the proof, however establishing $s(W_n) \in K_n$ requires a short additional argument. Define the support set $\mathrm{Supp}(W) \subseteq \mathbb{R}_d$ of an abstract work function W to be the set of points x possessing a subgradient $v \in \nabla W(x)$ with $|v| < 1$. For a work function W and convex body K, set

$$W^K(x) = \min_{y \in K} W(y) + ||y - x||.$$

If W is the work function for some sequence of requests, then making an additional request of K results in the new work function W^K.

Proposition 6.6.1. *Supp$(W^K) \subseteq K$ holds for any work function W and convex body K.*

Proof. Suppose $x \notin K$ and set

$$y \in \arg\min_{y_0 \in K} (W(y_0) + ||y_0 - x||).$$

For any z on the segment $\overline{yx}$, it follows that $W(x) - W(z) = ||x - z||$. This implies that no v with $|v| < 1$ can be a subgradient in $\nabla W_n(x)$. □

Corollary 6.6.2. *The algorithm $x_n = s(W_n)$ is d competitive for chasing convex bodies, and $O(\sqrt{d\log N})$ competitive in Euclidean space.*

Proof. Proposition 6.6.1 and the primal definition (6.3.4) together imply $s(W_n) \in K_n$, i.e. the algorithm is valid. The d-competitiveness follows from Theorem 31 and the argument of Proposition 6.2.1 while the $O(\sqrt{d\log N})$ competitive ratio in Euclidean space follows from the argument of

Theorem 32. □

6.6.2 Steiner Points of Level Sets

This final subsection has two main objectives. Theorem 33 states that the functional Steiner point of any work function can be expressed as the Steiner point of large level sets. Corollary 6.6.5 states that the Steiner point of any level set of the work function W_n is inside K_n for convex body chasing. As we discuss at the end, Corollary 6.6.5 is related to the algorithm for chasing convex bodies given by [AGGT21]. Denote level sets by

$$\Omega_{W,R} = \{x : W(x) \leq R\}.$$

It is easy to see that for any work function W and $R \geq \min_x W(x)$,

$$W^{\Omega_{W,R}}(x) = \begin{cases} W(x), & \text{for } x \in \Omega_{W,R} \\ d(x, \Omega_{W,R}) + R, & \text{for } x \notin \Omega_{W,R}. \end{cases}$$

Theorem 33. *For any work function W and $R \geq \min_x W(x)$, it holds that $s(\Omega_{W,R}) = s\left(W^{\Omega_{W,R}}\right)$ and $\lim_{R\to\infty} s(\Omega_{W,R}) = s(W)$. Moreover if Supp$(W) \subseteq \Omega_{W,R}$ then $s(\Omega_{W,R}) = s(W)$.*

Proof. The dual definitions (6.3.2), (6.3.5) imply

$$s(\Omega_{W,R}) - s(W) = d \fint_{\theta\in\partial B_1^*} \left(W^*(\theta) + h_{\Omega_{W,R}}(\theta)\right) n(\theta) \mathrm{d}\mu(\theta). \tag{6.6.1}$$

Also for any $\theta \in \partial B_1^*$,

$$\begin{aligned} \left(W^{\Omega_{W,R}}\right)^*(\theta) &= \inf_{w\in X} \left(W^{\Omega_{W,R}}(w) - \langle w, \theta\rangle\right) \\ &= \inf_{w\in\partial\Omega_{W,R}} \left(W^{\Omega_{W,R}}(w) - \langle w, \theta\rangle\right) \\ &= R - h_{\Omega_{W,R}}(\theta). \end{aligned}$$

It follows from the symmetry $\theta \leftrightarrow -\theta$ that

$$\fint_{\theta\in\partial B_1^*} n(\theta)\mathrm{d}\mu(\theta) = 0.$$

Combining the above yields

$$s(\Omega_{W,R}) = s\left(W^{\Omega_{W,R}}\right).$$

We proceed similarly for the second claim. For any $\theta \in \partial B_1^*$,

$$\begin{aligned} W^*(\theta) &= \inf_{w \in X} (W(w) - \langle \theta, w \rangle) \\ &= \lim_{R \to \infty} \inf_{w \in \Omega_{W,R}} (W(w) - \langle \theta, w \rangle) \\ &= \lim_{R \to \infty} \inf_{w \in \partial\Omega_{W,R}} (W(w) - \langle \theta, w \rangle) \\ &= \lim_{R \to \infty} \left(R - h_{\Omega_{W,R}}(\theta) \right). \end{aligned}$$

Because $W(x) - ||x||$ is uniformly bounded it follows that the expression

$$W^*(\theta) + h_{\Omega_{W,R}}(\theta) - R$$

is uniformly bounded for $(\theta, R) \in (\partial B_1^* \times \mathbb{R}^+)$. As just shown it tends to 0 as $R \to \infty$. The bounded convergence theorem therefore implies

$$\lim_{R \to \infty} \fint_{\theta \in \partial B_1^*} \left| W^*(\theta) + h_{\Omega_{W,R}}(\theta) - R \right| \mathrm{d}\mu(\theta) = 0.$$

Combining with equation (6.6.1) shows that $\lim_{R \to \infty} ||s(\Omega_{W,R}) - s(W)|| = 0$, proving the second assertion. The last assertion is proved similarly after observing that $\mathrm{Supp}(W) \subseteq \Omega_{W,R}$ implies

$$\begin{aligned} W^*(\theta) &= \inf_{w \in X} (W(w) - \langle \theta, w \rangle) \\ &= \lim_{\lambda \uparrow 1} \inf_{w \in X} (W(w) - \langle \lambda\theta, w \rangle) \\ &= \lim_{\lambda \uparrow 1} \inf_{w \in \Omega_{W,R}} (W(w) - \langle \lambda\theta, w \rangle) \\ &= R - h_{\Omega_{W,R}}(\theta). \end{aligned}$$

□

Proposition 6.6.3. *Supp*$(W^{\Omega_{W,R}}) \subseteq$ *Supp*(W) *holds for any* $R \geq \min_x W(x)$.

Proof. Because $\Omega_{W,R}$ is a level set,

$$W^{\Omega_{W,R}}(x) = \begin{cases} W(x), & \text{for } x \in \Omega_{W,R} \\ d(x, \Omega_{W,R}) + R, & \text{for } x \notin \Omega_{W,R} \end{cases}$$

Proposition 6.6.1 combined with the fact that W and $W^{\Omega_{W,R}}$ agree inside $\Omega_{W,R}$ imply that the only possible new support points are on the boundary $\partial\Omega_{W,R}$. Fix a boundary point $y \in \partial\Omega_{W,R} \backslash \mathrm{Supp}(W)$.

Because $y \notin \text{Supp}(W)$, there exists a sequence $(y_i)_{i\in\mathbb{N}} \to y$ satisfying

$$W(y) - W(y_i) \geq (1 - o(1))||y - y_i||.$$

Such a sequence of points y_i must eventually satisfy $W(y_i) \leq W(y)$ and therefore $y_i \in \Omega_{W,R}$, implying $W(y_i) = W^{\Omega_{W,R}}(y_i)$. Hence

$$W^{\Omega_{W,R}}(y) - W^{\Omega_{W,R}}(y_i) \geq (1 - o(1))||y - y_i||.$$

This implies $y \notin \text{Supp}(W^{\Omega_{W,R}})$, completing the proof.

□

Corollary 6.6.4. *Let $W = \widehat{W}^K$ for a work function $\widehat{W}$ and convex body K. For any $R \geq \min_x W(x)$,*

$$s(\Omega_{W,R}) = s\left(W^{\Omega_{W,R}}\right) \in K.$$

Proof. Propositions 6.6.1 and 6.6.3 show that

$$\text{Supp}\left(W^{\Omega_{W,R}}\right) \subseteq \text{Supp}(W) \subseteq K.$$

The primal definition (6.3.4) of the functional Steiner point now implies $s(W^{\Omega_{W,R}}) \in K$.

□

Corollary 6.6.5. *Let W_n be the work function for convex body requests $(K_1, \ldots, K_n)$. Then*

$$s(W_n^{\Omega_{W_n,R}}) \in K_n$$

for any $R \geq \min_x W_n(x)$.

Proof. Immediate from Corollary 6.6.4 with $\widehat{W} = W_{n-1}$ and $K = K_n$.

□

Remark 6.6.1. [AGGT21] solved chasing convex bodies in Euclidean space by using the algorithm $x_n = s\big(W_n^{\Omega_{W_n,R_n}}\big)$ with $R_n = 2^{\lceil \log_2(\min_x W_n(x))\rceil}$. This defines a selector by Corollary 6.6.5. Estimating the movement cost is not difficult because the sets $W_n^{\Omega_{W_n,R}}$ decrease for fixed R. Note that $\text{diam}(\Omega_{W_n,R}) \leq 2R$ because of the inequality $W_t(x) \geq ||x||$ (recall Proposition 6.3.4). Using Theorem 30, the movement from each fixed R value is at most $O(\min(dR, R\sqrt{d \log T}))$. Summing over the geometric sequence of R values yields the same upper bound as in Theorems 31 and 32 up to a constant factor.

[AGGT21] prove that $s\big(W_n^{\Omega_{W_n,R_n}}\big) \in K_n$ using reflectional symmetries that may not exist in arbitrary normed spaces. Corollary 6.6.5 implies that their algorithm also works for general norms.

6.7 Proof of Lemma 6.3.6

Proof. We prove the result for all $v \in B_1^*$ where $\nabla W_t^*(v)$ exists. This includes almost all v by Alexandrov's theorem. Moreover it ensures the conjugate point $v_t^* = \arg\min_{w\in X} W(w) - \langle v, w\rangle$ is well-defined and that W_t is strictly convex at v_t^* [Roc70, Corollary 25.1.2]. We write:

$$\begin{aligned} W_{t+\delta}(v) &= \min_{x_s:[0,t+\delta]\to X} \left(\int_0^{t+\delta} (f_s(x_s) + ||x_s'||\mathrm{d}s - \langle v, x_{t+\delta}\rangle \right) \\ &= \min_{x_s:[t,t+\delta]\to X} \left(W_t(x_t) + \int_t^{t+\delta} f_s(x_s) + ||x_s'||\mathrm{d}s - \langle v, x_{t+\delta}\rangle \right) \end{aligned}$$

For small $\delta \in (0, \varepsilon)$, we show $W_{t+\delta}^*(v) = W_t^*(v) + \delta f_t(v_t^*) + o(\delta)$. For the upper bound,

$$\begin{aligned} W_{t+\delta}(v_t^*) &\le W_t(v_t^*) + \int_t^{t+\delta} f_s(v_t^*)\mathrm{d}s \\ &= W_t(v_t^*) + \delta f_t(v_t^*) + o(\delta) \end{aligned}$$

holds by taking $x_s = v_t^*$ constant for $s \in [t, t+\delta)$ and recalling the assumption that $f_s(x)$ is continuous on $s \in [t, t+\delta)$. Since $v_t^* = \arg\min_x \big(W_t(x) - \langle x, v\rangle\big)$, the upper bound follows from

$$\begin{aligned} W_{t+\delta}^*(v) &\le W_{t+\delta}(v_t^*) - \langle v, v_t^*\rangle \\ &\le W_t(v_t^*) + \delta f_t(v_t^*) + o(\delta) - \langle v, v_t^*\rangle \\ &= W_t^*(v) + \delta f_t(v_t^*) + o(\delta). \end{aligned}$$

For the lower bound, the strict convexity of W_t at v_t^* implies

$$W_t(x) = W_t(v_t^*) + \langle v, x - v_t^*\rangle + \gamma(||x - v_t^*||)$$

where $\gamma : \mathbb{R}^+ \to \mathbb{R}^+$ is continuous and increasing with unique minimum $F(0) = 0$. Therefore any path $x_s : [0, t+\delta] \to X$ satisfies:

$$\begin{aligned} &W_t(x_t) + \int_t^{t+\delta} f_s(x_s) + ||x_s'||\mathrm{d}s - \langle v, x_{t+\delta}\rangle \\ &\qquad \ge W_t(v_t^*) + \langle v, x_t - v_t^*\rangle + \gamma(||x_t - v_t^*||) + \int_t^{t+\delta} f_s(x_s) + ||x_s'||\mathrm{d}s - \langle v, x_{t+\delta}\rangle. \end{aligned}$$

The observation $\int_t^{t+\delta} ||x_s'|| \mathrm{d}s \geq ||x_{t+\delta} - x_t|| \geq \langle v, x_{t+\delta} - x_t \rangle$ implies

$$\begin{aligned} W_t(x_t) + \int_t^{t+\delta} f_s(x_s) + ||x_s'|| \mathrm{d}s - \langle v, x_{t+\delta} \rangle \geq & W_t(x_t) - \langle v, v_t^* \rangle + f(||x_t - v_t^*||) + \int_t^{t+\delta} f_s(x_s) \mathrm{d}s \\ \geq & W_t(v_t^*) - \langle v, v_t^* \rangle + \gamma(||x_t - v_t^*||) + \int_t^{t+\delta} f_s(x_s) \mathrm{d}s \\ \geq & W_t^*(v) + \gamma(||x_t - v_t^*||) + \int_t^{t+\delta} f_s(x_s) \mathrm{d}s. \end{aligned}$$

Because $W_{t+\delta}(v) = W_t(v) + O(\delta)$, we see that for $\delta \to 0$ small we must have $||x_t - v_t^*|| = o_{\delta \to 0}(1)$ for any optimal trajectory x_s witnessing the correct value $W_{t+\delta}$. Additionally,

$$\int_t^{t+\delta} ||x_s'|| \mathrm{d}s + \langle v, x_t - x_{t+\delta} \rangle \geq (1 - |v|) \int_t^{t+\delta} ||x_s'|| \mathrm{d}s \geq (1 - |v|) \sup_{s \in [t, t+\delta]} |x_t - x_s|.$$

which similarly implies $\sup_{s \in [t, t+\delta]} ||x_t - x_s|| = o(1)$ for any optimal trajectory since $||v|| < 1$. It follows that all optimal trajectories satisfy $\int_t^{t+\delta} f_s(x_s) \mathrm{d}s = \delta f_t(v_t^*) + o(\delta)$. This concludes the proof.

□

Chapter 7

A Universal Law of Robustness via Isoperimetry

7.1 Introduction

Solving n equations generically requires only n unknowns[1]. However, the revolutionary deep learning methodology revolves around highly overparametrized models, with many more than n parameters to learn from n training data points. We propose an explanation for this enigmatic phenomenon, showing in great generality that finding a *smooth* function to fit d-dimensional data requires at least nd parameters. In other words, overparametrization by a factor of d is *necessary* for *smooth* interpolation, suggesting that perhaps the large size of the models used in deep learning is a *necessity* rather than a weakness of the framework. Another way to phrase the result is as a *tradeoff* between the size of a model (as measured by the number of parameters) and its "robustness" (as measured by its Lipschitz constant): either one has a small model (with n parameters) which must then be non-robust, or one has a robust model (constant Lipschitz) but then it must be very large (with nd parameters). Such a tradeoff was conjectured for the specific case of two-layer neural networks and Gaussian data in [BLN21]. Our result shows that in fact it is a *universal* phenomenon, which applies to essentially any parametrized function class (including in particular deep neural networks) as well as a much broader class of data distributions. As in [BLN21] we obtain an entire tradeoff curve between size and robustness: our universal law of robustness states that, for any function class smoothly parametrized by p parameters, and for any d-dimensional dataset satisfying mild regularity conditions, any function in this class that fits the data *below the noise level* must have its

[1] As in, for instance, the inverse function theorem in analysis or Bézout's theorem in algebraic geometry. See also [YSJ19, BELM20] for versions of this claim with neural networks.

(Euclidean) Lipschitz constant larger than $\sqrt{\frac{nd}{p}}$.

Theorem 34 (Informal version of Theorem 37)**.** *Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^d \to \mathbb{R}$ and let $(x_i, y_i)_{i=1}^n$ be i.i.d. input-output pairs in $\mathbb{R}^d \times [-1, 1]$. Assume that:*

1. *$\mathcal{F}$ admits a Lipschitz parametrization by p real parameters, each of size at most $poly(n, d)$.*
2. *The distribution μ of the covariates x_i satisfies isoperimetry (or is a mixture theoreof).*
3. *The expected conditional variance of the output (i.e., the "noise level") is strictly positive, denoted $\sigma^2 \equiv \mathbb{E}^\mu[Var[y|x]] > 0$.*

Then, with high probability over the sampling of the data, one has simultaneously for all $f \in \mathcal{F}$:

$$\frac{1}{n}\sum_{i=1}^n (f(x_i) - y_i)^2 \leq \sigma^2 - \varepsilon \;\Rightarrow\; \mathrm{Lip}(f) \geq \widetilde{\Omega}\left(\frac{\varepsilon}{\sigma}\sqrt{\frac{nd}{p}}\right).$$

Remark 7.1.1. For the distributions μ we have in mind, for instance uniform on the unit sphere, there exists with high probability some $O(1)$-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ satisfying $f(x_i) = y_i$ for all i. Indeed, with probability $1 - e^{-\Omega(d)}$ we have $||x_i - x_j|| \geq 1$ for all $1 \leq i \neq j \leq n$ so long as $n \leq poly(d)$. In this case we may apply the Kirszbraun extension theorem to find a suitable f regardless of the labels y_i. More explicitly we may fix a smooth bump function $g : \mathbb{R}^+ \to \mathbb{R}$ with $g(0) = 1$ and $g(x) = 0$ for $x \geq 1$, and then interpolate using the sum of radial basis functions

$$f(x) = \sum_{i=1}^n g(||x - x_i||)y_i.$$

In fact this construction requires only $p = n(d + 1)$ parameters to specify the values $(x_i, y_i)_{i\in[n]}$ and thus determine the function f. Hence $p = n(d + 1)$ parameters suffice for robust interpolation, i.e. Theorem 37 is essentially best possible for $L = O(1)$. A similar construction shows the same conclusion for any $p \in [\widetilde{\Omega}(n), nd]$, essentially tracing the entire tradeoff curve. This is because one can first project onto a fixed subspace of dimension $\tilde{d} = p/n$, and the projected inputs x_i now have pairwise distances at least $\Omega\left(\sqrt{\tilde{d}/d}\right)$ with high probability. The analogous construction on the projected points now requires only $p = \tilde{d}n$ parameters and has Lipschitz constant $L = O\left(\sqrt{d/\tilde{d}}\right) = O\left(\sqrt{\frac{nd}{p}}\right)$.

7.1.1 Speculative implication for real data

To put Theorem 34 in context, we compare to the empirical results presented in [MMS+18]. In the latter work, they consider the MNIST dataset which consists of $n = 6 \times 10^4$ images in dimension

$28^2 = 784$. They trained robustly different architectures, and reported in Figure 4 the size of the architecture versus the obtained robust test accuracy (third plot from the left). One can see a sharp transition from roughly 10% accuracy to roughly 90% accuracy at around 2×10^5 parameters (capacity scale 4 in their notation). Moreover the robust accuracy keeps climbing up with more parameters, to roughly 95% accuracy at roughly 3×10^6 parameters.

How can we compare these numbers to the law of robustness? There are a number of difficulties that we discuss below, and we emphasize that this discussion is highly speculative in nature, though we find that, with a few leaps of faith, our universal law of robustness sheds light on the potential parameter regimes of interest for robust deep learning.

The first difficulty is to evaluate the "correct" dimension of the problem. Certainly the number of pixels per image gives an upper bound, however one expects that the data lies on something like a lower dimensional sub-manifold. Optimistically, we hope that Theorem 34 will continue to apply for an appropriate *effective dimension* which may be rather smaller than the literal number of pixels. This hope is partially justified by the fact that isoperimetry holds in many less-than-picturesque situations, some of which are stated in the next subsection.

Estimating the effective dimension of data manifolds is an interesting problem and has attracted some study in its own right. For instance [FdRL17, PZA+21] both predict that MNIST has effective dimension slightly larger than 10, which is consistent with our numerical discussion at the end of this subsection. The latter also predicts an effective dimension of about 40 for ImageNet. It is unclear how accurate these estimates are for our setting. One concrete issue is that from the point of view of isoperimetry, a "smaller" manifold (e.g. a sphere with radius $r < 1$) will behave as though it has a larger effective dimension (e.g. d/r^2 instead of d). Thus we expect the "scale" of the mixture components to also be relevant for studying real datasets through our result.

Another difficulty is to estimate/interpret the noise value σ^2. From a theoretical point of view, this noise assumption is necessary for otherwise there could exist a smooth classifier with perfect accuracy in $\mathcal{F}$, defeating the point of any lower bound on the size of $\mathcal{F}$. We tentatively would like to think of σ^2 as capturing the contribution of the "difficult" part of the learning problem, that is σ^2 could be thought of as the non-robust generalization error of reasonably good models, so a couple of % of error in the case of MNIST. With that interpretation, one gets "below the noise level" in MNIST with a training error of a couple of %. We believe that versions of the law of robustness might hold without noise; these would need to go beyond representational power and consider the dynamics of learning algorithms.

Finally another subtlety to interpret the empirical results of [MMS+18] is that there is a mismatch between what they measure and our quantities of interest. Namely the law of robustness talks about two things: the training error, and the worst-case robustness (i.e., the Lipschitz constant). On the other hand [MMS+18] measures the *robust generalization error*. Understanding the interplay

between those three quantities is a fantastic open problem. Here we take the perspective that a small robust generalization error should imply a small training error and a small Lipschitz constant. Another important mismatch is that we stated our universal law of robustness for Lipschitzness in ℓ_2, while the experiments in [MMS+18] are for robustness in ℓ_∞. We believe that a variant of the law of robustness remains true for ℓ_∞, a belief again partially justified by how broad isoperimetry is (see next subsection).

With all the caveats described above, we can now look at the numbers as follows: in the [MMS+18] experiments, smooth models with accuracy below the noise level are attained with a number of parameters somewhere in the range $2 \times 10^5 - 3 \times 10^6$ parameters (possibly even larger depending on the interpretation of the noise level), while the law of robustness would predict any such model must have at least nd parameters, and this latter quantity should be somewhere in the range $10^6 - 10^7$ (corresponding to an effective dimension between 15 and 150). While far from perfect, the law of robustness prediction is far more accurate than the classical rule of thumb # parameters $\simeq$ # equations (which here would predict a number of parameters of the order 10^4).

Perhaps more interestingly, one could apply a similar reasoning to the ImageNet dataset, which consists of 1.4×10^7 images of size roughly 2×10^5. Estimating that the effective dimension is a couple of order of magnitudes smaller than this size, the law of robustness predicts that to obtain good robust models on ImageNet one would need at least $10^{10} - 10^{11}$ parameters. This number is larger than the size of current neural networks trained robustly for this task, which sports between $10^8 - 10^9$ parameters. Thus, we arrive at the tantalizing possibility that robust models for ImageNet do not exist yet simply because we are a couple orders of magnitude off in the current scale of neural networks trained for this task.

7.1.2 Related work

Theorem 34 is a direct follow-up to the conjectured law of robustness in [BLN21] for (arbitrarily weighted) two-layer neural networks with Gaussian data. Our result does not actually prove their conjecture, because we assume here polynomially bounded weights. While this assumption is reasonable from a practical perspective, it remains mathematically interesting to prove the full conjecture for the two-layer case. We prove however in Section 7.6 that the polynomial weights assumption is necessary as soon as one considers three-layer neural networks. Let us also mention the [GCL+19, Theorem 6.1] which showed a lower bound $\Omega(nd)$ on the VC dimension of any function class which can robustly interpolate *arbitrary* labels on *all* well-separated input sets $(x_1, \ldots, x_n)$. We also note that a relation between high-dimensional phenomenon such as concentration and adversarial examples has been hypothesized before, such as in [GMF+18].

In addition to [MMS+18], several recent works have experimentally studied the relationship between a neural network scale and its achieved robustness, see e.g., [NBA+18, XY20, GQU+20].

It has been consistently reported that larger networks help tremendously for robustness, beyond what is typically seen for classical non-robust accuracy. We view our universal law of robustness as putting this empirical observation on a more solid footing: scale is actually *necessary* to achieve robustness.

The law of robustness setting is closely related to the interpolation setting: in the former case one considers models optimizing "beyond the noise level", while in the latter case one studies models with perfect fit on the training data. The study of generalization in this interpolation regime has been a central focus of learning theory in the last few years (see e.g., [BHMM19, MM19, BLLT20, NKB$^+$20]), as it seemingly contradicts classical theory about regularization. More broadly though, generalization remains a mysterious phenomon in deep learning, and the exact interplay between the law of robustness' setting (interpolation regime/worst-case robustness) and (robust) generalization error is a fantastic open problem. Interestingly, we note that one could potentially avoid the conclusion of the law of robustness (that is, that large models are necessary for robustness), with early stopping methods that could stop the optimization once the noise level is reached. In fact, this theoretically motivated suggestion has already been empirically tested and confirmed in the recent work [RWK20], showing again a close tie between the conclusions one can draw from the law of robustness and actual practical settings.

Classical lower bounds on the gradient of a function include Poincaré type inequalities, but they are of a qualitately different nature compared to the law of robustness lower bound. We recall that a measure μ on $\mathbb{R}^d$ satisfies a Poincaré inequality if for any function f, one has $\mathbb{E}^\mu[\|\nabla f\|^2] \geq C \cdot \mathrm{Var}(f)$ (for some constant $C > 0$). In our context, such a lower bound for an interpolating function f has essentially no consequence since the variance f could be exponentially small. In fact this is tight, as one easily use similar constructions to those in [BLN21] to show that one can interpolate with an exponentially small expected norm squared of the gradient (in particular it is crucial in the law of robustness to consider the Lipschitz constant, i.e., the supremum of the norm of the gradient). On the other hand, our isoperimetry assumption is related to a certain strenghtening of the Poincaré inequality known as log-Sobolov inequality (see e.g., [Led01]). If the covariate measure satisfies only a Poincaré inequality, then we could prove a weaker law of robustness of the form $\mathrm{Lip} \gtrsim \frac{n\sqrt{d}}{p}$ (using for example the concentration result obtained in [BL97]). For the case of two-layer neural networks there is another natural notion of smoothness (different from ℓ_p norms of the gradient) that can be considered, known as the Barron norm. In [BELM20] it is shown that for such a notion of smoothness there is no tradeoff à la the law of robustness, namely one can simultaneously be optimal both in terms of Barron norm and in terms of the network size. More generally, it is an interesting challenge to understand for which notions of smoothness there is a tradeoff with size.

7.1.3 Isoperimetry

Concentration of measure and isoperimetry are perhaps the most ubiquitous features of high-dimensional geometry. In short, they assert in many cases that Lipschitz functions on high-dimensional space concentrate tightly around their mean. Our result assumes that the distribution μ of the covariates x_i satisfies such an inequality in the following sense.

Definition 7.1.1. *A probability measure μ on $\mathbb{R}^d$ satisfies c-isoperimetry if for any bounded L-Lipschitz $f : \mathbb{R}^d \to \mathbb{R}$, and any $t \geq 0$,*

$$\mathbb{P}[|f(x) - \mathbb{E}[f]| \geq t] \leq 2e^{-\frac{dt^2}{2cL^2}}. \tag{7.1.1}$$

In general, if a scalar random variable X satisfies $\mathbb{P}[|X| \geq t] \leq 2e^{-t^2/C}$ then we say X is C-subgaussian. Hence isoperimetry states that the output of any Lipschitz function is $O(1)$-subgaussian under suitable rescaling. Distributions satisfying $O(1)$-isoperimetry include high dimensional Gaussians $\mu = \mathcal{N}\left(0, \frac{I_d}{d}\right)$ and uniform distributions on spheres and hypercubes (normalized to have diameter 1). Isoperimetry also holds for mild perturbations of these idealized scenarios, including[2]:

- The sum of a Gaussian and an independent random vector of small norm [CCNW21].
- Strongly log-concave measures in any normed space [BL00, Proposition 3.1].
- Manifolds with positive Ricci curvature [Gro86, Theorem 2.2].

Due to the last condition above, we believe our results are realistic even under the *manifold hypothesis* that high-dimensional data tends to lie on a lower-dimensional submanifold. This viewpoint on learning has been studied for decades, see e.g. [HS89, KL93, RS00, TDSL00, NM10, FMN16]. We also note that our formal theorem (Theorem 37) actually applies to distributions that can be written as a mixture of distributions satisfying isoperimetry. Let us also point out that from a technical perspective, our proof is not tied to the Euclidean norm and applies essentially whenever Definition 7.1.1 holds. The main difficulty in extending the law of robustness to e.g. the earth-mover distance seems to be identifying realistic cases which satisfy isoperimetry.

Our proofs will repeatedly use the following simple fact:

Proposition 7.1.2. *If $X_1, \ldots, X_n$ are independent, C-subgaussian, with mean 0, then $X_{av} = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ is $18C$-subgaussian.*

Proof. By [vH14, Exercise 3.1 part d.],

$$\mathbb{E}\left[e^{X_i^2/3C}\right] \leq 2, \quad i \in [n].$$

[2]The first two examples satisfy a logarithmic Sobolev inequality, which implies isoperimetry [Led99, Proposition 2.3].

It is immediate by Hölder that the same bound holds for X_{av} in place of X_i, and using [vH14, Exercise 3.1 parts e. and c.] now implies the first claim. The second claim follows similarly, since by convexity we have

$$\mathbb{E}[e^{Y^2/3C}] \leq \mathbb{E}[e^{X_1^2/3C}] \leq 2.$$

□

7.2 A finite approach to the law of robustness

For the function class of two-layer neural networks, [BLN21] investigated several approaches to prove the law of robustness. At a high level, the proof strategies there relied on various ways to measure how "large" the set of two-layer neural networks can be (specifically, they tried a geometric approach based on relating to multi-index models, a statistical approach based on the Rademacher complexity, and an algebraic approach for the case of polynomial activations).

In this chapter we take here a different route: we shift the focus from the function class $\mathcal{F}$ to an *individual* function $f \in \mathcal{F}$. Namely, our proof starts by asking the following question: for a fixed function f, what is the probability that it would give a good approximate fit on the (random) data? For simplicity, consider for a moment the case where we require f to actually interpolate the data (i.e., perfect fit), and say that y_i are random ± 1 labels. The key insight is that isoperimetry implies that *either* the 0-level set of f *or* the 1-level set of f must have probability smaller than $\exp\left(-\frac{d}{\mathrm{Lip}(f)^2}\right)$. Thus, the probability that f fits all the n points is at most $\exp\left(-\frac{nd}{\mathrm{Lip}(f)^2}\right)$ so long as both labels $y_i \in \{-1, 1\}$ actually appear a constant fraction of the time. In particular, using an union bound[3], for a finite function class $\mathcal{F}$ of size N with L-Lipschitz functions, the probability that there exists a function $f \in \mathcal{F}$ fitting the data is at most

$$N \exp\left(-\frac{nd}{L^2}\right) = \exp\left(\log(N) - \frac{nd}{L^2}\right).$$

Thus we see that, if $L \ll \sqrt{\frac{nd}{\log(N)}}$, then the probability of finding a fitting function in $\mathcal{F}$ is very small. This basically concludes the proof, since via a standard discretization argument, for a smoothly parametrized family with p (bounded) parameters one expects $\log(N) = \tilde{O}(p)$.

We now give the formal proof, which applies in particular to approximate fit rather than exact fit in the argument above. The only difference is that we will identify a well-chosen subgaussian random variable in the problem. We start with the finite function class case:

Theorem 35. *Let (x_i, y_i) be i.i.d. input-output pairs in $\mathbb{R}^d \times [-1, 1]$ such that:*

[3] In this informal argument we ignore the possibility that the labels y_i are not well-balanced. Note that the probability of this rare event is not amplified by a union bound over $f \in \mathcal{F}$.

1. *The distribution μ of the covariates x_i can be written as $\mu = \sum_{\ell=1}^k \alpha_\ell \mu_\ell$, where each μ_ℓ satisfies c-isoperimetry and $\alpha_\ell \geq 0, \sum_{\ell=1}^k \alpha_\ell = 1$.*

2. *The expected conditional variance $\sigma^2 \equiv \mathbb{E}^\mu[Var[y|x]] > 0$ of the output is strictly positive.*

Then one has:

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \leq \sigma^2 - \varepsilon\right) \leq 4k \exp\left(-\frac{n\varepsilon^2}{8^3 k}\right) + 2\exp\left(\log(|\mathcal{F}|) - \frac{\varepsilon^2 nd}{10^4 cL^2}\right).$$

We start with a lemma showing that, to optimize beyond the noise level one must necessarily correlate with the noise part of the labels. In what follows we denote $g(x) = \mathbb{E}[y|x]$ for the target function, and $z_i = y_i - g(x_i)$ for the noise part of the observed labels (namely y_i is the sum of the target function $g(x_i)$ and the noise term z_i).

Lemma 7.2.1. *One has*

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \leq \sigma^2 - \varepsilon\right) \leq 2\exp\left(-\frac{n\varepsilon^2}{8^3}\right) + \mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^n f(x_i) z_i \geq \frac{\varepsilon}{4}\right).$$

Proof. The sequence (z_i^2) is i.i.d., with mean σ^2, and such that $|z_i|^2 \leq 4$. Thus Hoeffding's inequality yields:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n z_i^2 \leq \sigma^2 - \frac{\varepsilon}{6}\right) \leq \exp\left(-\frac{n\varepsilon^2}{8^3}\right). \tag{7.2.1}$$

On the other hand the sequence $(z_i g(x_i))$ is i.i.d., with mean 0 (since $\mathbb{E}[z_i|x_i] = 0$), and such that $|z_i g(x_i)| \leq 2$. Thus Hoeffding's inequality yields:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n z_i g(x_i) \leq -\frac{\varepsilon}{6}\right) \leq \exp\left(-\frac{n\varepsilon^2}{8^3}\right). \tag{7.2.2}$$

Let us write $Z = \frac{1}{\sqrt{n}}(z_1, \ldots, z_n), G = \frac{1}{\sqrt{n}}(g(x_1), \ldots, g(x_n))$, and $F = \frac{1}{\sqrt{n}}(f(x_1), \ldots, f(x_n))$. We claim that if $\|Z\|^2 \geq \sigma^2 - \frac{\varepsilon}{6}$ and $\langle Z, G\rangle \geq -\frac{\varepsilon}{6}$, then for any $f \in \mathcal{F}$ one has

$$\|G + Z - F\|^2 \leq \sigma^2 - \varepsilon \Rightarrow \langle F, Z\rangle \geq \frac{\varepsilon}{4}.$$

This claim together with (7.2.1) and (7.2.2) conclude the proof. On the other hand the claim itself directly follows from:

$$\sigma^2 - \varepsilon \geq \|G + Z - F\|^2 = \|Z + G - F\|^2 = \|Z\|^2 + 2\langle Z, G - F\rangle + \|G - F\|^2 \geq \sigma^2 - \frac{\varepsilon}{2} - 2\langle Z, F\rangle.$$

□

We can now proceed to the proof of Theorem 35:

Proof. First note that without loss of generality we can assume that the range of any function in $\mathcal{F}$ is included in $[-1,1]$ (indeed clipping the values improves both the fit to any $y \in [-1,1]$ and the Lipschitz constant). We also assume without loss of generality that all functions in $\mathcal{F}$ are L-Lipschitz.

For clarity let us start with the case $k = 1$. By the isoperimetry assumption we have that $\sqrt{\frac{d}{c}}\frac{f(x_i)-\mathbb{E}[f]}{L}$ is 2-subgaussian. Since $|z_i| \leq 2$, we also have that $\sqrt{\frac{d}{c}}\frac{(f(x_i)-\mathbb{E}[f])z_i}{L}$ is 8-subgaussian. Moreover, the latter random variable has mean zero since $\mathbb{E}[z|x] = 0$. Thus by Proposition 7.1.2 (and $8 \times 18 = 12^2$) we have:

$$\mathbb{P}\left(\sqrt{\frac{d}{cnL^2}}\sum_{i=1}^{n}(f(x_i) - \mathbb{E}[f])z_i \geq t\right) \leq 2\exp\left(-(t/12)^2\right) .$$

Rewriting (and noting $12 \times 8 \leq 10^2$), we find:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - \mathbb{E}[f])z_i \geq \frac{\varepsilon}{8}\right) \leq 2\exp\left(-\frac{\varepsilon^2 nd}{10^4 cL^2}\right) . \tag{7.2.3}$$

Since we assumed that the range of the functions is in $[-1,1]$ we have $\mathbb{E}[f] \in [-1,1]$ and hence:

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[f]z_i \geq \frac{\varepsilon}{8}\right) \leq \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} z_i\right| \geq \frac{\varepsilon}{8}\right) . \tag{7.2.4}$$

(This step is the analog of requiring the labels y_i to be well-balanced in the example of perfect interpolation.) By Hoeffding's inequality, the above quantity is smaller than $2\exp(-n\varepsilon^2/8^3)$ (recall that $|z_i| \leq 2$). Thus we obtain with a union bound:

$$\begin{aligned}\mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n} f(x_i)z_i \geq \frac{\varepsilon}{4}\right) &\leq |\mathcal{F}| \cdot \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - \mathbb{E}[f])z_i \geq \frac{\varepsilon}{8}\right) + \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} z_i\right| \geq \frac{\varepsilon}{8}\right) \\ &\leq 2|\mathcal{F}| \cdot \exp\left(-\frac{\varepsilon^2 nd}{10^4 cL^2}\right) + 2\exp\left(-\frac{n\varepsilon^2}{8^3}\right) .\end{aligned}$$

Together with Lemma 7.2.1 this concludes the proof for $k = 1$.

We now turn to the case $k > 1$. We first sample the mixture component $\ell_i \in [k]$ for each data point $i \in [n]$, and we now reason conditioned on these mixture components. Let $S_\ell \subset [n]$ be the set of data points sampled from mixture component $\ell \in [k]$, that is $x_i, i \in S_\ell$, is i.i.d. from μ_ℓ. We now have that $\sqrt{\frac{d}{c}}\frac{f(x_i)-\mathbb{E}^{\mu_{\ell_i}}[f]}{L}$ is 1-subgaussian (notice that the only difference is that now we need to center by $\mathbb{E}^{\mu_{\ell_i}}[f]$, which depends on the mixture component). In particular using the same

reasoning as for (7.2.3) we obtain (crucially note that Proposition 7.1.2 does not require the random variables to be identically distributed):

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-\overset{\mu_{\ell_i}}{\mathbb{E}}[f])z_i\geq\frac{\varepsilon}{8}\right)\leq 2\exp\left(-\frac{\varepsilon^2 nd}{9^4cL^2}\right). \tag{7.2.5}$$

Next we want to appropriately modify (7.2.4). To do so note that:

$$\max_{m_1,\ldots,m_k\in[-1,1]}\sum_{i=1}^{n}m_{\ell_i}z_i=\sum_{\ell=1}^{k}\left|\sum_{i\in S_\ell}z_i\right|,$$

so that we can rewrite (7.2.4) as:

$$\mathbb{P}\left(\exists f\in\mathcal{F}:\frac{1}{n}\sum_{i=1}^{n}\overset{\mu_{\ell_i}}{\mathbb{E}}[f]z_i\geq\frac{\varepsilon}{8}\right)\leq\mathbb{P}\left(\frac{1}{n}\sum_{\ell=1}^{k}\left|\sum_{i\in S_\ell}z_i\right|\geq\frac{\varepsilon}{8}\right).$$

Now note that $\sum_{\ell=1}^{k}\sqrt{|S_\ell|}\leq\sqrt{nk}$ and thus we have:

$$\mathbb{P}\left(\frac{1}{n}\sum_{\ell=1}^{k}\left|\sum_{i\in S_\ell}z_i\right|\geq\frac{\varepsilon}{8}\right)\leq\mathbb{P}\left(\sum_{\ell=1}^{k}\left|\sum_{i\in S_\ell}z_i\right|\geq\frac{\varepsilon}{8}\sqrt{\frac{n}{k}}\sum_{\ell=1}^{k}\sqrt{|S_\ell|}\right)\leq\sum_{\ell=1}^{k}\mathbb{P}\left(\left|\sum_{i\in S_\ell}z_i\right|\geq\frac{\varepsilon}{8}\sqrt{\frac{n}{k}}\sqrt{|S_\ell|}\right).$$

Finally by Hoeffding's inequality, we have for any $\ell\in[k]$, $\mathbb{P}\left(\left|\sum_{i\in S_\ell}z_i\right|\geq t\sqrt{|S_\ell|}\right)\leq 2\exp\left(-\frac{t^2}{8}\right)$, and thus the last display is bounded from above by $2k\exp\left(-\frac{n\varepsilon^2}{8^3k}\right)$. The proof can now be concluded as in the case $k=1$. □

In fact the above result can be further improved for small σ using the following Lemma 7.2.2. The intuition is that the naive estimate in (7.2.5) of

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-\overset{\mu_{\ell_i}}{\mathbb{E}}[f])z_i\geq\frac{\varepsilon}{8}\right)$$

was loose. Indeed $\mathbb{E}[z_i^2]\leq\sigma^2$, but (7.2.5) did not take advantage of this and only used that $|z_i|\leq 2$ almost surely. For instance, if the variables x_i and z_i were independent, then the sum

$$\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-\overset{\mu_{\ell_i}}{\mathbb{E}}[f])z_i$$

would have subgaussian constant proportional to $\frac{1}{n}\sum_{i=1}^{n}z_i^2$ after conditioning on $(z_1,\ldots,z_n)$. Since $\frac{1}{n}\sum_{i=1}^{n}z_i^2=O(\sigma^2)$ with high probability, the desired improvement would follow.

However because the variables x_i and z_i are not independent, it is not obvious how to improve

on the bound (7.2.5). Our strategy is to carefully construct noisy realizations w_i of z_i and then argue that conditioning on w_i can affect the distribution of x_i by at most a constant factor $\left(\frac{10}{\sigma}\right)^3$. Thus conditioning on $(w_1, \ldots, w_n)$ only changes the distribution of $(x_1, \ldots, x_n)$ by a factor $\left(\frac{10}{\sigma}\right)^{3n}$. The argument above for independent (x_i, z_i) now goes through up to this distortion factor, yielding the result below. We defer details to the Appendix.

Lemma 7.2.2. *In the setting of Theorem 35, we have*

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - \overset{\mu_{\ell_i}}{\mathbb{E}}[f])z_i \geq \frac{\varepsilon}{8}\right) \leq \exp\left(-\frac{n\sigma^2}{4}\right) + \left(\frac{20}{\sigma}\right)^{3n}\exp\left(\log|\mathcal{F}| - \frac{\varepsilon^2 nd}{8^6 cL^2\sigma^2}\right).$$

By using Lemma 7.2.2 in place of (7.2.5) when proving Theorem 35, one readily obtains the following.

Theorem 36. *In the setting of Theorem 35, we have*

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \leq \sigma^2 - \varepsilon\right) \leq (4k{+}1)\exp\left(-\frac{n\varepsilon^2}{8^3 k}\right) + \left(\frac{20}{\sigma}\right)^{3n}\exp\left(\log|\mathcal{F}| - \frac{\varepsilon^2 nd}{8^6 cL^2\sigma^2}\right).$$

Proof. Using Lemma 7.2.2 in place of (7.2.5) when proving Theorem 35 immediately implies

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \leq \sigma^2 - \varepsilon\right) \leq 4k\exp\left(-\frac{n\varepsilon^2}{8^3 k}\right) + \exp\left(-\frac{n\sigma^2}{4}\right) + \left(\frac{20}{\sigma}\right)^{3n}\exp\left(\log|\mathcal{F}| - \frac{\varepsilon^2 nd}{8^6 cL^2\sigma^2}\right).$$

It remains to observe that $\frac{\varepsilon^2}{8^3 k} \leq \frac{\sigma^2}{4}$ since $\varepsilon \leq \sigma^2 \leq 1$. □

Finally we can now state and prove the formal version of the informal Theorem 34 from the introduction. We remark that we now impose a mild lower bound on the dimension d depending only on ε and σ, which is used to account for the factor $\left(\frac{20}{\sigma}\right)^{3n}$. It is not necessary if Theorem 35 is used in place of Theorem 36 (which would sacrifice a factor σ in the resulting lower bound on $\mathrm{Lip}(f)$).

Theorem 37. *Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^d \to \mathbb{R}$ and let $(x_i, y_i)_{i=1}^n$ be i.i.d. input-output pairs in $\mathbb{R}^d \times [-1, 1]$. Fix $\varepsilon, \delta \in (0, 1)$. Assume that:*

1. *The function class can be written as $\mathcal{F} = \{f_{\boldsymbol{w}}, \boldsymbol{w} \in \mathcal{W}\}$ with $\mathcal{W} \subset \mathbb{R}^p$, $\mathrm{diam}(\mathcal{W}) \leq W$ and for any $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathcal{W}$,*

$$||f_{\boldsymbol{w}_1} - f_{\boldsymbol{w}_2}||_\infty \leq J||\boldsymbol{w}_1 - \boldsymbol{w}_2||.$$

2. *The distribution μ of the covariates x_i can be written as $\mu = \sum_{\ell=1}^k \alpha_\ell \mu_\ell$, where each μ_ℓ satisfies c-isoperimetry, $\alpha_\ell \geq 0, \sum_{\ell=1}^k \alpha_\ell = 1$, and k is such that $9^4 k \log(8k/\delta) \leq n\varepsilon^2$.*

3. *The expected conditional variance of the output is strictly positive, denoted $\sigma^2 \equiv \mathbb{E}^\mu[Var[y|x]] > 0$.*

4. *The dimension d is large compared to ε and σ:*

$$d \geq \frac{8^8 cL^2 \log(20/\sigma)}{\varepsilon^2}.$$

Then, with probability at least $1-\delta$ with respect to the sampling of the data, one has simultaneously for all $f \in \mathcal{F}$:

$$\frac{1}{n}\sum_{i=1}^n (f(x_i)-y_i)^2 \leq \sigma^2 - \varepsilon \;\Rightarrow\; \mathrm{Lip}(f) \geq \frac{\varepsilon}{2^{11}\sigma\sqrt{c}} \sqrt{\frac{nd}{p\log(60WJ\varepsilon^{-1}) + \log(4/\delta)}}\,. \tag{7.2.6}$$

Moreover if $\mathcal{W}$ consists only of k-sparse vectors with $||w||_0 \leq k$, then the above inequality improves to

$$\frac{1}{n}\sum_{i=1}^n (f(x_i)-y_i)^2 \leq \sigma^2 - \varepsilon \;\Rightarrow\; \mathrm{Lip}(f) \geq \frac{\varepsilon}{2^{11}\sigma\sqrt{c}} \sqrt{\frac{nd}{k\log(60WJp\varepsilon^{-1}) + \log(4/\delta)}}\,. \tag{7.2.7}$$

Proof. Define $\mathcal{W}_L \subseteq \mathcal{W}$ by

$$\mathcal{W}_L \equiv \{\boldsymbol{w} \in \mathcal{W} : \mathrm{Lip}(f_{\boldsymbol{w}}) \leq L\}.$$

Denote $\mathcal{W}_{L,\varepsilon} \subseteq \mathcal{W}_L$ for an $\frac{\varepsilon}{8J}$-net of $\mathcal{W}_L$. We have in particular $|\mathcal{W}_\varepsilon| \leq (60WJ\varepsilon^{-1})^p$. We apply Theorem 36 to $\mathcal{F}_{L,\varepsilon} \equiv \{f_{\boldsymbol{w}}, \boldsymbol{w} \in \mathcal{W}_{L,\varepsilon}\}$:

$$\begin{aligned}
&\mathbb{P}\left(\exists f \in \mathcal{F}_{L,\varepsilon} : \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \leq \sigma^2 - \frac{\varepsilon}{2}\right) \\
&\leq (4k+1)\exp\left(-\frac{n\varepsilon^2}{9^4 k}\right) + \left(\frac{20}{\sigma}\right)^{3n} \exp\left(p\log(60WJ\varepsilon^{-1}) - \frac{2\varepsilon^2 nd}{8^7 cL^2\sigma^2}\right).
\end{aligned}$$

Observe that if $\|f-g\|_\infty \leq \frac{\varepsilon}{8}$ and $\|y\|_\infty, \|f\|_\infty, \|g\|_\infty \leq 1$, then

$$\frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \leq \frac{\varepsilon}{2} + \frac{1}{n}\sum_{i=1}^n (y_i - g(x_i))^2.$$

(We may again assume without loss of generality that all functions in $\mathcal{F}$ map to $[-1,1]$.) Thus we

obtain for any $L > 0$:

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \leq \sigma^2 - \varepsilon \text{ and } \mathrm{Lip}(f) \leq L\right) \tag{7.2.8}$$
$$\leq (4k+1)\exp\left(-\frac{n\varepsilon^2}{10^4 k}\right) + \left(\frac{20}{\sigma}\right)^{3n}\exp\left(p\log(60WJ\varepsilon^{-1}) - \frac{2\varepsilon^2 nd}{8^7 cL^2\sigma^2}\right).$$

The first assumption ensures that for any $\boldsymbol{w} \in \mathcal{W}_L$, there is $\boldsymbol{w}' \in \mathcal{W}_{L,\varepsilon}$ with $\|f_{\boldsymbol{w}} - f_{\boldsymbol{w}'}\|_\infty \leq \frac{\varepsilon}{8}$. The final assumption ensures that

$$\left(\frac{20}{\sigma}\right)^{3n}\exp\left(-\frac{\varepsilon^2 nd}{8^7 cL^2\sigma^2}\right) \leq 1. \tag{7.2.9}$$

Finally we use the second assumption to show the probability in (7.2.8) just above is at most δ when $L = \frac{\varepsilon}{2^{11}\sigma\sqrt{c}}\sqrt{\frac{nd}{p\log(60WJ\varepsilon^{-1})+\log(4/\delta)}}$. The first term is estimated via

$$(4k+1)\exp\left(-\frac{n\varepsilon^2}{9^4 k}\right) \leq \frac{(4k+1)\delta}{8k} \leq \frac{3\delta}{4}.$$

Using the inequality $2^{22} \geq 8^7$, we find

$$\left(\frac{20}{\sigma}\right)^{3n}\exp\left(p\log(60WJ\varepsilon^{-1}) - \frac{2\varepsilon^2 nd}{8^7 cL^2\sigma^2}\right) \overset{(7.2.9)}{\leq} \exp\left(p\log(60WJ\varepsilon^{-1}) - \frac{\varepsilon^2 nd}{8^7 cL^2\sigma^2}\right)$$
$$\leq e^{-\log(4/\delta)} = \frac{\delta}{4}.$$

Combining these estimates on (7.2.8) concludes the proof of (7.2.6).

To show (7.2.7), the proof proceeds identically after the improved estimate $|\mathcal{W}_\varepsilon| \leq (60WJp\varepsilon^{-1})^k$. To obtain this estimate, note that the number of k-subsets $S \subseteq \binom{[p]}{k}$ is at most p^k. Letting $\mathcal{W}_S$ consist of those $w \in \mathcal{W}$ with $w_i = 0$ for all $i \notin S$, the size of an ε-net $\mathcal{W}_{S,\varepsilon}$ for $\mathcal{W}_S$ is $|\mathcal{W}_{S,\varepsilon}| \leq (60WJ\varepsilon^{-1})^k$. Therefore the union

$$\bigcup_{S\subseteq\binom{[p]}{k}} \mathcal{W}_{S,\varepsilon}$$

is an ε-net of $\mathcal{W}$ of size at most $(60WJp\varepsilon^{-1})^k$ as claimed above. □

7.3 Deep neural networks

We now specialize the law of robustness (Theorem 37) to multi-layer neural networks. We consider a rather general class of depth D neural networks described as follows. First, we require that the neurons are partitioned into layers $\mathcal{L}_1, \ldots, \mathcal{L}_D$, and that all connections are from $\mathcal{L}_i \to \mathcal{L}_j$ for some

$i < j$. This includes the basic feed-forward case in which only connections $\mathcal{L}_i \to \mathcal{L}_{i+1}$ are used as well as more general skip connections. We specify (in the natural way) a neural network by matrices W_j of shape $|\mathcal{L}_j| \times \sum_{i<j} |\mathcal{L}_i|$ for each $1 \leq j \leq D$, as well as 1-Lipschitz non-linearities $\sigma_{j,\ell}$ and scalar biases $b_{j,\ell}$ for each (j, ℓ) satisfying $\ell \in |\mathcal{L}_j|$. We use fixed non-linearities $\sigma_{j,\ell}$ as well as a fixed architecture, in the sense that each matrix entry $W_j[k, \ell]$ is either always 0 or else it is variable (and similarly for the bias terms).

To match the notation of Theorem 37, we identify the parametrization in terms of the matrices (W_j) and bias terms $(b_{j,\ell})$ to a single p-dimensional vector $\boldsymbol{w}$ as follows. A variable matrix entry $W_j[k, \ell]$ is set to $w_{a(j,k,\ell)}$ for some fixed index $a(j, k, \ell) \in [p]$, and a variable bias term $b_{j,\ell}$ is set to $w_{a(j,\ell)}$ for some $a(j, \ell) \in [p]$. Thus we now have a parametrization $\boldsymbol{w} \in \mathbb{R}^p \mapsto f_{\boldsymbol{w}}$ where $f_{\boldsymbol{w}}$ is the neural network represented by the parameter vector $\boldsymbol{w}$. Importantly, note that our formulation allows for weight sharing (in the sense that a shared weight is counted only as a single parameter). For example, this is important to obtain an accurate count of the number of parameters in convolutional architectures.

In order to apply Theorem 37 to this class of functions we need to estimate the Lipschitz constant of the parametrization $\boldsymbol{w} \mapsto f_{\boldsymbol{w}}$. To do this we introduce three more quantities. First, we shall assume that all the parameters are bounded in magnitude by W, that is we consider the set of neural networks parametrized by $\boldsymbol{w} \in [-W, W]^p$. Next, for the architecture under consideration, denote Q for the maximum number of matrix entries/bias terms that are tied to a single parameter w_a for some $a \in [p]$. Finally we define

$$B(\boldsymbol{w}) = \prod_{j \in [D]} \max(\|W_j\|_{op}, 1).$$

Observe that $B(\boldsymbol{w})$ is an upper bound on the Lipschitz constant of the network itself, i.e., the map $x \mapsto f_{\boldsymbol{w}}(x)$. It turns out that a uniform control on it also controls the Lipschitz constant of the *parametrization* $\boldsymbol{w} \mapsto f_{\boldsymbol{w}}$. Namely we have the following lemma:

Lemma 7.3.1. *Let $x \in \mathbb{R}^d$ such that $\|x\| \leq R$, and $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^p$ such that $B(\boldsymbol{w}_1), B(\boldsymbol{w}_2) \leq \overline{B}$. Then one has*

$$|f_{\boldsymbol{w}_1}(x) - f_{\boldsymbol{w}_2}(x)| \leq \overline{B}^2 QR\sqrt{p}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|.$$

Moreover for any $\boldsymbol{w} \in [-W, W]^p$ with $W \geq 1$, one has

$$B(\boldsymbol{w}) \leq (W\sqrt{pQ})^D.$$

Proof. Fix an input x and define g_x by $g_x(\boldsymbol{w}) = f_{\boldsymbol{w}}(x)$. A standard gradient calculation for multi-layer neural networks directly shows that $\|\nabla g_x(\boldsymbol{w})\|_\infty \leq B(\boldsymbol{w})QR$ so that $\|\nabla g_x(\boldsymbol{w})\| \leq B(\boldsymbol{w})QR\sqrt{p}$. Since the matrix operator norm is convex (and nonnegative) it follows that $B(\boldsymbol{w}) \leq B(\boldsymbol{w}_1)B(\boldsymbol{w}_2) \leq$

$\overline{B}^2$ on the entire segment $[\boldsymbol{w}_1, \boldsymbol{w}_2]$ by multiplying over layers. Thus $\|\nabla g_x(\boldsymbol{w})\| \leq \overline{B}^2 QR\sqrt{p}$ on that segment, which concludes the proof of the first claimed inequality. The second claimed inequality follows directly from $\|W_j\|_{op} \leq \|W_j\|_2 \leq W\sqrt{pQ}$. □

Lemma 7.3.1 shows that when applying Theorem 37 to our class of neural networks one can always take $J = R(WQp)^D$ (assuming that the covariate measure μ is supported on the ball of radius R). Thus in this case the law of robustness (under the assumptions of Theorem 37) directly states that with high probability, any neural network in our class that fits the training data well below the noise level must also have:

$$\mathrm{Lip}(f) \geq \tilde{\Omega}\left(\sqrt{\frac{nd}{Dp}}\right), \tag{7.3.1}$$

where $\tilde{\Omega}$ hides logarithmic factors in W, p, R, Q, and the probability of error δ. Thus we see that the law of robustness, namely that the number of parameters should be at least nd for a smooth model with low training error, remains intact for constant depth neural networks. If taken at face value, the lower bound (7.3.1) suggests that it is better in practice to distribute the parameters towards *depth* rather than *width*, since the lower bound is decreasing with D. On the other hand, we note that (7.3.1) can be strengthened to:

$$\mathrm{Lip}(f) \geq \tilde{\Omega}\left(\sqrt{\frac{nd}{p\log(\overline{B})}}\right), \tag{7.3.2}$$

for the class of neural networks such that $B(\boldsymbol{w}) \leq \overline{B}$. In other words the dependence on the depth all but disappears by simply assuming that the quantity $B(\boldsymbol{w})$ (a natural upper bound on the Lipschitz constant of the network) is polynomially controlled. Interestingly many works have suggested to keep $B(\boldsymbol{w})$ under control, either for regularization purpose (for example [BFT17] relates $B(\boldsymbol{w})$ to the Rademacher complexity of multi-layer neural networks) or to simply control gradient explosion during training, see e.g., [ASB16, CBG$^+$17, MHRB17, MKKY18, JCC$^+$19, YM17]. Moreover, in addition to being well-motivated in practice, the assumption that $\overline{B}$ is polynomially controlled seems also somewhat unavoidable in theory, since $B(\boldsymbol{w})$ is an *upper bound* on the Lipschitz constant $\mathrm{Lip}(f_{\boldsymbol{w}})$. Thus a theoretical construction showing that the lower bound in (7.3.1) is tight (at some large depth D) would necessarily need to have an exponential gap between $\mathrm{Lip}(f_{\boldsymbol{w}})$ and $B(\boldsymbol{w})$. We are not aware of any such example, and it would be interesting to fully elucidate the role of depth in the law of robustness (particularly if it could give recommendation on how to best distribute parameters in a neural network).

7.4 Generalization Perspective

The law of robustness can be phrased in a slightly stronger way, as a generalization bound for classes of Lipschitz functions based on data-dependent Rademacher complexity. In particular, this perspective applies to any Lipschitz loss function, whereas our analysis in the main text was specific to the squared loss. We define the data-dependent Rademacher complexity $\mathrm{Rad}_{n,\mu}(\mathcal{F})$ by

$$\mathrm{Rad}_{n,\mu}(\mathcal{F}) = \frac{1}{n} \mathop{\mathbb{E}}^{\sigma_i, x_i} \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \sigma_i f(x_i) \right| \right] \tag{7.4.1}$$

where the values $(\sigma_i)_{i \in [n]}$ are i.i.d. symmetric Rademacher variables in $\{-1, 1\}$ while the values $(x_i)_{i \in [n]}$ are i.i.d. samples from μ.

Lemma 7.4.1. *Suppose $\mu = \sum_{i=1}^{k} \alpha_i \mu_i$ is a mixture of c-isoperimetric distributions. For finite $\mathcal{F}$ consisting of L-Lipschitz f with $|f(x)| \leq 1$ for all $(f, x) \in \mathcal{F} \times \mathbb{R}^d$, we have*

$$Rad_{n,\mu}(\mathcal{F}) \leq O\left(\max\left(\sqrt{\frac{k}{n}}, L\sqrt{\frac{c \log(|\mathcal{F}|)}{nd}} \right) \right). \tag{7.4.2}$$

The proof is identical to that of Theorem 35. Although we do not pursue it in detail, Lemma 7.2.2 easily extends to a sharpening of this result to general $\sigma_i \in [-1, 1]$ when $\mathbb{E}[\sigma_i^2]$ is small, even if σ_i and x_i are not independent. We only require that the n pairs $((\sigma_i, x_i))_{i \in [n]}$ are i.i.d. and that the distribution of σ_i given x_i is symmetric. To see that the latter symmetry condition is natural, recall the quantity $\mathrm{Rad}_{n,\mu}$ classically controls generalization due to the symmetrization trick, in which one writes $\sigma_i = y_i - y_i'$ for y_i' a resampled label for x_i. (In the modified proof, one would construct a noisy copy w_i of σ_i as in Lemma 7.5.1 and in a symmetric way, and then condition on $(|w_i|)_{i \in [n]}$ to preserve the symmetry, to replace the fact that f is not explicitly centered as in Lemma 7.2.2.)

Note that $\mathrm{Rad}_{n,\mu}(\mathcal{F})$ simply measures the ability of functions in $\mathcal{F}$ to correlate with random noise. Using standard machinery it implies the following generalization bound:

Corollary 7.4.2. *For any loss function $\ell(t, y)$ which is bounded and 1-Lipschitz in its first argument and any $\delta \in [0, 1]$, in the setting of Lemma 7.4.1 we have with probability at least $1 - \delta$ the uniform convergence bound:*

$$\sup_{f \in \mathcal{F}} \left| \mathop{\mathbb{E}}^{(x,y) \sim \mu} [\ell(f(x), y)] - \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) \right| \leq O\left(\max\left(\sqrt{\frac{k}{n}}, L\sqrt{\frac{c \log(|\mathcal{F}|)}{nd}}, \sqrt{\frac{\log(1/\delta)}{n}} \right) \right).$$

Proof. Using McDiarmid's concentration inequality it is enough to bound the left hand side in

expectation over (x_i, y_i). Using the symmetrization trick, one reduces this task to upper bound

$$\mathop{\mathbb{E}}^{x_i,y_i,\sigma_i} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(f(x_i), y_i)\,.$$

Fixing the pairs (x_i, y_i) and using the contraction lemma (see e.g., [SSBD14, Theorem 26.9]) the above quantity is upper bounded by $\mathrm{Rad}_{n,\mu}(\mathcal{F})$ which concludes the proof. □

Of course, one can again use an ε-net to obtain an analogous result for continuously parametrized function classes. The law of robustness, now for a general loss function, follows as a corollary (the argument is similar to [Proposition 1, [BELM20]]). Let us point out that many works have studied the Rademacher complexity of function classes such as neural networks – see e.g. [BFT17], or [YKB19] in the context of adversarial examples. The new feature of our result is that isoperimetry of the covariates yields improved generalization guarantees.

7.5 Proof of Lemma 7.2.2

We begin with a preliminary lemma.

Lemma 7.5.1. *Let z be a random variable such that $z \in [-2, 2]$ holds almost surely and $\mathbb{E}[z] = 0$, and fix $\sigma \in (0, 1]$. Let $\Gamma = 10\lceil \log(1 + \sigma^{-1})\rceil$. There exists a random variable w such that:*

1. *$\mathbb{E}[w|z] = z$.*
2. *$|w| \in \{4, 2, 1, \ldots, 2^{-\Gamma}, 0\}$ almost surely.*
3. *$\mathbb{E}[w^2] \le 16\mathbb{E}[z^2] + 4\sigma^2$.*
4. *The inequality*

$$\frac{\mathbb{P}[z \in A | w = a]}{\mathbb{P}[z \in A]} \le (10/\sigma)^3.$$

holds almost surely for all $a \in \{\pm 4, \pm 2, \pm 1, \ldots, \pm 2^{-\Gamma}, 0\}$ and $A \subseteq \mathbb{R}$ such that $\mathbb{P}[z \in A] > 0$.

In the proof, we will say that (a, b) is a martingale coupling if $\mathbb{E}[b|a] = a$. Thus, the first conclusion above states that (z, w) is a martingale coupling.

Proof of Lemma 7.5.1. For $a > 0$, let $r(a) = 2^{-\lfloor \log_2(a)\rfloor}$ be the smallest power of 2 larger than a. It is easy to see that there exists a martingale coupling $(z, \tilde{z})$ such that both $\mathbb{P}[\tilde{z} = 0|z] = \frac{1}{10}$ and

$$|\tilde{z}| \in \{2r(|z|), 0\}$$

holds almost surely. For such $\tilde{z}$, we have

$$\mathbb{E}[\tilde{z}^2|z] \leq 4r(|z|)^2 \leq 16z^2.$$

In the case $2r(|z|) \leq 2^{-\Gamma}$, we may create a second martingale coupling $(\tilde{z}, \hat{z})$ such that

$$|\hat{z}| \in \{2^{-\Gamma}, 0\}$$

where $\tilde{z} = 0$ if and only if $\hat{z} = 0$. Moreover we require that $\tilde{z}$ and z are independent conditionally on $\hat{z}$.

$$\mathbb{E}[\hat{z}^2 - \tilde{z}^2|\tilde{z}] \leq 2^{-2\Gamma} \leq \sigma^2.$$

Letting $\hat{z} = \tilde{z}$ if $2r(|z|) \geq 2^{-\Gamma}$, we conclude that $(z, \hat{z})$ is a martingale coupling such that:

- $\hat{z} \in \left\{\max\left(2^{-\Gamma}, 2r(|z|)\right)\right\}$.
- $\mathbb{P}[\hat{z} = 0|z] = 1/10$.
- $\mathbb{E}[\hat{z}^2] \leq 16\mathbb{E}[z^2] + \sigma^2$.

Finally we let $(\hat{z}, w)$ be a martingale coupling where $\hat{z} = w$ almost surely when $\hat{z} \neq 0$. If $\hat{z} = 0$, then we take

$$w = \hat{Z} \in \left\{ \pm 4, \pm 2, \pm 1, \ldots, \pm 2^{-\Gamma}, 0\right\}$$

independently of $(z, \tilde{z})$ with

$$\mathbb{P}[\hat{Z} = 2^k] = \mathbb{P}[\hat{Z} = -2^k] = \min\left(\frac{1}{2\Gamma + 6}, \frac{\sigma^2}{2^{2k}\Gamma}\right), \quad k \in \{2, 1, 0, -1, \ldots, -\Gamma\}.$$

The term $\frac{1}{2\Gamma+6}$ ensures these probabilities sum to at most 1, and we take $\hat{Z} = 0$ with the remaining probability. It is easy to see that (z, w) thus constructed is indeed a martingale coupling, verifying the first desired statement.

The second statement to be proved, namely that $|w| \in \left\{4, 2, 1, \ldots, 2^{-\Gamma}, 0\right\}$, holds by construction. For the third, we have

$$\begin{aligned}
\mathbb{E}[w^2] &\leq \mathbb{E}[\hat{z}^2] + \mathbb{E}[\hat{Z}^2] \\
&\leq 16\mathbb{E}[z^2] + \sigma^2 + \frac{2}{10}\sum_{k=-\Gamma}^{2} \frac{\sigma^2}{2^{2k}\Gamma} 2^{2k} \\
&\leq 16\mathbb{E}[z^2] + 4\sigma^2.
\end{aligned}$$

Finally for the fourth, we observe that for any value $a \in \left\{ \pm 4, \pm 2, \pm 1, \ldots, \pm 2^{-\Gamma}, 0\right\}$, we have almost

surely

$$\mathbb{P}[w = a|z] \geq \frac{1}{10} \min\left(\frac{1}{2\Gamma + 6}, \frac{\sigma^2}{(a^2+1)\Gamma}\right) \geq \frac{\sigma^2}{100\Gamma}.$$

(Here we write $a^2 + 1$ instead of a^2 simply to avoid division by 0 when $a = 0$.) Using Bayes' rule, we find

$$\begin{aligned}\frac{\mathbb{P}[z \in A|w = a]}{\mathbb{P}[z \in A]} &= \frac{\mathbb{P}[w = a|z \in A]}{\mathbb{P}[w = a]} \\ &\leq \frac{1}{\mathbb{E}^z\left[\mathbb{P}[w = a|z]\right]} \\ &\leq \frac{100\Gamma}{\sigma^2}.\end{aligned}$$

for any $a \in \left\{\pm 4, \pm 2, \pm 1, \ldots, \pm 2^{-\Gamma}, 0\right\}$ and any set $A \subseteq \mathbb{R}$ such that $\mathbb{P}(z \in A) > 0$. Observing that $\Gamma \leq \frac{10}{\sigma}$ for $\sigma \leq 1$ implies the result. □

To prove Lemma 7.2.2, we will apply Lemma 7.5.1 to construct variables w_i from each z_i. These can be viewed as noisy realizations of z_i (with a delicate choice of noise). The last conclusion of Lemma 7.5.1 ensures that $(x_1, \ldots, x_n)$ is independent of $(w_1, \ldots, w_n)$ up to a "likelihood distortion factor" $e^{O(n)}$. Since we are in the end concerned with probabilities of order $e^{-\Omega(nd)}$, the factor $e^{O(n)}$ can be absorbed. We then argue that

$$\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - \overset{\mu_{\ell_i}}{\mathbb{E}}[f])w_i$$

is $O\left(\frac{cL^2 \cdot \frac{1}{n}\sum_{i=1}^n w_i^2}{nd}\right)$-subgaussian "modulo" the likelihood distortion. Moreover the first conclusion of Lemma 7.5.1 ensures that this sum dominates the quantity of interest

$$\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - \overset{\mu_{\ell_i}}{\mathbb{E}}[f])z_i.$$

Proof of Lemma 7.2.2. For each $i \in [n]$, apply Lemma 7.5.1 to z_i conditionally on the component μ_{ℓ_i} (and independently of x_i) to construct w_i. In applying Lemma 7.5.1, we take the same value σ as in Lemma 7.2.2. Then by the first guarantee,

$$\mathbb{E}[(f(x_i) - \overset{\mu_{\ell_i}}{\mathbb{E}}[f])w_i \mid \mathcal{G}] = (f(x_i) - \overset{\mu_{\ell_i}}{\mathbb{E}}[f])z_i$$

where $\mathcal{G}$ is the sigma algebra generated by $(x_i, w_i)_{i\in[n]}$. Hence the difference

$$D \equiv \sum_{i\in[n]}(f(x_i) - \overset{\mu_{\ell_i}}{\mathbb{E}}[f])(w_i - z_i)$$

is, conditionally on $\mathcal{G}$, an independent sum of random variables $A_i \equiv (f(x_i) - \mathbb{E}^{\mu_{\ell_i}}[f])(w_i - z_i)$ satisfying:

- $\mathbb{E}[A_i \mid \mathcal{G}] = 0$.
- $|A_i| \le 20$ almost surely.

Applying the Berry-Esseen theorem conditionally on $\mathcal{G}$ and letting $V_i = \mathrm{Var}[A_i \mid \mathcal{G}]$ be the conditional variance of A_i, we find

$$\left|\mathbb{P}[D \ge 0 \mid \mathcal{G}] - \frac{1}{2}\right| \le 20\left(\sum_{i=1}^n V_i\right)^{-1/2}.$$

With $V = \sum_{i=1}^n V_i$, we have

$$\begin{aligned}\mathbb{P}[D \ge -400 \mid \mathcal{G}] &\ge \mathbb{P}[D \ge 0 \mid \mathcal{G}] \\ &\ge \frac{1}{2} - 20V^{-1/2} \\ &\ge \frac{1}{4}\end{aligned}$$

whenever $V \ge 10^4$. On the other hand, the Chebyshev inequality implies

$$\mathbb{P}[D \ge -400 \mid \mathcal{G}] \ge \frac{1}{4}$$

when $V \le 10^4$. Hence in either case, $\mathbb{P}[D \ge -400 \mid \mathcal{G}]$ holds. For $\varepsilon \ge \frac{6400}{n}$ we conclude:

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (f(x_i) - \overset{\mu_{\ell_i}}{\mathbb{E}}[f])w_i \ge \frac{\varepsilon}{16} \;\middle|\; \mathcal{G}\right) &\ge \frac{1}{4}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (f(x_i) - \overset{\mu_{\ell_i}}{\mathbb{E}}[f])z_i \ge \frac{\varepsilon}{16} + \frac{400}{n} \;\middle|\; \mathcal{G}\right) \\ &\ge \frac{1}{4}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (f(x_i) - \overset{\mu_{\ell_i}}{\mathbb{E}}[f])z_i \ge \frac{\varepsilon}{8} \;\middle|\; \mathcal{G}\right).\end{aligned} \tag{7.5.1}$$

We now turn to upper-bounding the left hand side of (7.5.1). By the last guarantee in Lemma 7.5.1, we find

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (f(x_i) - \overset{\mu_{\ell_i}}{\mathbb{E}}[f])w_i \ge \frac{\varepsilon}{16}\right) \le \left(\frac{10}{\sigma}\right)^{3n} \cdot \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (f(\hat{x}_i) - \overset{\mu_{\ell_i}}{\mathbb{E}}[f])w_i \ge \frac{\varepsilon}{16}\right) \tag{7.5.2}$$

where $\hat{x}_i \sim \mu_{\ell_i}$ is an independent copy of x_i in the same component. Indeed, the bound on the Radon–Nikodym derivative transfers to the law of x_i since x_i and w_i are conditionally independent given z_i. The final step is to apply subgaussian concentration results to the right side of (7.5.2). This is very easy thanks to the independence between $\hat{x}_i$ and w_i.

First, for $i \in [n]$ we have $w_i \in [-4, 4]$ almost surely. Next, let $\sigma_\ell^2 = \mathbb{E}[z_i^2 | \ell_i = \ell]$ be the average

label noise level over $x \sim \mu_\ell$, so that $\sum_{\ell=1}^k \alpha_\ell \sigma_\ell^2 = \sigma^2$. Then Lemma 7.5.1 ensures $\mathbb{E}[w_i^2|\ell_i] \le 16\sigma_{\ell_i}^2 + 4\sigma^2$ and so $\mathbb{E}[w_i^2] \le 20\sigma^2$. It follows that the random variables $B_i \equiv w_i^2 - \mathbb{E}[w_i^2]$ satisfy

$$|B_i| \le 16$$

and

$$\begin{aligned}\mathbb{E}[B_i^2] &\le \mathbb{E}[w_i^4] + \mathbb{E}[w_i^2]^2 \\ &\le 2 \times 20\sigma^2 \times 16 \\ &= 640\sigma^2.\end{aligned}$$

Bernstein's inequality implies

$$\begin{aligned}\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n w_i^2 \ge 40\sigma^2\right] &= \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n B_i \ge 20\sigma^2\right] \\ &\le \exp\left(-\frac{400\sigma^4 n^2}{2 \times (640 n\sigma^2) + 320\sigma^2}\right) \\ &\le \exp\left(-\frac{n\sigma^2}{4}\right). \end{aligned} \tag{7.5.3}$$

Assuming $\frac{1}{n}\sum_{i=1}^n w_i^2 \le 40\sigma^2$, the value

$$\frac{1}{n}\sum_{i=1}^n (f(\hat{x}_i) - \overset{\mu_{\ell_i}}{\mathbb{E}}[f]) w_i$$

is $\frac{720cL^2\sigma^2}{nd}$-subgaussian by Proposition 7.1.2. Therefore conditioned on $(w_i)_{i\in[n]}$ such that $\frac{1}{n}\sum_{i=1}^n w_i^2 \le 40\sigma^2$ holds, we have (using $8^6 \ge 16^2 \times 720$) that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (f(\hat{x}_i) - \overset{\mu_{\ell_i}}{\mathbb{E}}[f]) w_i \ge \frac{\varepsilon}{16}\right) \le 2\exp\left(-\frac{\varepsilon^2 nd}{8^6 cL^2\sigma^2}\right). \tag{7.5.4}$$

Next for a finite function class $\mathcal{F}$ we write

$$\begin{aligned}&\mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^n (f(x_i) - \overset{\mu_{\ell_i}}{\mathbb{E}}[f]) z_i \ge \frac{\varepsilon}{8}\right) \\ &\overset{(7.5.3)}{\le} |\mathcal{F}| \cdot \sup_{f\in\mathcal{F}} \sup_{\substack{(w_1,\dots,w_n): \\ \frac{1}{n}\sum_{i=1}^n w_i^2 \le 40\sigma^2}} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (f(x_i) - \overset{\mu_{\ell_i}}{\mathbb{E}}[f]) z_i \ge \frac{\varepsilon}{8} \;\middle|\; (w_i)_{i\in[n]}\right) + \exp\left(-\frac{n\sigma^2}{4}\right).\end{aligned}$$

We now estimate the main term on the right side above. For any $f \in \mathcal{F}$ and any sequence $\vec{w} =$

$(w_1, \dots, w_n)$ such that $\frac{1}{n}\sum_{i=1}^n w_i^2 \le 40\sigma^2$, we have

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (f(x_i) - \overset{\mu_{\ell_i}}{\mathbb{E}}[f])z_i \ge \frac{\varepsilon}{8} \,\Big|\, \vec{w}\right) &\overset{(7.5.1)}{\le} 4\cdot\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (f(x_i) - \overset{\mu_{\ell_i}}{\mathbb{E}}[f])w_i \ge \frac{\varepsilon}{16} \,\Big|\, \vec{w}\right) \\ &\overset{(7.5.2)}{\le} 4\left(\frac{10}{\sigma}\right)^{3n}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (f(\hat{x}_i) - \overset{\mu_{\ell_i}}{\mathbb{E}}[f])w_i \ge \frac{\varepsilon}{16} \,\Big|\, \vec{w}\right) \\ &\overset{(7.5.4)}{\le} \left(\frac{20}{\sigma}\right)^{3n}\exp\left(-\frac{\varepsilon^2 nd}{8^6 cL^2\sigma^2}\right).\end{aligned}$$

Combining the final two displays completes the proof. □

7.6 Necessity of Polynomially Bounded Weights

In [BLN21] it was conjectured that the law of robustness should hold for the class of *all* two-layer neural networks. In this chapter we prove that in fact it holds for arbitrary smoothly parametrized function classes, as long as the parameters are of size at most polynomial in the dimension d. In this section we demonstrate that this polynomial size restriction is necessary for bounded depth neural networks.

First we note that *some* restriction on the size of the parameters is certainly necessary in the most general case. Indeed one can build a single-parameter family, where the single real parameter is used to approximately encode all Lipschitz functions from a compact set in $\mathbb{R}^d$ to $[-1, 1]$, simply by brute-force enumeration. In particular no tradeoff between number of parameters and attainable Lipschitz constant would exist for this function class.

Showing a counter-example to the law of robustness with unbounded parameters and "reasonable" function classes is slightly harder. Here we build a three-layer neural network, with a single fixed nonlinearity $\sigma : \mathbb{R} \to \mathbb{R}$, but the latter is rather complicated and we do not know how to describe it explicitly (it is based on the Kolmogorov-Arnold theorem). It would be interesting to give similar constructions using other function classes such as ReLU networks.

Theorem 38. *For each $d \in \mathbb{Z}^+$ there is a continuous function $\sigma : \mathbb{R} \to \mathbb{R}$ and a sequence $(b_\ell)_{\ell \le 2^{2^d}}$ such that the following holds. The function Φ_a defined by*

$$\Phi_a(x) = \sum_{\ell=1}^{2^{2^d}} \sigma(a-\ell)\sum_{i=1}^{2d}\sigma\left(b_\ell + \sum_{j=1}^{d}\sigma(x_j + b_\ell)\right), \qquad |a| \le 2^{2^d} \tag{7.6.1}$$

is always $O(d^{3/2})$-Lipschitz, and the parametrization $a \to \Phi_a$ is 1-Lipschitz. Moreover for $n \le \frac{2^d}{100}$, given i.i.d. uniform points $x_1, \dots, x_n \in \mathbb{S}^{d-1}$ and random labels $y_1, \dots, y_n \in \{-1, 1\}$, with probability

$1 - e^{-\Omega(d)}$ *there exists* $\ell \in [2^{2^d}]$ *such that* $\Phi_\ell(x_i) = y_i$ *for at least* $\frac{3n}{4}$ *of the values* $i \in [n]$.

Proof. For each coordinate $i \in [d]$, define the slab $\texttt{slab}_i = \{x \in \mathbb{S}^{d-1} : |x_i| \leq \frac{1}{100d^{3/2}}\}$ and set $\texttt{slab} = \bigcup_{i\in[d]} \texttt{slab}_i$. Then it is not difficult to see that $\mu(\texttt{slab}) \leq \frac{1}{10}$. We partition $\mathbb{S}^{d-1}\backslash\texttt{slab}$ into its 2^d connected components, which are characterized by their sign patterns in $\{-1,1\}^d$; this defines a piece-wise constant function $\gamma : \mathbb{S}^{d-1}\backslash\texttt{slab} \to \{-1,1\}^d$. If we sample the points $x_1, \ldots, x_n$ sequentially, each point has probability at least $\frac{4}{5}$ to be in a new cell - this implies that with probability $1 - e^{-\Omega(n)}$, at least $\frac{3n}{4}$ are in a unique cell. It therefore suffices to give a construction that achieves $\Phi(x_i) = y_i$ for all $x_i \notin \texttt{slab}$ such that $\gamma(x_i) \neq \gamma(x_j)$ for all $j \in [n]\backslash\{i\}$. We do this now.

For each of the 2^{2^d} functions $g_\ell : \{-1,1\}^d \to \{-1,1\}$, we now obtain the partial function $\tilde{h}_\ell = g_\ell \circ \gamma : \mathbb{S}^{d-1}\backslash\texttt{slab} \to \{-1,1\}$. By the Kirszbraun extension theorem, $\tilde{h}_\ell$ extends to an $O(d^{3/2})$-Lipschitz function $h_\ell : \mathbb{S}^{d-1} \to [-1,1]$ on the whole sphere. The Kolmogorov-Arnold theorem guarantees the existence of an exact representation

$$\Phi_\ell(x) = \sum_{i=1}^{2d} \sigma_\ell \left(\sum_{j=1}^{d} \sigma_\ell(x_j) \right) \tag{7.6.2}$$

of h_ℓ by a two-layer neural network for some continuous function $\sigma_\ell : \mathbb{R} \to \mathbb{R}$ depending on ℓ. It suffices to give a single neural network capable of computing all functions $(\Phi_\ell)_{\ell=1}^{2^{2^d}}$. We extend the definition of Φ_a to any $a \in \mathbb{R}$ via:

$$\Phi_a(x) = \sum_{\ell=1}^{2^{2^d}} \sigma(a - \ell)\Phi_\ell(x) \tag{7.6.3}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ satisfies $\sigma(x) = (1 - |x|)_+$ for $|x| \leq 2^{2^d}$. This ensures that (7.6.3) extends (7.6.2). To express Φ_a using only a single non-linearity, we prescribe further values for σ. Let

$$U = 2^{2^d} + d \cdot \max_{x\in[-1,1],\ell\in[2^{2^d}]} |\sigma_\ell(x)|$$

so that $\left|\sum_{j=1}^{d} \sigma_\ell(x_j)\right| \leq U$ for all $x \in \mathbb{S}^{d-1}$. Define real numbers $b_\ell = 10\ell U + 2^{2^d}$ for $\ell \in [2^{2^d}]$ and for all $|x| \leq U$ set

$$\sigma(x + b_\ell) = \sigma_\ell(x).$$

Due to the separation of the values b_ℓ such a function σ certainly exists. Then we have

$$\Phi_\ell(x) = \sum_{i=1}^{2d} \sigma\left(b_\ell + \sum_{j=1}^{d} \sigma(x_j + b_\ell)\right).$$

Therefore with this choice of non-linearity σ and (data-independent) constants b_ℓ, some function Φ_ℓ fits at least $\frac{3n}{4}$ of the n data points with high probability, and the functions Φ_a are parametrized in a 1-Lipschitz way by a single real number $a \leq 2^{2^d}$. □

Remark 7.6.1. The representation (7.6.1) is a three-layer neural network because the $\sigma(a - \ell)$ terms are just matrix entries for the final layer.

Remark 7.6.2. The construction above can be made more efficient, using only $O(n \cdot 2^n)$ uniformly random functions $g_\ell : \{-1, 1\}^d \to \{-1, 1\}$ instead of all 2^{2^ℓ}. Indeed by the coupon collector problem, this results in all functions from $\{\gamma(x_i) : i \in [n]\} \to \{-1, 1\}$ being expressable as the restriction of some g_ℓ, with high probability.

Bibliography

[AAA03] Noga Alon, Baruch Awerbuch, and Yossi Azar. The online set cover problem. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 100–105, 2003.

[AAZB+17] Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1195–1199, 2017.

[ABA13] Antonio Auffinger and Gérard Ben Arous. Complexity of random smooth functions on the high-dimensional sphere. *The Annals of Probability*, 41(6):4214–4247, 2013.

[ABAČ13] Antonio Auffinger, Gérard Ben Arous, and Jiří Černý. Random matrices and complexity of spin glasses. *Communications on Pure and Applied Mathematics*, 66(2):165–201, 2013.

[ABC+19] CJ Argue, Sébastien Bubeck, Michael B Cohen, Anupam Gupta, and Yin Tat Lee. A nearly-linear bound for chasing nested convex bodies. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 117–122. SIAM, 2019.

[ABE+05] Sanjeev Arora, Eli Berger, Hazan Elad, Guy Kindler, and Muli Safra. On non-approximability for quadratic programs. In *Proceedings of 46th FOCS*, pages 206–215. IEEE, 2005.

[ABN+16] Antonios Antoniadis, Neal Barcelo, Michael Nugent, Kirk Pruhs, Kevin Schewior, and Michele Scquizzato. Chasing convex bodies and functions. In *LATIN 2016: Theoretical Informatics*, pages 68–81. Springer, 2016.

[AC15] Antonio Auffinger and Wei-Kuo Chen. The Parisi formula has a unique minimizer. *Communications in Mathematical Physics*, 335(3):1429–1444, 2015.

[AC17a] Antonio Auffinger and Wei-Kuo Chen. On the energy landscape of spherical spin glasses. *Advances in Mathematics*, 330, 02 2017.

[AC17b] Antonio Auffinger and Wei-Kuo Chen. Parisi formula for the ground state energy in the mixed p-spin model. *The Annals of Probability*, 45(6b):4617–4631, 2017.

[ACO08] Dimitris Achlioptas and Amin Coja-Oghlan. Algorithmic barriers from phase transitions. In *Proceedings of 49th FOCS*, pages 793–802, 2008.

[ACORT11] Dimitris Achlioptas, Amin Coja-Oghlan, and Federico Ricci-Tersenghi. On the solution-space geometry of random constraint satisfaction problems. *Random Structures & Algorithms*, 38(3):251–268, 2011.

[ACZ17] Antonio Auffinger, Wei-Kuo Chen, and Qiang Zeng. The SK model is Full-step Replica Symmetry Breaking at zero temperature. arXiv:1703.06872, 2017.

[ADS12] Sami Assaf, Persi Diaconis, and Kannan Soundararajan. Riffle shuffles with biased cuts. *Discrete Mathematics & Theoretical Computer Science*, 2012.

[AG97] G Ben Arous and Alice Guionnet. Large deviations for wigner's law and voiculescu's non-commutative entropy. *Probability theory and related fields*, 108(4):517–542, 1997.

[AGGT21] CJ Argue, Anupam Gupta, Guru Guruganesh, and Ziye Tang. Chasing convex bodies with linear competitive ratio. *Journal of the ACM (JACM)*, 68(5):1–10, 2021.

[AH87] Michael Aizenman and Richard Holley. Rapid convergence to equilibrium of stochastic Ising models in the Dobrushin Shlosman regime. In *Percolation theory and ergodic theory of infinite particle systems*, pages 1–11. Springer, 1987.

[AJK+21] Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Entropic Independence I: Modified Log-Sobolev Inequalities for Fractionally Log-Concave Distributions and High-Temperature Ising Models. *arXiv preprint arXiv:2106.04105*, 2021.

[Ald83] David Aldous. Random walks on finite groups and rapidly mixing markov chains. In *Séminaire de Probabilités XVII 1981/82*, pages 243–297. Springer, 1983.

[Ald90] David J Aldous. The random walk construction of uniform spanning trees and uniform labelled trees. *SIAM Journal on Discrete Mathematics*, 3(4):450–465, 1990.

[ALGV19] Nima Anari, Kuikui Liu, Shayan Oveis Gharan, and Cynthia Vinzant. Log-concave polynomials ii: high-dimensional walks and an fpras for counting bases of a matroid. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 1–12, 2019.

[ALR87] Michael Aizenman, Joel L Lebowitz, and David Ruelle. Some rigorous results on the Sherrington–Kirkpatrick spin glass model. *Communications in Mathematical Physics*, 112(1):3–20, 1987.

[AM20] Ahmed El Alaoui and Andrea Montanari. Algorithmic thresholds in mean field spin glasses. *arXiv preprint arXiv:2009.11481*, 2020.

[AMMN19] Gerard Ben Arous, Song Mei, Andrea Montanari, and Mihai Nica. The landscape of the spiked tensor model. *Communications on Pure and Applied Mathematics*, 72(11):2282–2330, 2019.

[AMS21] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Optimization of mean-field spin glasses. *Ann. Probab.*, 49(6):2922–2960, 2021.

[AMS22] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from the Sherrington-Kirkpatrick Gibbs measure via Algorithmic Stochastic Localization. *Foundations of Computer Science (FOCS)*, 2022.

[ASB16] Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *International Conference on Machine Learning*, pages 1120–1128. PMLR, 2016.

[AWZ20] Gérard Ben Arous, Alexander S Wein, and Ilias Zadik. Free energy wells and overlap gap property in sparse PCA. In *Conference on Learning Theory*, pages 479–482. PMLR, 2020.

[BADG06] Gérard Ben Arous, Amir Dembo, and Alice Guionnet. Cugliandolo-Kurchan equations for dynamics of spin-glasses. *Probability Theory and Related Fields*, 136(4):619–660, 2006.

[BAGJ20] Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Bounding flows for spherical spin glass dynamics. *Communications in Mathematical Physics*, 373(3):1011–1048, 2020.

[BAJ18] Gérard Ben Arous and Aukosh Jagannath. Spectral gap estimates in mean field spin glasses. *Communications in Mathematical Physics*, 361(1):1–52, 2018.

[Bal97] Keith Ball. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31:1–58, 1997.

[BASZ20] Gérard Ben Arous, Eliran Subag, and Ofer Zeitouni. Geometry and temperature chaos in mixed spherical spin glasses at low temperature: the perturbative regime. *Communications on Pure and Applied Mathematics*, 73(8):1732–1828, 2020.

[BB00] Avrim Blum and Carl Burch. On-line learning and the metrical task system problem. *Machine Learning*, 39(1):35–58, 2000.

[BB19] Roland Bauerschmidt and Thierry Bodineau. A very simple proof of the LSI for high temperature spin systems. *Journal of Functional Analysis*, 276(8):2582–2588, 2019.

[BBE+18] Nikhil Bansal, Martin Böhm, Marek Eliáš, Grigorios Koumoutsos, and Seeun William Umboh. Nested convex bodies are chaseable. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1253–1260. SIAM, 2018.

[BBH+12] Boaz Barak, Fernando G.S.L. Brandão, Aram W Harrow, Jonathan Kelner, David Steurer, and Yuan Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 307–326. ACM, 2012.

[BBM06] Yair Bartal, Béla Bollobás, and Manor Mendel. Ramsey-type theorems for metric spaces with applications to online problems. *Journal of Computer and System Sciences*, 72(5):890–921, 2006.

[BBMN15] Nikhil Bansal, Niv Buchbinder, Aleksander Madry, and Joseph Naor. A polylogarithmic-competitive algorithm for the k-server problem. *Journal of the ACM (JACM)*, 62(5):1–49, 2015.

[BCKM98] Jean-Philippe Bouchaud, Leticia F Cugliandolo, Jorge Kurchan, and Marc Mézard. Out of equilibrium dynamics in spin-glasses and other glassy systems. *Spin glasses and random fields*, pages 161–223, 1998.

[BCLL19] Sébastien Bubeck, Michael B Cohen, James R Lee, and Yin Tat Lee. Metrical task systems on trees via mirror descent and unfair gluing. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 89–97. SIAM, 2019.

[BČNS21] David Belius, Jiří Černý, Shuta Nakajima, and Marius Schmidt. Triviality of the geometry of mixed p-spin spherical hamiltonians with external field. *arXiv preprint arXiv:2104.06345*, 2021.

[BD92] Dave Bayer and Persi Diaconis. Trailing the dovetail shuffle to its lair. *Annals of Applied Probability*, 2(2):294–313, 1992.

[BD98a] Sebastian Böcker and Andreas W.M. Dress. Recovering symbolically dated, rooted trees from symbolic ultrametrics. *Advances in mathematics*, 138(1):105–125, 1998.

[BD98b] Kenneth S Brown and Persi Diaconis. Random walks and hyperplane arrangements. *Annals of Probability*, pages 1813–1854, 1998.

[BD11] Joseph Blitzstein and Persi Diaconis. A sequential importance sampling algorithm for generating random graphs with prescribed degrees. *Internet Mathematics*, 6(4):489–522, 2011.

[BELM20] Sebastien Bubeck, Ronen Eldan, Yin Tat Lee, and Dan Mikulincer. Network size and size of the weights in memorization with two-layers neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 4977–4986, 2020.

[BFT17] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[BGK+15] Nikhil Bansal, Anupam Gupta, Ravishankar Krishnaswamy, Kirk Pruhs, Kevin Schewior, and Cliff Stein. A 2-competitive algorithm for online convex optimization with switching costs. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2015)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2015.

[BH21] Guy Bresler and Brice Huang. The Algorithmic Phase Transition of Random k-SAT for Low Degree Polynomials. In *Proceedings of 62nd FOCS*, pages 298–309, 2021.

[BHMM19] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[BHR99] Pat Bidigare, Phil Hanlon, and Dan Rockmore. A combinatorial description of the spectrum for the tsetlin library and its generalization to hyperplane arrangements. *Duke Mathematical Journal*, 99(1):135–174, 1999.

[BKL+20] Sébastien Bubeck, Bo'az Klartag, Yin Tat Lee, Yuanzhi Li, and Mark Sellke. Chasing Nested Convex Bodies Nearly Optimally. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1496–1508. SIAM, 2020.

[BL97] Sergey Bobkov and Michel Ledoux. Poincaré's inequalities and talagrand's concentration phenomenon for the exponential distribution. *Probability Theory and Related Fields*, 107(3):383–400, 1997.

[BL00] Sergey G Bobkov and Michel Ledoux. From brunn-minkowski to brascamp-lieb and to logarithmic sobolev inequalities. *Geometric & Functional Analysis GAFA*, 10(5):1028–1052, 2000.

[BLLS19] Sébastien Bubeck, Yin Tat Lee, Yuanzhi Li, and Mark Sellke. Competitively Chasing Convex Bodies. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 861–868, 2019.

[BLLT20] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[BLM15] Mohsen Bayati, Marc Lelarge, and Andrea Montanari. Universality in polytope phase transitions and message passing algorithms. *The Annals of Applied Probability*, 25(2):753–822, 2015.

[BLMN05] Yair Bartal, Nathan Linial, Manor Mendel, and Assaf Naor. On metric Ramsey-type phenomena. *Annals of Mathematics*, 162:643–710, 2005.

[BLN21] Sébastien Bubeck, Yuanzhi Li, and Dheeraj M Nagaraj. A Law of Robustness for Two-Layers Neural Networks. In *Conference on Learning Theory*, pages 804–820. PMLR, 2021.

[BLS92] Allan Borodin, Nathan Linial, and Michael E Saks. An optimal on-line algorithm for metrical task system. *Journal of the ACM (JACM)*, 39(4):745–763, 1992.

[BM11a] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

[BM11b] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. on Inform. Theory*, 57:764–785, 2011.

[BMN19] Raphaël Berthier, Andrea Montanari, and Phan-Minh Nguyen. State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of the IMA*, 2019.

[BN19] Megan Bernstein and Evita Nestoridi. Cutoff for random to random card shuffle. *The Annals of Probability*, 47(5):3303–3320, 2019.

[Bol14] Erwin Bolthausen. An iterative construction of solutions of the TAP equations for the Sherrington–Kirkpatrick model. *Communications in Mathematical Physics*, 325(1):333–366, 2014.

[Bor75] Christer Borell. The Brunn-Minkowski inequality in Gauss space. *Inventiones mathematicae*, 30(2):207–216, 1975.

[Bro89] Andrei Z Broder. Generating random spanning trees. In *FOCS*, volume 89, pages 442–447, 1989.

[BS21] Sébastien Bubeck and Mark Sellke. A Universal Law of Robustness via Isoperimetry. *Advances in Neural Information Processing Systems*, 34, 2021.

[Bub15] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[Bur09] Krzysztof Burdzy. Differentiability of stochastic flow of reflected brownian motions. *Electronic Journal of Probability*, 14:2182–2240, 2009.

[Can80] E Rodney Canfield. Application of the berry-esséen inequality to combinatorial estimates. *Journal of Combinatorial Theory, Series A*, 28(1):17–25, 1980.

[CBG+17] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, pages 854–863. PMLR, 2017.

[CCM21] Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572*, 2021.

[CCNW21] Hong-Bin Chen, Sinho Chewi, and Jonathan Niles-Weed. Dimension-free log-sobolev inequalities for mixture distributions. *Journal of Functional Analysis*, 281(11):109236, 2021.

[CDHL05] Yuguo Chen, Persi Diaconis, Susan P Holmes, and Jun S Liu. Sequential Monte Carlo methods for statistical analysis of tables. *Journal of the American Statistical Association*, 100(469):109–120, 2005.

[CDHS17] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. "Convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *International Conference on Machine Learning*, pages 654–663. PMLR, 2017.

[CDHS18] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

[CDHS19] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, pages 1–50, 2019.

[CDM+19] Sitan Chen, Michelle Delcourt, Ankur Moitra, Guillem Perarnau, and Luke Postle. Improved bounds for randomly sampling colorings via linear programming. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2216–2234. SIAM, 2019.

[CE22] Yuansi Chen and Ronen Eldan. Localization schemes: A framework for proving mixing bounds for Markov chains. *Foundations of Computer Science (FOCS)*, 2022.

[CFM21] Michael Celentano, Zhou Fan, and Song Mei. Local convexity of the TAP free energy and AMP convergence for $\mathbf{Z}_2$-synchronization. *arXiv preprint arXiv:2106.11428*, 2021.

[CGPR19] Wei-Kuo Chen, David Gamarnik, Dmitry Panchenko, and Mustazee Rahman. Suboptimality of local algorithms for a class of max-cut problems. *The Annals of Probability*, 47(3):1587–1618, 2019.

[CGW18] Niangjun Chen, Gautam Goel, and Adam Wierman. Smoothed online convex optimization in high dimensions via online balanced descent. In *Conference On Learning Theory*, pages 1574–1594. PMLR, 2018.

[Cha09] Sourav Chatterjee. Disorder chaos and multiple valleys in spin glasses. *arXiv preprint arXiv:0907.3381*, 2009.

[Cha14] Sourav Chatterjee. *Superconcentration and related topics*, volume 15. Springer, 2014.

[Che13a] Wei-Kuo Chen. Disorder chaos in the Sherrington–Kirkpatrick model with external field. *The Annals of Probability*, 41(5):3345–3391, 2013.

[Che13b] Wei-Kuo Chen. The Aizenman-Sims-Starr scheme and Parisi formula for mixed p-spin spherical models. *Electronic Journal of Probability*, 18, 2013.

[Che17] Wei-Kuo Chen. Variational representations for the Parisi functional and the two-dimensional Guerra–Talagrand bound. *The Annals of Probability*, 45(6A):3929–3966, 2017.

[Che21] Yuansi Chen. An almost constant lower bound of the isoperimetric coefficient in the kls conjecture. *Geometric and Functional Analysis*, 31(1):34–61, 2021.

[CHHS15] Wei-Kuo Chen, Hsi-Wei Hsieh, Chii-Ruey Hwang, and Yuan-Chung Sheu. Disorder chaos in the spherical mean-field model. *Journal of Statistical Physics*, 160(2):417–429, 2015.

[CHL18] Wei-Kuo Chen, Madeline Handschy, and Gilad Lerman. On the energy landscape of the mixed even p-spin model. *Probability Theory and Related Fields*, 171(1-2):53–95, 2018.

[CIS76] Boris S. Cirel'son, Ildar A. Ibragimov, and V.N. Sudakov. Norms of Gaussian sample functions. In *Proceedings of the Third Japan—USSR Symposium on Probability Theory*, pages 20–41. Springer, 1976.

[CK94] Leticia F. Cugliandolo and Jorge Kurchan. On the out-of-equilibrium relaxation of the Sherrington-Kirkpatrick model. *Journal of Physics A: Mathematical and General*, 27(17):5749, 1994.

[CKP+14] Patrick Charbonneau, Jorge Kurchan, Giorgio Parisi, Pierfrancesco Urbani, and Francesco Zamponi. Exact theory of dense amorphous hard spheres in high dimension. iii. the full replica symmetry breaking solution. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(10):P10009, 2014.

[CL21] Wei-Kuo Chen and Wai-Kit Lam. Universality of approximate message passing algorithms. *Electronic Journal of Probability*, 26:1–44, 2021.

[CLR03] Andrea Crisanti, Luca Leuzzi, and Tommaso Rizzo. The complexity of the spherical p-spin spin glass model, revisited. *The European Physical Journal B-Condensed Matter and Complex Systems*, 36(1):129–136, 2003.

[CLR05] Andrea Crisanti, Luca Leuzzi, and Tommaso Rizzo. Complexity in mean-field spin-glass models: Ising p-spin. *Physical Review B*, 71(9):094202, 2005.

[COE15] Amin Coja-Oghlan and Charilaos Efthymiou. On independent sets in random graphs. *Random Structures & Algorithms*, 47(3):436–486, 2015.

[CP18] Wei-Kuo Chen and Dmitry Panchenko. Disorder chaos in some diluted spin glass models. *The Annals of Applied Probability*, 28(3):1356–1378, 2018.

[CPS19] Wei-Kuo Chen, Dmitry Panchenko, and Eliran Subag. The generalized tap free energy ii. *arXiv preprint arXiv:1903.01030*, 2019.

[CPS22] Wei-Kuo Chen, Dmitry Panchenko, and Eliran Subag. Generalized tap free energy. *Communications on Pure and Applied Mathematics*, n/a(n/a), 2022.

[CR02] Andrea Crisanti and Tommaso Rizzo. Analysis of the ∞-replica symmetry breaking solution of the Sherrington-Kirkpatrick model. *Physical Review E*, 65(4):046137, 2002.

[CS92] Andrea Crisanti and H-J Sommers. The spherical p-spin interaction spin glass model: the statics. *Zeitschrift für Physik B Condensed Matter*, 87(3):341–354, 1992.

[CS04] I Csiszar and PC Shields. Information theory and statistics: a tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4):417–417, 2004.

[CS17] Wei-Kuo Chen and Arnab Sen. Parisi formula, disorder chaos and fluctuation for the ground state energy in the spherical mixed p-spin models. *Communications in Mathematical Physics*, 350(1):129–173, 2017.

[CS21] Sourav Chatterjee and Leila Sloman. Average Gromov hyperbolicity and the Parisi ansatz. *Advances in Mathematics*, 376:107417, 2021.

[CT21] Wei-Kuo Chen and Si Tang. On Convergence of the Cavity and Bolthausen's TAP Iterations to the Local Magnetization. *Communications in Mathematical Physics*, 386(2):1209–1242, 2021.

[DAM17] Yash Deshpande, Emmanuel Abbe, and Andrea Montanari. Asymptotic mutual information for the balanced binary stochastic block model. *Information and Inference: A Journal of the IMA*, 6(2):125–170, 2017.

[DEZ15] Jian Ding, Ronen Eldan, and Alex Zhai. On multiple peaks and moderate deviations for the supremum of a gaussian field. *The Annals of Probability*, 43(6):3468–3493, 2015.

[DFP92] Persi Diaconis, James Allen Fill, and Jim Pitman. Analysis of top to random shuffles. *Combinatorics, Probability & Computing*, 1:135–155, 1992.

[DG95] Persi Diaconis and Anil Gangolli. Rectangular arrays with fixed margins. In *Discrete probability and algorithms*, pages 15–41. Springer, 1995.

[DG98] Martin Dyer and Catherine Greenhill. A more rapidly mixing markov chain for graph colorings. *Random Structures & Algorithms*, 13(3-4):285–317, 1998.

[DI91] Paul Dupuis and Hitoshi Ishii. On lipschitz continuity of the solution mapping to the skorokhod problem, with applications. *Stochastics: An International Journal of Probability and Stochastic Processes*, 35(1):31–62, 1991.

[Dia03] Persi Diaconis. Mathematical developments from the analysis of riffle shuffling. *Groups, combinatorics & geometry (Durham, 2001)*, pages 73–97, 2003.

[DKW56] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

[DMM10] David L. Donoho, Arian Maleki, and Andrea Montanari. Message Passing Algorithms for Compressed Sensing: I. Motivation and Construction. In *Proceedings of IEEE Inform. Theory Workshop*, Cairo, 2010.

[DS81] Persi Diaconis and Mehrdad Shahshahani. Generating a random permutation with random transpositions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57(2):159–179, 1981.

[Dur19] Rick Durrett. *Probability: Theory and Examples*, volume 49. Cambridge university press, 2019.

[DV13] Daniel Dadush and Santosh S Vempala. Near-optimal deterministic algorithms for volume computation via m-ellipsoids. *Proceedings of the National Academy of Sciences*, 110(48):19237–19245, 2013.

[EAM22] Ahmed El Alaoui and Andrea Montanari. An information-theoretic view of stochastic localization. *IEEE Transactions on Information Theory*, 2022.

[EKZ21] Ronen Eldan, Frederic Koehler, and Ofer Zeitouni. A spectral condition for spectral gap: fast mixing in high-temperature ising models. *Probability Theory and Related Fields*, pages 1–17, 2021.

[Eld13] Ronen Eldan. Thin shell implies spectral gap up to polylog via a stochastic localization scheme. *Geometric and Functional Analysis*, 23(2):532–569, 2013.

[Eld16] Ronen Eldan. Skorokhod embeddings via stochastic flows on the space of gaussian measures. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 52, pages 1259–1280. Institut Henri Poincaré, 2016.

[Eld20] Ronen Eldan. Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation. *Probability Theory and Related Fields*, 176(3):737–755, 2020.

[Eld22] Ronen Eldan. Analysis of high-dimensional distributions using pathwise methods. In *Proceedings of ICM*, 2022.

[ES22] Ronen Eldan and Omer Shamir. Log concavity and concentration of Lipschitz functions on the Boolean hypercube. *Journal of Functional Analysis*, page 109392, 2022.

[Fan22] Zhou Fan. Approximate message passing algorithms for rotationally invariant matrices. *The Annals of Statistics*, 50(1):197–224, 2022.

[FdRL17] Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):1–8, 2017.

[FKP94] Alan Frieze, Ravi Kannan, and Nick Polson. Sampling from log-concave distributions. *The Annals of Applied Probability*, pages 812–837, 1994.

[FL93] Joel Friedman and Nathan Linial. On convex body chasing. *Discrete & Computational Geometry*, 9(3):293–321, 1993.

[FM03] Amos Fiat and Manor Mendel. Better algorithms for unfair metrical task systems and applications. *SIAM Journal on Computing*, 32(6):1403–1422, 2003.

[FMN16] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

[FS18] Charles Fefferman and Pavel Shvartsman. Sharp finiteness principles for Lipschitz selections. *Geometric and Functional Analysis*, 28(6):1641–1705, 2018.

[Ful98] Jason Fulman. The combinatorics of biased riffle shuffles. *Combinatorica*, 18(2):173–184, 1998.

[Fyo13] Yan V. Fyodorov. High-dimensional random fields and random matrix theory. *arXiv preprint arXiv:1307.2379*, 2013.

[Gam21] David Gamarnik. The overlap gap property: A topological barrier to optimizing over random structures. *Proceedings of the National Academy of Sciences*, 118(41), 2021.

[GCL+19] Ruiqi Gao, Tianle Cai, Haochuan Li, Cho-Jui Hsieh, Liwei Wang, and Jason D Lee. Convergence of adversarial training in overparametrized neural networks. *Advances in Neural Information Processing Systems*, 32:13029–13040, 2019.

[GDG+19] Alexander Gasnikov, Pavel Dvurechensky, Eduard Gorbunov, Evgeniya Vorontsova, Daniil Selikhanovych, César A Uribe, Bo Jiang, Haoyue Wang, Shuzhong Zhang, Sébastien Bubeck, et al. Near optimal methods for minimizing convex functions with lipschitz p-th derivatives. In *Conference on Learning Theory*, pages 1392–1393. PMLR, 2019.

[GJ19] Reza Gheissari and Aukosh Jagannath. On the spectral gap of spherical spin glass dynamics. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 55, pages 756–776. Institut Henri Poincaré, 2019.

[GJ21] David Gamarnik and Aukosh Jagannath. The overlap gap property and approximate message passing algorithms for p-spin models. *The Annals of Probability*, 49(1):180–205, 2021.

[GJS19] David Gamarnik, Aukosh Jagannath, and Subhabrata Sen. The overlap gap property in principal submatrix recovery. *arXiv preprint arXiv:1908.09959*, 2019.

[GJW20a] David Gamarnik, Aukosh Jagannath, and Alexander S. Wein. Low-degree hardness of random optimization problems. In *Proceedings of 61st FOCS*, pages 131–140. IEEE, 2020.

[GJW20b] David Gamarnik, Aukosh Jagannath, and Alexander S Wein. Low-degree hardness of random optimization problems. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 131–140, 2020.

[GJW21] David Gamarnik, Aukosh Jagannath, and Alexander S. Wein. Circuit lower bounds for the p-spin optimization problem. *arXiv preprint arXiv:2109.01342*, 2021.

[GK21a] David Gamarnik and Eren C. Kizildag. Algorithmic obstructions in the random number partitioning problem. *arXiv preprint arXiv:2103.01369*, 2021.

[GK21b] David Gamarnik and Eren C. Kizildag. Computing the partition function of the Sherrington–Kirkpatrick model is hard on average. *The Annals of Applied Probability*, 31(3):1474–1504, 2021.

[GL18] David Gamarnik and Quan Li. Finding a large submatrix of a gaussian random matrix. *Annals of Statistics*, 46(6A):2511–2561, 2018.

[GLSW19] Gautam Goel, Yiheng Lin, Haoyuan Sun, and Adam Wierman. Beyond online balanced descent: An optimal algorithm for smoothed online optimization. *Advances in Neural Information Processing Systems*, 32:1875–1885, 2019.

[GMF+18] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018.

[GQU+20] Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020.

[Gra66] Ronald L Graham. Bounds for certain multiprocessing anomalies. *Bell system technical journal*, 45(9):1563–1581, 1966.

[Gro86] Mikhael Gromov. Isoperimetric inequalities in riemannian manifolds. In *Asymptotic Theory of Finite Dimensional Spaces*, volume 1200, pages 114–129. Springer Berlin, 1986.

[Gro91] Edward F Grove. The harmonic online k-server algorithm is competitive. In *Proceedings of the twenty-third annual ACM symposium on Theory of computing*, pages 260–266, 1991.

[GS14] David Gamarnik and Madhu Sudan. Limits of local algorithms over sparse random graphs. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 369–376. ACM, 2014.

[GS17] David Gamarnik and Madhu Sudan. Performance of sequential local algorithms for the random NAE-K-sat problem. *SIAM Journal on Computing*, 46(2):590–619, 2017.

[GSV05] D. Guo, S. Shamai, and S. Verdú. Mutual information and minimum mean-square error in gaussian channels. *IEEE Trans. Inform. Theory*, 51:1261–1282, 2005.

[Gue01] Francesco Guerra. Sum rules for the free energy in the mean field spin glass model. *Fields Institute Communications*, 30(11), 2001.

[GZ17] David Gamarnik and Ilias Zadik. Sparse high-dimensional linear regression. algorithmic barriers and a local search algorithm. *arXiv preprint arXiv:1711.04952*, 2017.

[GZ19] David Gamarnik and Ilias Zadik. The landscape of the planted clique problem: Dense subgraphs and the overlap gap property. *arXiv preprint arXiv:1904.07174*, 2019.

[HS89] Trevor Hastie and Werner Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989.

[HS22] Brice Huang and Mark Sellke. Tight Lipschitz Hardness for Optimizing Mean Field Spin Glasses. *Foundations of Computer Science (FOCS)*, 2022.

[IW77] Nobuyuki Ikeda and Shinzo Watanabe. A comparison theorem for solutions of stochastic differential equations and its applications. *Osaka Journal of Mathematics*, 14(3):619–633, 1977.

[Jag17] Aukosh Jagannath. Approximate ultrametricity for random measures and applications to spin glasses. *Communications on Pure and Applied Mathematics*, 70(4):611–664, 2017.

[JCC+19] Haoming Jiang, Zhehui Chen, Minshuo Chen, Feng Liu, Dingding Wang, and Tuo Zhao. On computation and generalization of gans with spectrum control. *Proc. of International Conference on Learning Representation (ICLR)*, 2019.

[JGN+17] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pages 1724–1732. PMLR, 2017.

[JLLV21] He Jia, Aditi Laddha, Yin Tat Lee, and Santosh Vempala. Reducing isotropy and volume to kls: an o*(n 3 ψ 2) volume algorithm. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 961–974, 2021.

[JM13] Adel Javanmard and Andrea Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2(2):115–144, 2013.

[JM15] Johan Jonasson and Benjamin Morris. Rapid mixing of dealer shuffles and clumpy shuffles. *Electronic Communications in Probability*, 20:1–11, 2015.

[JS89] Mark Jerrum and Alistair Sinclair. Approximating the permanent. *SIAM journal on computing*, 18(6):1149–1178, 1989.

[JT16] Aukosh Jagannath and Ian Tobasco. A dynamic programming approach to the parisi functional. *Proceedings of the American Mathematical Society*, 144(7):3135–3150, 2016.

[JT17] Aukosh Jagannath and Ian Tobasco. Low temperature asymptotics of spherical mean field spin glasses. *Communications in Mathematical Physics*, 352(3):979–1017, 2017.

[JVV86] Mark R Jerrum, Leslie G Valiant, and Vijay V Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical computer science*, 43:169–188, 1986.

[Kiv21] Pax Kivimae. The ground state energy and concentration of complexity in spherical bipartite models. *arXiv preprint arXiv:2107.13138*, 2021.

[KL93] Nanda Kambhatla and Todd K Leen. Fast nonlinear dimension reduction. In *IEEE International Conference on Neural Networks*, pages 1213–1218. IEEE, 1993.

[KMRT+07] Florent Krzakala, Andrea Montanari, Federico Ricci-Tersenghi, Guilhem Semerjian, and Lenka Zdeborová. Gibbs states and the set of solutions of random constraint satisfaction problems. *Proceedings of the National Academy of Sciences*, 104(25):10318–10323, 2007.

[KP95] Elias Koutsoupias and Christos H Papadimitriou. On the k-server conjecture. *Journal of the ACM (JACM)*, 42(5):971–983, 1995.

[KP21] Bo'az Klartag and Eli Putterman. Spectral monotonicity under Gaussian convolution. *arXiv preprint arXiv:2107.09496*, 2021.

[KPS18] Ravi Kumar, Manish Purohit, and Zoya Svitkina. Improving Online Algorithms via ML Predictions. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 9684–9693, 2018.

[Kup05] Ivan Kupka. Continuous selections for Lipschitz multifunctions. *Acta Mathematica Universitatis Comenianae*, 74(1):133–141, 2005.

[Lac16] Hubert Lacoin. Mixing time and cutoff for the adjacent transposition shuffle and the simple exclusion. *The Annals of Probability*, 44(2):1426–1487, 2016.

[Lal96] Steven P Lalley. Cycle structure of riffle shuffles. *The Annals of Probability*, pages 49–73, 1996.

[Lal00] Steven P Lalley. On the rate of mixing for p-shuffles. *Annals of Applied Probability*, pages 1302–1321, 2000.

[Led99] Michel Ledoux. Concentration of measure and logarithmic sobolev inequalities. In *Seminaire de probabilites XXXIII*, pages 120–216. Springer, 1999.

[Led01] M. Ledoux. The concentration of measure phenomenon. In *Mathematical Surveys and Monographs*, volume 89. American Mathematical Society, Providence, RI, 2001.

[LFN18] Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[LMP15] Alessandra Lunardi, Michele Miranda, and Diego Pallara. Infinite dimensional analysis. In *19th Internet Seminar*, volume 2016, 2015.

[LS77] Robert Shevilevich Liptser and Al'bert Nikolaevich Shiriaev. *Statistics of random processes: General theory*, volume 394. Springer, 1977.

[LS84] Pierre-Louis Lions and Alain-Sol Sznitman. Stochastic differential equations with reflecting boundary conditions. *Communications on Pure and Applied Mathematics*, 37(4):511–537, 1984.

[LS93] László Lovász and Miklós Simonovits. Random walks in a convex body and an improved volume algorithm. *Random structures & algorithms*, 4(4):359–412, 1993.

[LS15] Eyal Lubetzky and Allan Sly. An exposition to information percolation for the ising model. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 24, pages 745–761, 2015.

[LS16] Eyal Lubetzky and Allan Sly. Information percolation for the Ising model: cutoff in three dimensions up to criticality. *J Am Math Soc*, pages 729–774, 2016.

[LS17] Eyal Lubetzky and Allan Sly. Universality of cutoff for the Ising model. *The Annals of Probability*, 45(6A):3664–3696, 2017.

[LST20] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Logsmooth gradient concentration and tighter runtimes for metropolized hamiltonian monte carlo. In *Conference on Learning Theory*, pages 2565–2597. PMLR, 2020.

[LV06] László Lovász and Santosh Vempala. Simulated annealing in convex bodies and an o*(n4) volume algorithm. *Journal of Computer and System Sciences*, 72(2):392–417, 2006.

[LV17] Yin Tat Lee and Santosh Srinivas Vempala. Eldan's stochastic localization and the kls hyperplane conjecture: an improved lower bound for expansion. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 998–1007. IEEE, 2017.

[LV18a] Yin Tat Lee and Santosh S Vempala. Stochastic localization+ stieltjes barrier= tight bound for log-sobolev. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1122–1129, 2018.

[LV18b] Thodoris Lykouris and Sergei Vassilvtiskii. Competitive caching with machine learned advice. In *International Conference on Machine Learning*, pages 3296–3305. PMLR, 2018.

[LWAT13] Minghong Lin, Adam Wierman, Lachlan LH Andrew, and Eno Thereska. Dynamic right-sizing for power-proportional data centers. *IEEE/ACM Transactions on Networking (TON)*, 21(5):1378–1391, 2013.

[Mas90] Pascal Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, pages 1269–1283, 1990.

[McK21] Benjamin McKenna. Complexity of bipartite spherical spin glasses. *arXiv preprint arXiv:2105.05043*, 2021.

[MHRB17] Zakaria Mhammedi, Andrew Hellicar, Ashfaqur Rahman, and James Bailey. Efficient orthogonal parametrisation of recurrent neural networks using householder reflections. In *International Conference on Machine Learning*, pages 2401–2409. PMLR, 2017.

[MKKY18] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *Proc. of International Conference on Learning Representation (ICLR)*, 2018.

[MM19] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 2019.

[MMS90] Mark S Manasse, Lyle A McGeoch, and Daniel D Sleator. Competitive algorithms for server problems. *Journal of Algorithms*, 11(2):208–230, 1990.

[MMS+18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *Proc. of International Conference on Learning Representation (ICLR)*, 2018.

[MMZ05] Marc Mézard, Thierry Mora, and Riccardo Zecchina. Clustering of solutions in the random satisfiability problem. *Physical Review Letters*, 94(19):197205, 2005.

[Mon21] Andrea Montanari. Optimization of the sherrington–kirkpatrick hamiltonian. *SIAM Journal on Computing*, (0):FOCS19–1, 2021.

[Mor09] Ben Morris. Improved mixing time bounds for the Thorp shuffle and L-reversal chain. *The Annals of Probability*, 37(2):453–477, 2009.

[Mor13] Ben Morris. Improved mixing time bounds for the thorp shuffle. *Combinatorics, Probability and Computing*, 22(1):118–132, 2013.

[MP12] Jason Miller and Yuval Peres. Uniformity of the uncovered set of random walk and cutoff for lamplighter chains. *The Annals of Probability*, 40(2):535–577, 2012.

[MPV87] Marc Mézard, Giorgio Parisi, and Miguel A. Virasoro. *Spin Glass Theory and Beyond*. World Scientific, 1987.

[MV21] Andrea Montanari and Ramji Venkataramanan. Estimation of low-rank matrices via approximate message passing. *The Annals of Statistics*, 49(1):321–345, 2021.

[NBA+18] Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. *arXiv preprint arXiv:1802.08760*, 2018.

[Nes21] Yurii Nesterov. Superfast second-order methods for unconstrained convex optimization. *Journal of Optimization Theory and Applications*, pages 1–30, 2021.

[NKB+20] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020.

[NM10] Hariharan Narayanan and Sanjoy Mitter. Sample complexity of testing the manifold hypothesis. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems-Volume 2*, pages 1786–1794, 2010.

[NSZ22] Danny Nam, Allan Sly, and Lingfu Zhang. Ising model on trees and factors of IID. *Communications in Mathematical Physics*, pages 1–38, 2022.

[Oks13] Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

[OSS07] Reinhold Oppermann, Manuel J Schmidt, and David Sherrington. Double Criticality of the Sherrington-Kirkpatrick Model at $T = 0$. *Physical Review Letters*, 98(12):127201, 2007.

[Pan] Dmitry Panchenko. Properties of the Parisi formula. https://drive.google.com/file/d/0B6JeBUquZ5BwRFpLVjdVd3IwV1E/view.

[Pan13a] Dmitry Panchenko. The Parisi ultrametricity conjecture. *Annals of Mathematics*, pages 383–393, 2013.

[Pan13b] Dmitry Panchenko. *The Sherrington-Kirkpatrick model*. Springer Science & Business Media, 2013.

[Pan14] Dmitry Panchenko. The Parisi formula for mixed p-spin models. *The Annals of Probability*, 42(3):946–958, 2014.

[Par79] Giorgio Parisi. Infinite number of order parameters for spin-glasses. *Physical Review Letters*, 43(23):1754, 1979.

[Par06] Giorgio Parisi. Computing the number of metastable states in infinite-range models. *arXiv preprint arXiv:cond-mat/0602349*, 2006.

[Pil14] Andrey Pilipenko. *An introduction to stochastic differential equations with reflection*, volume 1. Universitätsverlag Potsdam, 2014.

[Pit97] Jim Pitman. Probabilistic bounds on the coefficients of polynomials with only real zeros. *Journal of Combinatorial Theory, Series A*, 77(2):279–303, 1997.

[PY89] Krzysztof Przeslawski and David Yost. Continuity properties of selectors. *Michigan Math. J*, 36(1):13, 1989.

[PY95] Krzysztof Przeslawski and David Yost. Lipschitz retracts, selectors, and extensions. *Michigan Mathematical Journal*, 42(3):555–571, 1995.

[PZA+21] Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.

[RM14] Emile Richard and Andrea Montanari. A statistical model for tensor pca. In *Advances in Neural Information Processing Systems*, pages 2897–2905, 2014.

[Roc70] R Tyrrell Rockafellar. *Convex analysis*, volume 36. Princeton University Press, 1970.

[RS00] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[RTV86] Rammal Rammal, Gérard Toulouse, and Miguel Angel Virasoro. Ultrametricity for physicists. *Reviews of Modern Physics*, 58(3):765, 1986.

[Rue87] David Ruelle. A mathematical reformulation of Derrida's REM and GREM. *Communications in Mathematical Physics*, 108(2):225–239, 1987.

[RV17a] Mustazee Rahman and Bálint Virág. Local algorithms for independent sets are half-optimal. *The Annals of Probability*, 45(3):1543–1577, 2017.

[RV17b] Mustazee Rahman and Balint Virag. Local algorithms for independent sets are half-optimal. *The Annals of Probability*, 45(3):1543–1577, 2017.

[RW94] L. Chris G. Rogers and David Williams. *Diffusions, Markov processes and martingales: Volume 2, Itô calculus*, volume 2. Cambridge university press, 1994.

[RWK20] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8093–8104. PMLR, 2020.

[RY13] Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*, volume 293. Springer Science & Business Media, 2013.

[Sch05] Oded Schramm. Compositions of random transpositions. *Israel Journal of Mathematics*, 147(1):221–243, 2005.

[Sel20] Mark Sellke. Chasing Convex Bodies Optimally. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1509–1518. SIAM, 2020.

[Sel21a] Mark Sellke. Approximate Ground States of Hypercube Spin Glasses are Near Corners. *Comptes Rendus. Mathématique*, 359(9):1097–1105, 2021.

[Sel21b] Mark Sellke. Optimizing Mean Field Spin Glasses with External Field. *arXiv preprint arXiv:2105.03506*, 2021.

[Sel22] Mark Sellke. Cutoff for the Asymmetric Riffle Shuffle. *Annals of Probability*, 2022.

[Shv84] Pavel Shvartsman. Lipshitz selections of multivalued mappings and traces of the Zygmund class of functions to an arbitrary compact, dokl. acad. nauk sssr 276 (1984), 559-562. In *English transl. in Soviet Math. Dokl*, volume 29, pages 565–568, 1984.

[Shv02] Pavel Shvartsman. Lipschitz selections of set-valued mappings and Helly's theorem. *The Journal of Geometric Analysis*, 12(2):289–324, 2002.

[SK75] David Sherrington and Scott Kirkpatrick. Solvable model of a spin-glass. *Physical review letters*, 35(26):1792, 1975.

[SL19] Ruoqi Shen and Yin Tat Lee. The randomized midpoint method for log-concave sampling. *Advances in Neural Information Processing Systems*, 32, 2019.

[SO08] Manuel J Schmidt and Reinhold Oppermann. Method for replica symmetry breaking at and near T= 0 with application to the Sherrington-Kirkpatrick model. *Physical Review E*, 77(6):061104, 2008.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[ST85] Daniel D Sleator and Robert E Tarjan. Amortized efficiency of list update and paging rules. *Communications of the ACM*, 28(2):202–208, 1985.

[Sta01] Richard P Stanley. Generalized riffle shuffles and quasisymmetric functions. *Annals of Combinatorics*, 5(3-4):479–491, 2001.

[Ste40] Jacob Steiner. From the center of curvature of plane curves. *Journal of Pure and Applied Mathematics*, 21:33–63, 1840.

[Sub17] Eliran Subag. The complexity of spherical p-spin models—a second moment approach. *The Annals of Probability*, 45(5):3385–3450, 2017.

[Sub18] Eliran Subag. Free energy landscapes in spherical spin glasses. *arXiv preprint arXiv:1804.10576*, 2018.

[Sub21] Eliran Subag. Following the Ground States of Full-RSB Spherical Spin Glasses. *Communications on Pure and Applied Mathematics*, 74(5):1021–1044, 2021.

[SZ81] Haim Sompolinsky and Annette Zippelius. Dynamic theory of the spin-glass phase. *Physical Review Letters*, 47(5):359, 1981.

[SZ21] Eliran Subag and Ofer Zeitouni. Concentration of the complexity of spherical pure p-spin models at arbitrary energies. *Journal of Mathematical Physics*, 62(12):123301, 2021.

[Tal06a] Michel Talagrand. Free energy of the spherical mean field model. *Probability Theory and Related Fields*, 134:339–382, 03 2006.

[Tal06b] Michel Talagrand. Free energy of the spherical mean field model. *Probability theory and related fields*, 134(3):339–382, 2006.

[Tal06c] Michel Talagrand. Parisi measures. *Journal of Functional Analysis*, 231(2):269–286, 2006.

[Tal06d] Michel Talagrand. The Parisi formula. *Annals of Mathematics*, pages 221–263, 2006.

[Tal10] Michel Talagrand. *Mean Field Models for Spin Glasses: Volume I*. Springer-Verlag, Berlin, 2010.

[TDSL00] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[Tho73] Edward O Thorp. Nonrandom shuffling with applications to the game of Faro. *Journal of the American Statistical Association*, 68(344):842–847, 1973.

[Ton02] Fabio Lucio Toninelli. About the Almeida-Thouless transition line in the Sherrington-Kirkpatrick mean-field spin glass model. *EPL (Europhysics Letters)*, 60(5):764, 2002.

[VdV98] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 1998.

[vH14] Ramon van Handel. Probability in high dimension. Technical report, Princeton University, 2014.

[Wei22] Alexander S Wein. Optimal low-degree hardness of maximum independent set. *Mathematical Statistics and Learning*, 2022.

[Wil96] David Bruce Wilson. Generating random spanning trees more quickly than the cover time. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 296–303, 1996.

[WZ20] Alexander Wei and Fred Zhang. Optimal Robustness-Consistency Trade-offs for Learning-Augmented Online Algorithms. In *Advances in Neural Information Processing Systems*, volume 33, pages 8042–8053, 2020.

[XY20] Cihang Xie and Alan Yuille. Intriguing properties of adversarial training at scale. In *International Conference on Learning Representations*, 2020.

[YKB19] Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *International conference on machine learning*, pages 7085–7094. PMLR, 2019.

[YM17] Yuichi Yoshida and Takeru Miyato. Spectral norm regularization for improving the generalizability of deep learning. *arXiv preprint arXiv:1705.10941*, 2017.

[YSJ19] Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Small relu networks are powerful memorizers: a tight analysis of memorization capacity. In *Advances in Neural Information Processing Systems*, pages 15532–15543, 2019.

[Zei15] Ofer Zeitouni. Gaussian Fields Notes for Lectures. `https://www.wisdom.weizmann.ac.il/~zeitouni/notesGauss.pdf`, 2015.

[Zha09] Yufei Zhao. Biased riffle shuffles, quasisymmetric functions, and the RSK algorithm. *https://yufeizhao.com/research/shuffling.pdf*, 2009.

[ZK07] Lenka Zdeborova and Florent Krzakala. Phase transitions in the coloring of random graphs. *Physical Review E*, 76(3):031131, 2007.

Appendix A

State evolution: Proof of Proposition 3.6.1

In this and the following sections we prove Proposition 4.2.3. It will be more convenient to restate things directly in terms of the Gaussian tensors defining the Hamiltonian H_N. Throughout, we denote by $\boldsymbol{W}^{(k)} \in (\mathbb{R}^N)^{\otimes k}$, $k \geq 2$ a sequence of standard Gaussian tensors defined as follows.

Let $\boldsymbol{W}^{(k)} \in (\mathbb{R}^N)^{\otimes k}$, $k \geq 2$, be a standard symmetric Gaussian tensor of order k with entries $\boldsymbol{W}^{(k)} \equiv (W^{(k)}_{i_1,\cdots,i_k})_{1\leq i_1,\cdots,i_k\leq N}$. Namely, if $\{G^{(k)}_{i_1,\ldots,i_k} : k \geq 2, 1 \leq i_1, \cdots, i_k \leq N\}$ is a collection of i.i.d. standard normal $\mathsf{N}(0,1)$ random variables, we set $\boldsymbol{W}^{(k)} \equiv N^{-(k-1)/2} \sum_{\pi\in S_k} \boldsymbol{G}^{(k)}_\pi$ where the sum is over the group of permutations of k objects, and $\boldsymbol{G}^{(k)}_\pi$ is obtained by permuting the indices of $\boldsymbol{G}^{(k)}$ according to π.

We write $\boldsymbol{A}^{(k)} = c_k \boldsymbol{W}^{(k)}$ for the rescaled tensors, and $\xi(t) = \sum_{k\geq 2} c_k^2 t^k$. Recall the notation $\boldsymbol{A}^{(p)}\{\boldsymbol{u}\} \in \mathbb{R}^N$, for a symmetric tensor $\boldsymbol{A}^{(p)} \in (\mathbb{R}^N)^{\otimes p}$:

$$\boldsymbol{A}^{(p)}\{\boldsymbol{u}\}_i = \frac{1}{(p-1)!} \sum_{1\leq i_1,\cdots,i_{p-1}\leq N} A^{(p)}_{i,i_1,\cdots,i_{p-1}} u_{i_1}\cdots u_{i_{p-1}}. \tag{A.0.1}$$

Analogously, if $\boldsymbol{T} \in (\mathbb{R}^N)^{\otimes(p-1)}$, $\boldsymbol{A}^{(p)}\{\boldsymbol{T}\} \in \mathbb{R}^N$ is the vector with components

$$\boldsymbol{A}^{(p)}\{\boldsymbol{T}\}_i = \frac{1}{(p-1)!} \sum_{1\leq i_1,\cdots,i_{p-1}\leq N} A^{(p)}_{i,i_1,\cdots,i_{p-1}} T_{i_1\ldots i_{p-1}}. \tag{A.0.2}$$

We will use the notation $\langle \boldsymbol{v}\rangle_N = N^{-1}\sum_{i\leq N} v_i$ and $\langle \boldsymbol{u},\boldsymbol{v}\rangle_N = N^{-1}\sum_{i\leq N} u_i v_i$ when $\boldsymbol{u},\boldsymbol{v} \in \mathbb{R}^N$ are vectors. The corresponding norm is $\|\boldsymbol{u}\|_N = \langle \boldsymbol{u},\boldsymbol{u}\rangle_N^{1/2}$. We will write $a_N \overset{p}{\simeq} b_N$ to mean that $a_N - b_N$ converges in probability to 0. Analogously, for two vectors $\boldsymbol{u}_N, \boldsymbol{v}_N$, we write $\boldsymbol{u}_N \overset{p}{\simeq} \boldsymbol{v}_N$

when $\|\boldsymbol{u}_N - \boldsymbol{v}_N\|_N$ converges in probability to 0. When $f : \mathbb{R}^{k+1} \to \mathbb{R}$ is a function of $k+1$ variables, and $\boldsymbol{v}^0, \boldsymbol{v}^1, \ldots, \boldsymbol{v}^k \in \mathbb{R}^N$ are $k+1$, we define $f(\boldsymbol{v}^0, \boldsymbol{v}^1, \ldots, \boldsymbol{v}^k) \in \mathbb{R}^N$ component-wise via

$$f(\boldsymbol{v}^0, \boldsymbol{v}^1, \ldots, \boldsymbol{v}^k)_i = f(v_i^0, \ldots, v_i^k). \tag{A.0.3}$$

Finally, for a sequence of vectors $\boldsymbol{x}^0, \boldsymbol{x}^1, \ldots$, we write $\boldsymbol{x}^{\leq t} = (\boldsymbol{x}^0, \boldsymbol{x}^1, \ldots, \boldsymbol{x}^t)$.

To deduce the state evolution result for mixed tensors, we analyze a slightly more general iteration where each homogenous p-tensor is tracked separately, while restricting ourselves to the case where the mixture ξ has finitely many components: $c_k = 0$ for all $k \geq D+1$ for some fixed $D \geq 2$. We then proceed by an approximation argument to extend the convergence to the general case $D = \infty$.

We begin by introducing the Gaussian process that captures the asymptotic behavior of AMP. For each $t \in \mathbb{N}$, let $f_t : \mathbb{R}^{t+1} \to \mathbb{R}$ be a Lipschitz function. Let $(U^{k,0})_{2 \leq k \leq D}$ a collection of random variables with bounded second moment, and $(U^{k,t})_{k \leq D, 1 \leq t \leq T}$ a centered Gaussian process, independent of $(U^{k,0})_{2 \leq k \leq D}$, with covariance defined by:

1. $U^{k,t}, U^{k',s}$ are independent whenever $k \neq k'$.
2. For each k, the covariance of $(U^{k,t})_{t \leq T}$ is defined recursively via

$$\mathbb{E}[U^{k,t+1} U^{k,s+1}] = k c_k^2 \, \mathbb{E}\left\{ f_t\left(X^0, \ldots, X^t\right) f_s\left(X^0, \ldots, X^s\right) \right\}^{k-1}, \tag{A.0.4}$$

$$X^t \equiv \sum_{k=2}^{D} U^{k,t}. \tag{A.0.5}$$

We are now in position to define the AMP algorithm. For each iteration t, the state of the algorithm is given by vectors $\boldsymbol{x}^t \in \mathbb{R}^N$, and $\boldsymbol{z}^{k,t} \in \mathbb{R}^N$, with $k \in \{2, \ldots, D\}$. (In the following we will often omit mentioning explicitly that k starts from 2 and simply write $k \leq D$.) We define the AMP mapping via

$$\mathsf{AMP}_t\left(\boldsymbol{x}^0, \ldots, \boldsymbol{x}^t\right)_k := \boldsymbol{A}^{(k)}\{f_t(\boldsymbol{x}^0, \ldots, \boldsymbol{x}^t)\} - \sum_{s \leq t} d_{t,s,k} f_{s-1}(\boldsymbol{x}^0, \ldots, \boldsymbol{x}^{s-1}), \tag{A.0.6}$$

$$\begin{aligned} d_{t,s,k} := c_k^2 \cdot k(k-1) \, \mathbb{E}\left\{ f_t\left(X^0, \ldots, X^t\right) f_{s-1}\left(X^0, \ldots, X^{s-1}\right) \right\}^{k-2} \\ \cdot \, \mathbb{E}\left\{ \frac{\partial f_t}{\partial x^s}(X^0, X^1, \ldots, X^t) \right\}. \end{aligned} \tag{A.0.7}$$

The *tensor AMP iteration* then reads

$$\boldsymbol{x}^t = \sum_{k=2}^{D} \boldsymbol{z}^{k,t}, \qquad \boldsymbol{z}^{k,t+1} = \mathsf{AMP}_t\left(\boldsymbol{x}^0, \ldots, \boldsymbol{x}^t\right)_k. \tag{A.0.8}$$

Theorem 39 (State Evolution for AMP)**.** *Let $\{\boldsymbol{W}^{(k)}\}_{k \geq 2}$ be independent standard Gaussian tensors*

with $\boldsymbol{W}^{(k)} \in (\mathbb{R}^N)^{\otimes k}$*, and define* $\boldsymbol{A}^{(k)} = c_k \boldsymbol{W}^{(k)}$*,* $\xi(t) = \sum_{k=2}^{D} c_k^2 t^k$*. Let* $f_0, f_1, \dots,$ *be a sequence of Lipschitz functions* $f_k : \mathbb{R}^{k+1} \to \mathbb{R}$*. Let* $\boldsymbol{z}^{2,0}, \cdots \boldsymbol{z}^{D,0} \in \mathbb{R}^N$ *be deterministic vectors and* $\boldsymbol{x}^0 = \sum_{k=2}^{D} \boldsymbol{z}^{k,0}$*. Assume that, the empirical distribution of the vectors* $(z_i^{2,0}, \cdots z_i^{D,0})$*,* $i \le N$ *converges in* W_2 *distance to the law of the vector* $(U^{k,0})_{2\le k\le D}$*.*

Let $\boldsymbol{x}^t, \boldsymbol{z}^{k,t}$*,* $t \ge 1$ *be given by the* tensor AMP *iteration. Then, for any* $T \ge 1$ *and for any pseudo-Lipschitz function* $\psi : \mathbb{R}^{D\times T} \to \mathbb{R}$*, we have*

$$\operatorname*{p-lim}_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \psi\big((z_i^{k,t})_{k\le D, t\le T}\big) = \mathbb{E}\left\{\psi\big((U^{k,t})_{k\le D, t\le T}\big)\right\}. \tag{A.0.9}$$

where $(U^{k,t})_{k\le D,t\le T}$ *is a centered Gaussian process, independent of* $(U^{k,0})_{2\le k\le D}$*, with covariance defined above.*

In the above proposition, W_2 refers to the Wasserstein, or optimal transport, distance between probability measures on $\mathbb{R}^D$ with quadratic cost $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^2$.

Proposition 3.6.1 in the special case $c_k = 0$ for all $k \ge D+1$ follows immediately from this theorem by considering $\psi((z^{k,t})_{k\le D,t\le T})$ only a function of $(\sum_{k\le D} z^{k,t})_{t\le T}$. We extend Proposition 3.6.1 to the general case $D = \infty$ in Section A.8.

A.1 Further definitions

We define the notations

$$\begin{aligned}
\boldsymbol{X}_t &= [\boldsymbol{x}_0 | \boldsymbol{x}_1 | \ \cdots \ | \boldsymbol{x}_t]\,, \\
\boldsymbol{Z}_{p,t}^k &= [\boldsymbol{z}_{p,0}^{\otimes k} | \boldsymbol{z}_{p,1}^{\otimes k} | \cdots | \boldsymbol{z}_{p,t}^{\otimes k}]\,,
\end{aligned}$$

where we replaced superscripts by subscripts for notational convenience. Given a $N \times (t+1)$ matrix, such as $\boldsymbol{X}_t$, and a tensor $\boldsymbol{A}^{(p)} \in (\mathbb{R}^N)^{\otimes p}$, we write $\boldsymbol{A}^{(p)}\{\boldsymbol{X}_t\}$ for the $N \times (t+1)$ matrix with columns $\boldsymbol{A}^{(p)}\{\boldsymbol{x}_0\}$, …, $\boldsymbol{A}^{(p)}\{\boldsymbol{x}_t\}$:

$$\boldsymbol{A}^{(p)}\{\boldsymbol{X}_t\} = \left[\boldsymbol{A}^{(p)}\{\boldsymbol{x}_0\} \middle| \boldsymbol{A}^{(p)}\{\boldsymbol{x}_1\}\right| \ \cdots \ \left|\boldsymbol{A}^{(p)}\{\boldsymbol{x}_t\}\right].$$

When $k = 1$ we omit k, e.g. $\boldsymbol{Z}_{p,t}^1 = \boldsymbol{Z}_{p,t}$. We will write $f_t(\boldsymbol{X}_t) = f_t(\mathbf{x}^0, \dots, \mathbf{x}^t)$, and we also set

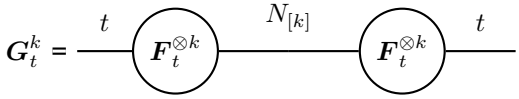
$$\boldsymbol{y}_{p,t+1}(\boldsymbol{Z}_{p,t}) = \boldsymbol{A}_p\left\{f_t(\boldsymbol{Z}_{p,t})\right\} = \boldsymbol{z}^{p,t+1} + \sum_{s\le t} d_{t,s,p} f_{s-1}(\boldsymbol{x}^0, \dots, \boldsymbol{x}^{s-1})\,, \tag{A.1.1}$$

$$\boldsymbol{Y}_{p,t} = [\boldsymbol{y}_{p,1} | \ \cdots \ | \boldsymbol{y}_{p,t}]\,, \qquad \boldsymbol{y}_t(\boldsymbol{Z}_{p,t}) = \sum_p \boldsymbol{y}_{p,t}(\boldsymbol{Z}_{p,t})\,. \tag{A.1.2}$$

For any positive integer k and $p \times T$ matrix $\mathtt{max}$ of length n vectors we define $\boldsymbol{F}_t^k(\mathtt{max})$ to be the length $t+1$ vector of k-tensors

$$\boldsymbol{F}_t^k(\mathtt{max}) = [f_0(\mathtt{max})^{\otimes k} | f_1(\mathtt{max})^{\otimes k} | \ \cdots \ | f_t(\mathtt{max})^{\otimes k}]. \tag{A.1.3}$$

We also define an associated $(t+1) \times (t+1)$ Gram matrix $\boldsymbol{G}_t^k = \boldsymbol{G}_t^k(\mathbf{M})$ via $(\boldsymbol{G}_t^k(\mathbf{M}))_{i,j} = \langle f_i(\mathbf{M}), f_j(\mathbf{M}) \rangle_N^k$. The matrix $\boldsymbol{G}_t^k$ can be represented by the following tensor network diagram:



We recall that in tensor networks, tensors correspond to vertices, and edges joining them to indices contracted between tensors. We use the convention of labeling vertices by the corresponding tensors, and edges by the dimension of the corresponding index. Since we often have indices with dimension N, we label the edges by $N_1, N_2, \ldots$ and so on. When two tensors are contracted along multiple indices of the same dimension (say N), we draw a single line between them labelled N_S where S is the set of contracted indices. For example, the middle edge in the above figure represents k edges with labels $N_1, \cdots, N_k$.

Finally, we let $\mathcal{F}_t$ denote the σ-algebra generated by all iterates up to time t:

$$\mathcal{F}_t = \sigma\big(\{\boldsymbol{z}_{p,s}\}_{p \le D, s \le t}\big) = \sigma(\{\boldsymbol{z}_{p,s}, \boldsymbol{x}_s, \mathbf{f}_s\}_{p \le D, s \le t}) \,. \tag{A.1.4}$$

A.2 Preliminary lemmas

Lemma A.2.1. *For any deterministic $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^N$ and standard Gaussian symmetric p-tensor $\boldsymbol{W}^{(p)} \in (\mathbb{R}^N)^{\otimes p}$ we have:*

1. *Letting $g_0 \sim \mathsf{N}(0,1)$ independently of $\mathbf{g} \sim \mathsf{N}(0, \boldsymbol{I}_N)$, we have*

$$\boldsymbol{W}^{(p)}\{\boldsymbol{u}\} \stackrel{\mathrm{d}}{=} \sqrt{p}\|\boldsymbol{u}\|_N^{p-1}\,\mathbf{g} + \sqrt{p(p-1)}\|\boldsymbol{u}\|_N^{p-2}\frac{\boldsymbol{u}}{\sqrt{N}}\,g_0 \,. \tag{A.2.1}$$

2. *Letting $g_0, g_1 \sim \mathsf{N}(0,1)$ independent, we have*

$$\sqrt{N}\langle \boldsymbol{v}, \boldsymbol{W}^{(p)}\{\boldsymbol{u}\}\rangle_N \stackrel{\mathrm{d}}{=} \sqrt{p}\|\boldsymbol{u}\|_N^{p-1}\|\boldsymbol{v}\|_N\,g_1 + \sqrt{p(p-1)}\|u\|_N^{p-2}\langle \boldsymbol{u}, \boldsymbol{v}\rangle_N\,g_0 \,. \tag{A.2.2}$$

3. $\langle \boldsymbol{W}^{(p)}\{\boldsymbol{u}\}, \boldsymbol{W}^{(p)}\{\boldsymbol{v}\}\rangle_N \stackrel{p}{\simeq} p\langle \boldsymbol{u}, \boldsymbol{v}\rangle_N^{p-1}$.

4. *For a deterministic symmetric tensor $\boldsymbol{T} \in (\mathbb{R}^N)^{\otimes p-1}$, the vector $\boldsymbol{W}^{(p)}\{\boldsymbol{T}\}$ is Gaussian, with*

zero mean and covariance

$$\mathbb{E}\{\boldsymbol{W}^{(p)}\{\boldsymbol{T}\}_i\boldsymbol{W}^{(p)}\{\boldsymbol{T}\}_j\} = \frac{p}{N^{p-1}}\,\|\boldsymbol{T}\|_F^2 + \frac{p(p-1)}{N^{p-1}}\sum_{i_1,\dots,i_{p-2}=1}^{N} T_{i,i_1,\dots,i_{p-1}}T_{j,i_1,\dots,i_{p-1}}\,. \tag{A.2.3}$$

5. *Let* $\mathbf{P}\in\mathbb{R}^{N\times N}$ *be the orthogonal projection onto a* d*-dimensional subspace* $S\subseteq\mathbb{R}^N$*.* $\|\mathbf{P}\boldsymbol{W}^{(p)}\{\boldsymbol{u}\}-\boldsymbol{W}^{(p)}\{\boldsymbol{u}\}\|_2/\|\boldsymbol{W}^{(p)}\{\boldsymbol{u}\}\|_2 \overset{p}{\to} 0$*.*

6. *Recall that the operator (injective) norm of a tensor is given by*

$$\|\boldsymbol{W}^{(p)}\|_{\rm op} \equiv \max_{\|\boldsymbol{u}_1\|\le 1,\dots,\|\boldsymbol{u}_p\|\le 1}\langle \boldsymbol{W}^{(p)}, \boldsymbol{u}_1\otimes\cdots\otimes\boldsymbol{u}_p\rangle,$$

or, equivalently for a symmetric tensor, by $\|\boldsymbol{W}^{(p)}\|_{\rm op}\equiv\max_{|\boldsymbol{u}\|\le 1}\langle\boldsymbol{W}^{(p)},\boldsymbol{u}^{\otimes p}\rangle$*. If* $\xi(t)<\infty$ *for some* $t>1$*, then there exists a constant* $C=C(\xi)$ *such that, with probability at least* $1-2e^{-N}$*,*

$$\|\boldsymbol{A}\|_{\rm op}\equiv\sum_{k=2}^{\infty}\frac{N^{k/2}}{k!}\|\boldsymbol{A}^{(k)}\|_{\rm op}=\sum_{k=2}^{\infty}\frac{c_kN^{k/2}}{k!}\|\boldsymbol{W}^{(k)}\|_{\rm op}\le CN\,. \tag{A.2.4}$$

Proof. All of these statements are the elementary Gaussian calculations. The only exception is the upper bound (A.2.4), which follows from the concentration bound

$$\mathbb{P}\Big(N^{(k-2)/2}\cdot\|\boldsymbol{W}^{(k)}\|_{\rm op}\ge k!\sqrt{\log k}+\frac{k!}{\sqrt{k}}s\Big)\le e^{-Ns^2/2k}\quad\forall s\ge 0.$$

The above is a restatement of [RM14, Lemma 2]. We conclude by using the fact $|c_k|\le c_*\alpha^k$ for some $\alpha<1$ and letting $s=k$. □

We next develop a formula for the conditional expectation of a Gaussian tensor $\boldsymbol{A}^{(p)}$ given a collection of linear observations. We set $\boldsymbol{D}$ to be the $t\times t\times t$ tensor with entries $D_{ijk}=1$ if $i=j=k$ and $D_{ijk}=0$ otherwise.

Lemma A.2.2. *Let* $\mathbb{E}\{\boldsymbol{A}^{(p)}|\mathcal{F}_t\}$ *be the conditional expectation of* $\boldsymbol{A}^{(p)}$ *given the* σ*-algebra* $\mathcal{F}_t=\sigma(\{\boldsymbol{z}_{p,s},\boldsymbol{x}_s,\mathbf{f}_s\}_{p\le D,s\le t})$ *generated by observations up to time* t*. Equivalently* $\mathbb{E}\{\boldsymbol{A}^{(p)}|\mathcal{F}_t\}$ *is the conditional expectation of* $\boldsymbol{A}^{(p)}$ *given the* t *linear (in* $\boldsymbol{A}^{(p)}$*) observations*

$$\boldsymbol{A}^{(p)}\{\mathbf{f}_s\}=\boldsymbol{y}_{p,s+1}\quad\text{for } s\in\{0,\dots,t-1\}. \tag{A.2.5}$$

Then we have for $i_1, i_2, \ldots, i_p \le n$,

$$\mathbb{E}[\boldsymbol{A}^{(p)}|\mathcal{F}_t]_{i_1,i_2,\ldots,i_p} = \frac{1}{p}\sum_{j=1}^{p}\sum_{0\le r,s\le t-1}(\hat{\boldsymbol{Z}}_{p,t})_{i_j,s}\cdot(\boldsymbol{G}^{-1}_{p-1,t-1})_{s,r}\cdot(\mathbf{f}_{r,i_1}\cdots\mathbf{f}_{r,i_{j-1}}\mathbf{f}_{r,i_{j+1}}\cdots\mathbf{f}_{r,i_p})\,. \tag{A.2.6}$$

Here, the matrix $\hat{\boldsymbol{Z}}_{p,t}\in\mathbb{R}^{N\times t}$ *is defined as the solution of a system of linear equations as follows. Define the linear operator* $\mathcal{T}_{p,t}:\mathbb{R}^{N\times t}\to\mathbb{R}^{N\times t}$ *by letting, for* $i\le N$, $0\le s\le t-1$*:*

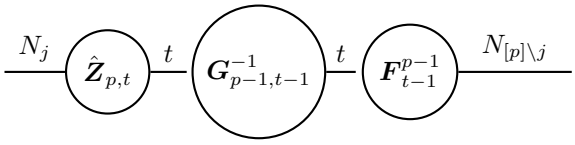
$$[\mathcal{T}_{p,t}(\boldsymbol{Z})]_{i,s} = \sum_{j\le N}\sum_{0\le r,r'\le t-1}(\mathbf{f}_{r'})_i(\mathbf{f}_s)_j(\boldsymbol{G}^{-1}_{p-1,t-1})_{r',r}(\boldsymbol{G}_{p-2,t-1})_{r',s}(\boldsymbol{Z})_{j,r}\,, \tag{A.2.7}$$

Then $\hat{\boldsymbol{Z}}_{p,t}$ *is the unique solution of the following linear equation (with* $\boldsymbol{Y}_{p,t}$ *defined as per Eq.* (A.1.1)*)*

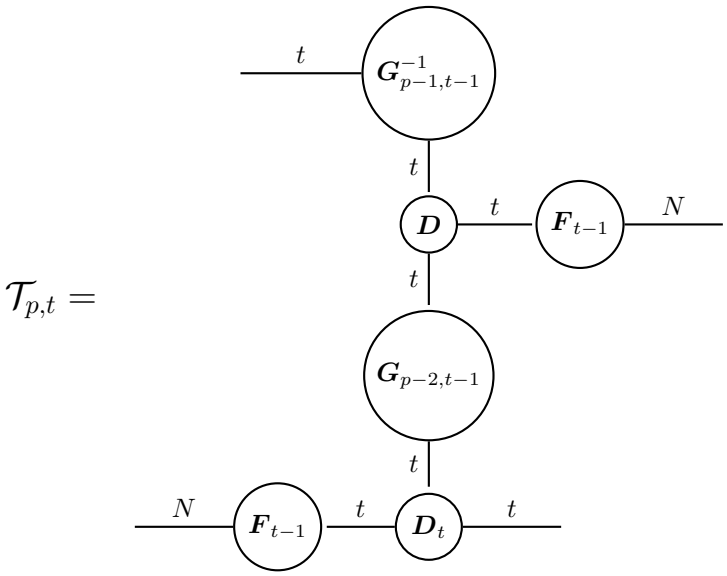
$$\hat{\boldsymbol{Z}}_{p,t} + (p-1)\mathcal{T}_{p,t}(\hat{\boldsymbol{Z}}_{p,t}) = \boldsymbol{Y}_{p,t}. \tag{A.2.8}$$

(Here, $\hat{\boldsymbol{Z}}_{p,t} = [\hat{\boldsymbol{z}}_{p,0},\cdots,\hat{\boldsymbol{z}}_{p,t-1}]$ *and* $\boldsymbol{Y}_{p,t} = [\hat{\boldsymbol{y}}_{p,1},\cdots,\hat{\boldsymbol{y}}_{p,t}]$ *have dimensions* $N\times t$*.)*

The above formulas for $\mathbb{E}\{\boldsymbol{A}^{(p)}|\mathcal{F}_t\}$ and $\mathcal{T}_{p,t}$ are somewhat difficult to parse. It is therefore useful to draw the associated tensor networks

$\mathbb{E}[\boldsymbol{A}^{(p)}|\mathcal{F}_t] = \frac{1}{p}\sum_j$ 

The operator $\mathcal{T}_{p,t}$ is represented by the following diagram, with input on the left and output on the right.



Proof of Lemma A.2.2. Let $\mathcal{V}_{p,t}$ be the affine space of symmetric tensors satisfying the constraint (A.2.5). The conditional expectation $\mathbb{E}[\boldsymbol{A}^{(p)}|\mathcal{F}_t]$ is the tensor with minimum Frobenius norm in the affine space $\mathcal{V}_{p,t}$. By Lagrange multipliers, there exist vectors $\boldsymbol{m}_1, \dots, \boldsymbol{m}_t \in \mathbb{R}^N$ such that $\mathbb{E}[\boldsymbol{A}^{(p)}|\mathcal{F}_t] = \hat{\boldsymbol{A}}^{(p)}$ takes the form

$$\hat{\boldsymbol{A}}_t^{(p)} := \sum_{s=0}^{t-1} \sum_{j=1}^{p} \underbrace{\mathbf{f}_s \otimes \cdots \otimes \mathbf{f}_s}_{j-1 \text{ times}} \otimes \boldsymbol{m}_s \otimes \underbrace{\mathbf{f}_s \otimes \cdots \otimes \mathbf{f}_s}_{p-j \text{ times}} . \tag{A.2.9}$$

Further, again by duality, if a tensor $\hat{\boldsymbol{A}}^{(p)}$ of this form (i.e., a choice of vectors $\boldsymbol{m}_1, \dots, \boldsymbol{m}_t$) satisfies the constraints $\hat{\boldsymbol{A}}^{(p)}\{\mathbf{f}_s\} = \boldsymbol{y}_{p,s+1}$ for $s < t$, then such a tensor is unique, and corresponds to $\mathbb{E}[\boldsymbol{A}^{(p)}|\mathcal{F}_t]$. Without loss of generality, we write

$$\boldsymbol{m}_r = \sum_{s=0}^{t-1} (\boldsymbol{G}^{-1}_{p-1,t-1})_{r,s} \hat{\boldsymbol{z}}_s \,, \quad \hat{\boldsymbol{Z}}_{p,t} = [\hat{\boldsymbol{z}}_1| \ \cdots \ |\hat{\boldsymbol{z}}_t] \,. \tag{A.2.10}$$

By direct calculation we obtain

$$\hat{\boldsymbol{A}}_t^{(p)}\{\mathbf{f}_s\} = \sum_{r=0}^{t-1} (\boldsymbol{G}_{p-1,t-1})_{s,r} \boldsymbol{m}_r + (p-1) \sum_{r=0}^{t-1} (\boldsymbol{G}_{p-2,t-1})_{s,r} \langle \mathbf{f}_s, \boldsymbol{m}_r \rangle \mathbf{f}_r \tag{A.2.11}$$

$$= \hat{\boldsymbol{z}}_s + (p-1) \sum_{r=0}^{t-1} (\boldsymbol{G}_{p-2,t-1})_{s,r} \langle \mathbf{f}_s, \boldsymbol{m}_r \rangle \mathbf{f}_r \,. \tag{A.2.12}$$

We next stack these vectors as columns of an $N \times t$ matrix. The first term obviously yields $\hat{\boldsymbol{Z}}_{p,t}$.

We claim that the second term coincides with $(p-1)\mathcal{T}_{p,t}(\hat{\boldsymbol{Z}}_{p,t})$ so that overall we get

$$\big[\hat{\boldsymbol{A}}_t^{(p)}\{\mathbf{f}_0\},\cdots,\hat{\boldsymbol{A}}_t^{(p)}\{\mathbf{f}_{t-1}\}\big] = \hat{\boldsymbol{Z}}_{p,t} + (p-1)\mathcal{T}_{p,t}(\hat{\boldsymbol{Z}}_{p,t})\,. \tag{A.2.13}$$

This in turns implies that the equation determining $\hat{\boldsymbol{Z}}_{p,t}$ takes the form (A.2.8). The desired claim is simply obtained by rearranging the order of sums in Eq. (A.2.12). □

A.3 Long AMP

As an intermediate step towards proving Theorem 39, we introduce a new iteration that we call Long AMP (LAMP), following [BMN19]. This iteration is less compact but simpler to analyze. For each $p \le D$, let $\mathcal{S}_{p,t} \subseteq (\mathbb{R}^N)^{\otimes p}$ be the linear subspace of tensors $\boldsymbol{T}$ that are symmetric and such that $\boldsymbol{T}\{\mathbf{f}_s\} = 0$ for all $s < t$. We denote by $\mathcal{P}_t^{\perp}(\boldsymbol{A}^{(p)})$ be the projection of $\boldsymbol{A}^{(p)}$ onto $\mathcal{S}_{p,t}$. We then define the LAMP mapping

$$\mathsf{LAMP}_t\left(\vec{v}^{\le t}\right)_p := \mathcal{P}_t^{\perp}(\boldsymbol{A}^{(p)})\{f_t(\vec{v}^0,\dots,\vec{v}^t)\} + \sum_{0\le s\le t} h_{t,s-1,p}\boldsymbol{q}^{p,s}, \tag{A.3.1}$$

$$h_{t,s,p} := \sum_{0\le r\le t-1} \big[\boldsymbol{G}_{p-1,t-1}^{-1}\big]_{s,r}\big[\boldsymbol{G}_{p-1,t}\big]_{r,t}, \quad h_{t,-1,p} = 0. \tag{A.3.2}$$

Here we use the same notations $\mathbf{f}_t = f_t(\boldsymbol{V}_t)$ and $\boldsymbol{G}_{k,t} = \boldsymbol{G}_{k,t}(\boldsymbol{V}_t) = (\langle \mathbf{f}_s, \mathbf{f}_r\rangle^k)_{s,r\le t}$ that we introduced for the case of AMP, however, these quantities are now different: they are computed using the vectors $\boldsymbol{v}^0,\dots,\boldsymbol{v}^t$.

$$\vec{v}^t = \sum_{p=2}^{D} \boldsymbol{q}^{p,t}\,, \qquad \boldsymbol{q}^{p,t+1} = \mathsf{LAMP}_t\left(\vec{v}^{\le t}\right)_p\,. \tag{A.3.3}$$

Our proof strategy will be similar to the one of [BMN19], and proceed along the following steps:

1. Prove state evolution for LAMP, under a non-degeneracy assumption.
2. Deduce state evolution for AMP, under the previous non-degeneracy assumption.
3. Deduce general state evolution for AMP, by perturbing the functions f_t slightly to give a non-degenerate instance.

We will use notations analogous to the ones introduced for AMP. In particular:

$$\boldsymbol{V}_t = [\vec{v}_1|\vec{v}_2|\dots|\vec{v}_t] \tag{A.3.4}$$

$$\mathbf{Q}_{p,t} = [\boldsymbol{q}_{p,1}^{\otimes p}|\boldsymbol{q}_{p,2}^{\otimes p}|\dots|\boldsymbol{q}_{p,t}^{\otimes p}]. \tag{A.3.5}$$

A.4 State Evolution for LAMP

Theorem 40. *Under the assumptions of Theorem 39, let $\boldsymbol{q}^{2,0}, \cdots \boldsymbol{q}^{D,0} \in \mathbb{R}^N$ be deterministic vectors and $\boldsymbol{v}^0 = \sum_{p=2}^{D} \mathbf{q}^{p,0}$. Assume that, the empirical distribution of the vectors $(q_i^{2,0}, \cdots, q_i^{D,0})$, $i \le N$ converges in W_2 distance to the law of the vector $(U^{p,0})_{2\le p\le D}$.*

Further assume that there exist a constant $C < \infty$ such that, for all $t \le T$,

(i) The matrices $\boldsymbol{G}_{p,t} = \boldsymbol{G}_{p,t}(\boldsymbol{V})$ are well-conditioned, i.e., $C^{-1} \le \sigma_{\min}(\boldsymbol{G}_{p,t}) \le \sigma_{\max}(\boldsymbol{G}_{p,t}) \le C$ for all $p \le D$, $t \le T$.

(ii) Let the linear operator $\mathcal{T}_{p,t} : \mathbb{R}^{N\times t} \to \mathbb{R}^{N\times t}$ be defined as per Eq. (A.2.7), with $\boldsymbol{G}_{p,t} = \boldsymbol{G}_{p,t}(\boldsymbol{V})$, and $\mathbf{f}_t = f_t(\boldsymbol{V})$, and define $\mathcal{L}_{p,t} = \mathbf{1} + (p-1)\mathcal{T}_{p,t}$. Then $C^{-1} \le \sigma_{\min}(\mathcal{L}_{p,t}) \le \sigma_{\max}(\mathcal{L}_{p,t}) \le C$.

Then the following statements hold for any $t \le T$ and sufficiently large N:

(a) Correct conditional law:

$$\boldsymbol{q}^{p,t+1}|_{\mathcal{F}_t} \stackrel{\mathrm{d}}{=} \mathbb{E}[\boldsymbol{q}^{p,t+1}|\mathcal{F}_t] + \mathcal{P}_t^{\perp}(\tilde{\boldsymbol{A}}^{(p)})\{f_t(\boldsymbol{V}_t)\}\,. \tag{A.4.1}$$

where $\tilde{\boldsymbol{A}}^{(p)}$ is a symmetric tensor distributed identically to $\boldsymbol{A}^{(p)}$ and independent of everything else, and $\mathcal{P}_t^{\perp}$ is the projection onto the subspace $\mathcal{S}_{p,t}$ defined in Section A.3. Further

$$\mathbb{E}[\boldsymbol{q}^{p,t+1}|\mathcal{F}_t] = \sum_{0\le s\le t} h_{t,s-1,p}\boldsymbol{q}^{p,s}\,. \tag{A.4.2}$$

Moreover, the vectors $(\boldsymbol{q}^{p,t+1})_{p\le D}$ are conditionally independent given $\mathcal{F}_t$.

(b) Approximate isometry: we have

$$\langle \boldsymbol{q}^{p,r+1}, \boldsymbol{q}^{p,s+1}\rangle_N \stackrel{p}{\simeq} pc_p^2\langle f_r(\boldsymbol{V}_r), f_s(\boldsymbol{V}_s)\rangle_N^{p-1}\,, \tag{A.4.3}$$

$$\langle \vec{v}^{r+1}, \vec{v}^{s+1}\rangle_N \stackrel{p}{\simeq} \xi'\big(\langle f_r(\boldsymbol{V}_r), f_s(\boldsymbol{V}_s)\rangle_N\big). \tag{A.4.4}$$

Moreover, both sides converge in probability to constants as $N \to \infty$, and for $p \ne p'$,

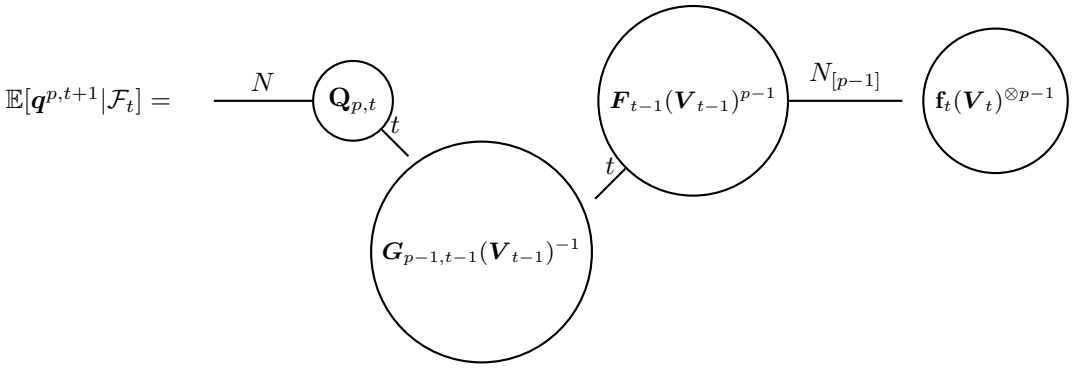
$$\langle \boldsymbol{q}^{p,r+1}, \boldsymbol{q}^{p',s+1}\rangle_N \stackrel{p}{\simeq} 0. \tag{A.4.5}$$

(c) State evolution: for any pseudo-Lipschitz function $\psi : \mathbb{R}^{D\times T} \to \mathbb{R}$, we have

$$\operatorname*{p-lim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \psi\big((q_i^{p,t})_{p\le D, t\le T}\big) = \mathbb{E}\left\{\psi\big((U^{p,t})_{p\le D,t\le T}\big)\right\}. \tag{A.4.6}$$

where $(U^{p,t})_{p\le D,1\le t\le T}$ is a centered Gaussian process, independent of $(U^{p,0})_{2\le p\le D}$, as defined in the statement of Theorem 39.

Note that the conditional expectation, as given by Eqs. (A.3.2), (A.4.2) can be represented by the following tensor network:



In the next section, we will prove these statements by induction on t. The crucial point we exploit is the representation (a).

As a preliminary remark, we emphasize that the iteration number t is bounded as $N \to \infty$, and therefore all numerical quantities not depending on N (but possibly on t) will be treated as constants. Further we will refer to the condition $C_T^{-1} \le \sigma_{\min}(\boldsymbol{G}_{k.t}) \le \sigma_{\max}(\boldsymbol{G}_{k,t}) \le C_T$ simply by saying that the matrices $\boldsymbol{G}_{k,t}$ are 'well conditioned'.

A.5 Proof of Theorem 40

The proof will be by induction over t. The base case is clear, so we focus on the inductive step. We assume the statements above for $t-1$ and prove them for t.

A.5.1 Proof of (a)

Note that $\mathcal{P}_t^{\perp}(\boldsymbol{A}^{(p)})$ is by construction independent of $\mathcal{F}_t$, and therefore we can replace $\boldsymbol{A}^{(p)}$ by a fresh independent matrix in Eq. (A.3.1), whence we get the desired expression.

A.5.2 Proof of (b): Approximate isometry

We will repeatedly apply Lemma A.2.1. We start with Eq. (A.4.3). As we are inducting on t, we may limit ourselves to considering inner products $\langle \boldsymbol{q}^{p,t+1}, \boldsymbol{q}^{p,u+1}\rangle_N$, for $u \le t$. We first state a useful lemma.

Lemma A.5.1. *For all $\boldsymbol{v} \in \mathbb{R}^N$ $\mathcal{F}_t$-measurable,*

$$\left\langle \boldsymbol{v}, \mathcal{P}_t^{\perp}(\boldsymbol{A}^{(p)})\{\mathbf{f}_t\} \right\rangle_N \overset{p}{\simeq} 0. \tag{A.5.1}$$

Moreover,

$$\mathcal{P}_t^{\perp}(\tilde{\boldsymbol{A}}^{(p)})\{(\mathbf{f}_t^{\otimes p-1})_{\perp}\} \overset{p}{\simeq} \tilde{\boldsymbol{A}}^{(p)}\{(\mathbf{f}_t^{\otimes p-1})_{\perp}\}\,. \tag{A.5.2}$$

Using the first assertion,Eq. (A.5.1), of the above lemma, we get, for $u \le t-1$,

$$\langle \boldsymbol{q}^{p,t+1}, \boldsymbol{q}^{p,u+1} \rangle_N \overset{p}{\simeq} \langle \mathbb{E}[\boldsymbol{q}^{p,t+1}|\mathcal{F}_t], \boldsymbol{q}^{p,u+1} \rangle_N. \tag{A.5.3}$$

We next use the formula in (a) for $\mathbb{E}[\boldsymbol{q}^{p,t+1}|\mathcal{F}_t]$ (together with the expression in Eq. (A.3.1)):

$$\begin{aligned}
\langle \mathbb{E}[\boldsymbol{q}^{p,t+1}|\mathcal{F}_t], \boldsymbol{q}^{p,u+1} \rangle_N &\overset{p}{\simeq} \left\langle \sum_{0\le r,s\le t-1} \boldsymbol{q}^{p,s+1} (\boldsymbol{G}^{-1}_{p-1,t-1})_{s,r} \langle \mathbf{f}_r, \mathbf{f}_t \rangle_N^{p-1}, \boldsymbol{q}^{p,u+1} \right\rangle_N && \text{(A.5.4)}\\
&= \sum_{0\le r,s\le t-1} \langle \boldsymbol{q}^{p,s+1}, \boldsymbol{q}^{p,u+1} \rangle_N (\boldsymbol{G}^{-1}_{p-1,t-1})_{s,r} \langle \mathbf{f}_r, \mathbf{f}_t \rangle_N^{p-1} && \text{(A.5.5)}\\
&\overset{p}{\simeq} pc_p^2 \sum_{0\le r,s\le t-1} (\boldsymbol{G}_{p-1,t-1})_{s,u} (\boldsymbol{G}^{-1}_{p-1,t-1})_{s,r} \langle \mathbf{f}_r, \mathbf{f}_t \rangle_N^{p-1} && \text{(A.5.6)}\\
&= pc_p^2 \langle \mathbf{f}_u, \mathbf{f}_t \rangle_N^{p-1}. && \text{(A.5.7)}
\end{aligned}$$

The third equality was obtained by the induction hypothesis. We next prove Eq. (A.4.3) when $u = t$. We set $(\mathbf{f}_t^{\otimes p-1})_{\|}$ to be the projection of $\mathbf{f}_t^{\otimes p-1}$ onto $\mathrm{span}(\mathbf{f}_s^{\otimes p-1})_{s<t}$ and $(\mathbf{f}_t^{\otimes p-1})_{\perp} = \mathbf{f}_t^{\otimes p-1} - (\mathbf{f}_t^{\otimes p-1})_{\|}$. We then have

$$\mathcal{P}_t^{\perp}(\tilde{\boldsymbol{A}}^{(p)})\{\mathbf{f}_t\} = \mathcal{P}_t^{\perp}(\tilde{\boldsymbol{A}}^{(p)})\{(\mathbf{f}_t^{\otimes p-1})_{\perp}\}\,,$$

where the right-hand side is defined according to Eq. (A.0.2). Using the second assertion, Eq. (A.5.2), of Lemma A.5.1 and Lemma A.2.1 (point 4), we have

$$\left\|\mathcal{P}_t^{\perp}(\tilde{\boldsymbol{A}}^{(p)})\{\mathbf{f}_t\}\right\|_N^2 \overset{p}{\simeq} \frac{pc_p^2}{N^{p-1}} \|(\mathbf{f}_t^{\otimes p-1})_{\perp}\|^2\,. \tag{A.5.8}$$

Further, again using $\mathcal{P}_t^{\perp}(\tilde{\boldsymbol{A}}^{(p)})\{(\mathbf{f}_t^{\otimes p-1})_{\perp}\} \overset{p}{\simeq} \tilde{\boldsymbol{A}}^{(p)}\{(\mathbf{f}_t^{\otimes p-1})_{\perp}\}$, and Lemma A.2.1 (point 2) we obtain

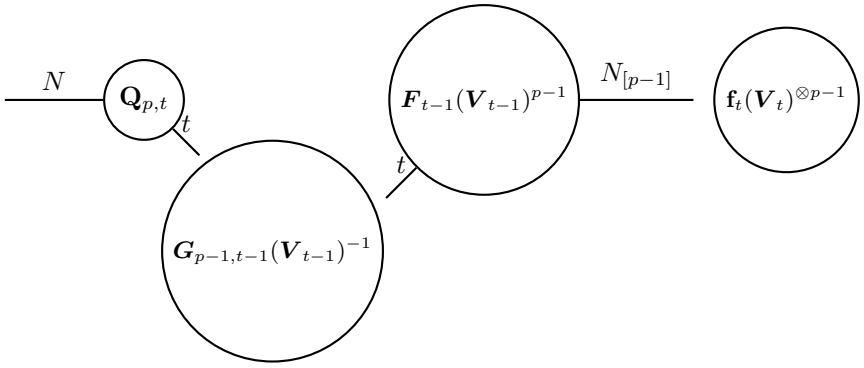
$$\langle \mathcal{P}_t^{\perp}(\tilde{\boldsymbol{A}}^{(p)})\{\mathbf{f}_t\}, \mathbb{E}[\boldsymbol{q}^{p,t+1}|\mathcal{F}_t] \rangle_N \overset{p}{\simeq} 0\,. \tag{A.5.9}$$

We next claim that

$$\left\| \mathbb{E}[\boldsymbol{q}^{p,t+1}|\mathcal{F}_t] \right\|_N^2 \overset{p}{\simeq} \frac{pc_p^2}{N^{p-1}} \|(\mathbf{f}_t^{\otimes p-1})_{\|}\|^2 \,. \tag{A.5.10}$$

In order to prove this, recall the expression for $\mathbb{E}[\boldsymbol{q}^{p,t+1}|\mathcal{F}_t]$ from part (a), and the corresponding tensor network diagram which we reproduce here

$\mathbb{E}[\boldsymbol{q}^{p,t+1}|\mathcal{F}_t] =$



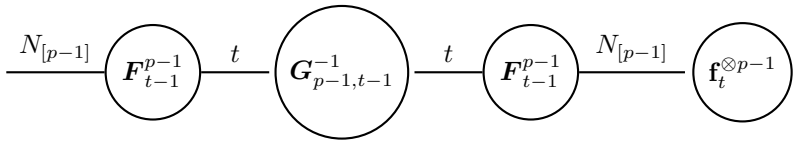
Further, by the formula for simple linear regression, we have

$$(\mathbf{f}_t^{\otimes p-1})_{\|} = \sum_{0\le s\le t-1} \alpha_{s,t} \mathbf{f}_s^{\otimes p-1}, \tag{A.5.11}$$

$$\alpha_{s,t} = \sum_{0\le r\le t-1} (\boldsymbol{G}_{p-1,t-1}^{-1})_{s,r} \langle \mathbf{f}_r, \mathbf{f}_t \rangle_N^{p-1} \,. \tag{A.5.12}$$

This can be represented by a tensor network as follows:

$(\mathbf{f}_t^{\otimes p-1})_{||} =$



However by part (b) of the inductive step, $\sqrt{p}c_p \boldsymbol{F}_{t-1}^{\otimes p-1}$ and $\mathbf{Q}_{p,t}$ are approximately unitarily equivalent in that $pc_p^2 \langle \mathbf{f}_r, \mathbf{f}_s \rangle_N^{p-1} \overset{p}{\simeq} \langle \boldsymbol{q}_{p,r+1}, \boldsymbol{q}_{p,s+1} \rangle_N$. Therefore the above expressions have approximately the same norm up to the factor $p^{1/2}c_p$, since they are linear combinations with the same coefficients:

$$\left\| \mathbb{E}[\boldsymbol{q}^{p,t+1}|\mathcal{F}_t] \right\|_N^2 \overset{p}{\simeq} \frac{pc_p^2}{N^{p-1}} \|(\mathbf{f}_t^{\otimes p-1})_{\|}\|^2. \tag{A.5.13}$$

Using together Eqs. (A.5.8), (A.5.9), and (A.5.13), we get

$$\begin{aligned}\langle \boldsymbol{q}^{p,t+1}, \boldsymbol{q}^{p,t+1}\rangle_N &\overset{p}{\simeq} \|\,\mathbb{E}[\boldsymbol{q}^{p,t+1}|\mathcal{F}_t]\|_N^2 + \frac{pc_p^2}{N^{p-1}}\|(\mathbf{f}_t^{\otimes p-1})_\perp\|_N^2 \\ &\overset{p}{\simeq} \frac{pc_p^2}{N^{p-1}}\langle \mathbf{f}_t^{\otimes p-1}, \mathbf{f}_t^{\otimes p-1}\rangle_N \\ &= pc_p^2\langle \mathbf{f}_t, \mathbf{f}_t\rangle_N^{p-1}\end{aligned}$$

finishing the proof of Eq. (A.4.3).

Next consider Eq. (A.4.5), i.e., approximate orthogonality of $\boldsymbol{q}^{p,r}$ and $\boldsymbol{q}^{p',r}$ for $p \neq p'$. This follows easily from the representation in point (a) which, together with Lemma A.2.1, inductively implies that the iterates $\boldsymbol{q}^{s,p}$ for different p are approximately orthogonal. Finally, Eq. (A.4.4) follows directly from Eq. (A.4.3) and (A.4.5). We now prove Lemma A.5.1.

Proof of Lemma A.5.1. Since $\boldsymbol{v}$ is $\mathcal{F}_t$-measurable, we can replace w.l.o.g. $\boldsymbol{A}^{(p)}$ with a fresh random tensor $\tilde{\boldsymbol{A}}$ independent of everything else. By Lagrange multipliers, there exists $(\boldsymbol{\lambda}_s)_{s\le t-1}$ vectors in $\mathbb{R}^N$ such that $\mathcal{P}_t^\perp(\tilde{\boldsymbol{A}}) = \tilde{\boldsymbol{A}} - \mathbf{Q}$, where

$$\mathbf{Q} = \frac{(p-1)!}{N^{p-1}}\sum_{s=0}^{t-1}\sum_{j=1}^{p}\underbrace{\mathbf{f}_s\otimes\cdots\otimes\mathbf{f}_s}_{j-1 \text{ times}}\otimes\boldsymbol{\lambda}_s\otimes\underbrace{\mathbf{f}_s\otimes\cdots\otimes\mathbf{f}_s}_{p-j \text{ times}}.$$

The vectors $(\boldsymbol{\lambda}_s)_{s\le t-1}$ are determined by the set of equations $\mathcal{P}_t^\perp(\tilde{\boldsymbol{A}})\{\mathbf{f}_s\} = 0$ for all $s \le t-1$ which are equivalent to

$$\sum_{r<t}(\boldsymbol{G}_{p-1,t-1})_{s,r}\boldsymbol{\lambda}_r + (p-1)\sum_{r<t}(\boldsymbol{G}_{p-2,t-1})_{s,r}\langle \mathbf{f}_s, \boldsymbol{\lambda}_r\rangle_N\mathbf{f}_r = \tilde{\boldsymbol{A}}\{\mathbf{f}_s\}\,.$$

Multiplying these equations by $\boldsymbol{G}_{p-1,t-1}^{-1}$ (recall that we assume $\boldsymbol{G}_{p-1,t-1}$ well conditioned with high probability), we obtain

$$\boldsymbol{\lambda}_s + (p-1)\sum_{r',r<t}(\boldsymbol{G}_{p-1,t-1}^{-1})_{s,r'}(\boldsymbol{G}_{p-2,t-1})_{r',r}\langle \mathbf{f}_{r'}, \boldsymbol{\lambda}_r\rangle_N\mathbf{f}_r = \sum_{r<t}(\boldsymbol{G}_{p-1,t-1}^{-1})_{s,r}\tilde{\boldsymbol{A}}\{\mathbf{f}_r\}\,. \tag{A.5.14}$$

This in particular implies that

$$\boldsymbol{\lambda}_s = \boldsymbol{\lambda}_s^0 + \boldsymbol{\lambda}_s^\parallel\,, \quad \boldsymbol{\lambda}_s^0 \equiv \sum_{r<t}(\boldsymbol{G}_{p-1,t-1}^{-1})_{s,r}\tilde{\boldsymbol{A}}\{\mathbf{f}_r\}\,,$$

where $\boldsymbol{\lambda}_s^\parallel \in \mathrm{span}((\mathbf{f}_r)_{r<t})$. We claim that $\|\boldsymbol{\lambda}^\parallel\|_N \overset{p}{\simeq} 0$, i.e., $\boldsymbol{\lambda}_s \overset{p}{\simeq} \boldsymbol{\lambda}_s^0$. Indeed, letting $\boldsymbol{\Lambda} \in \mathbb{R}^{N\times t}$ be the matrix with columns $(\boldsymbol{\lambda}_s)_{s<t}$, and $\boldsymbol{\Lambda}^0$ the matrix with columns $(\boldsymbol{\lambda}_s^0)_{s<t}$ Eq. (A.5.14) can be written

as

$$\mathcal{L}_{p,t}^{\mathsf{T}}(\mathbf{\Lambda}) = \mathbf{\Lambda}^0 \,.$$

Here we recall $\mathcal{L}_{p,t} = \mathbf{1} + (p-1)\mathcal{T}_{p.t}$ and $\mathcal{T}_{p,t} \in \mathbb{R}^{Nt \times Nt}$ is defined in Eq. (A.2.7). Substituting the decomposition $\mathbf{\Lambda} = \mathbf{\Lambda}^0 + \mathbf{\Lambda}^{\|}$ in the above, we obtain

$$\mathcal{L}_{p,t}^{\mathsf{T}}(\mathbf{\Lambda}^{\|}) = -(p-1)\mathcal{T}_{p,t}^{\mathsf{T}}(\mathbf{\Lambda}^0) \,.$$

Since by assumption $\mathcal{L}_{p,t}$ is well conditioned, it is sufficient to prove that $\mathcal{T}_{p,t}^{\mathsf{T}}(\mathbf{\Lambda}^0) \overset{p}{\simeq} 0$. Let $\boldsymbol{c}_0, \cdots, \boldsymbol{c}_{t-1} \in \mathbb{R}^N$ be the columns of the matrix $\mathcal{T}_{p,t}^{\mathsf{T}}(\mathbf{\Lambda}^0)$. Since $\boldsymbol{c}_s \in \mathrm{span}((\mathbf{f}_r)_{r<t})$ for all $s < t$, and the Gram matrix $\boldsymbol{G}_{1,t-1} = (\langle \mathbf{f}_r, \mathbf{f}_s \rangle)_{r,s<t}$ is well conditioned, it is sufficient to check that $\langle \mathbf{f}_r, \boldsymbol{c}_s \rangle_N \overset{p}{\simeq} 0$ for each $s, r < t$. This is in turn equivalent to

$$\sum_{r',r<t} (\boldsymbol{G}_{p-1,t-1}^{-1})_{s,r'} (\boldsymbol{G}_{p-2,t-1})_{r',r} \langle \mathbf{f}_{r'}, \boldsymbol{\lambda}_r^0 \rangle_N \langle \mathbf{f}_r, \mathbf{f}_q \rangle_N \overset{p}{\simeq} 0 \,,$$

for all $s, q < t$. Finally, this last claim follows by substituting the expression for $\boldsymbol{\lambda}_r^0$, and using the fact that $\langle \mathbf{f}_s, \tilde{\boldsymbol{A}}\{\mathbf{f}_q\} \rangle_N \overset{p}{\simeq} 0$ for all $r, q \le t$, by Lemma A.2.1.

We are now in position to prove the claim of this lemma. For the first assertion, we have

$$\begin{aligned}
\langle \boldsymbol{v}, \mathbf{Q}\{\mathbf{f}_t\} \rangle_N &= \sum_{s \le t-1} \langle \mathbf{f}_s, \mathbf{f}_t \rangle_N^{p-1} \langle \boldsymbol{\lambda}_s, \boldsymbol{v} \rangle_N + (p-1) \sum_{s \le t-1} \langle \mathbf{f}_s, \mathbf{f}_t \rangle_N^{p-2} \langle \boldsymbol{\lambda}_s, \mathbf{f}_t \rangle_N \langle \mathbf{f}_s, \boldsymbol{v} \rangle_N \\
&\overset{p}{\simeq} \sum_{s \le t-1} \langle \mathbf{f}_s, \mathbf{f}_t \rangle_N^{p-1} \langle \boldsymbol{\lambda}_s^0, \boldsymbol{v} \rangle_N + (p-1) \sum_{s \le t-1} \langle \mathbf{f}_s, \mathbf{f}_t \rangle_N^{p-2} \langle \boldsymbol{\lambda}_s^0, \mathbf{f}_t \rangle_N \langle \mathbf{f}_s, \boldsymbol{v} \rangle_N .
\end{aligned}$$

Further, using Lemma A.2.1 (point 2), we have for any $\boldsymbol{u} \in \mathbb{R}^N$ which is $\mathcal{F}_t$-measurable,

$$\langle \boldsymbol{\lambda}_s^0, \boldsymbol{u} \rangle_N = \sum_{r \le t-1} (\boldsymbol{G}_{p-1,t-1}^{-1})_{s,r} \langle \tilde{\boldsymbol{A}}\{\mathbf{f}_r\}, \boldsymbol{u} \rangle_N \overset{p}{\simeq} 0.$$

Whence $\left\langle \boldsymbol{v}, \mathbf{Q}\{\mathbf{f}_t\} \right\rangle_N \overset{p}{\simeq} 0$, and, $\left\langle \boldsymbol{v}, \mathcal{P}_t^{\perp}(\tilde{\boldsymbol{A}})\{\mathbf{f}_t\} \right\rangle_N = \left\langle \boldsymbol{v}, \tilde{\boldsymbol{A}}\{\mathbf{f}_t\} \right\rangle_N - \left\langle \boldsymbol{v}, \mathbf{Q}\{\mathbf{f}_t\} \right\rangle_N \overset{p}{\simeq} 0$.

We next prove the second assertion of the lemma, Eq. (A.5.2).

Note that $\tilde{\boldsymbol{A}}\{(\mathbf{f}_t^{\otimes p-1})_{\perp}\} - \mathcal{P}_t^{\perp}(\tilde{\boldsymbol{A}})\{(\mathbf{f}_t^{\otimes p-1})_{\perp}\} = \mathbf{Q}\{(\mathbf{f}_t^{\otimes p-1})_{\perp}\}$ and

$$\mathbf{Q}\{(\mathbf{f}_t^{\otimes p-1})_{\perp}\} = \frac{(p-1)}{N^{p-1}} \sum_{s<t} \langle \boldsymbol{\lambda}_s \otimes \mathbf{f}_s^{\otimes (p-2)}, (\mathbf{f}_t^{\otimes (p-1)})_{\perp} \rangle \, \mathbf{f}_s \equiv (p-1) \sum_{s \le t} c_s \mathbf{f}_s \,.$$

Since the Gram matrix $\boldsymbol{G}_{1,t-1} = (\langle \mathbf{f}_s, \mathbf{f}_r \rangle)_{s,r<t}$ is well conditioned, in order to show $\|\mathbf{Q}\{(\mathbf{f}_t^{\otimes p-1})_{\perp}\}\|_N \overset{p}{\simeq} 0$, it is sufficient to check that each of the coefficients $c_s \overset{p}{\simeq} 0$ for each s. Notice that $(\mathbf{f}_t^{\otimes (p-1)})_{\perp} =$

$\sum_{r\le t}\beta_r \mathbf{f}_r^{\otimes(p-1)}$, where the β_s are bounded thanks to the fact that $\boldsymbol{G}_{p-1,t-1}$ is well conditioned. Using $\boldsymbol{\lambda}_s \overset{p}{\simeq} \boldsymbol{\lambda}_s^0$, we get

$$\begin{aligned} c_s &\overset{p}{\simeq} \frac{1}{N^{p-1}}\langle \boldsymbol{\lambda}_s^0 \otimes \mathbf{f}_s^{\otimes(p-2)}, (\mathbf{f}_t^{\otimes p-1})_\perp\rangle \\ &= \frac{1}{N^{p-1}}\sum_{r\le t}\sum_{q<t}\beta_r(\boldsymbol{G}^{-1}_{p-1,t-1})_{s,r'}\langle \tilde{\boldsymbol{A}}\{\mathbf{f}_q\}\otimes \mathbf{f}_s^{\otimes(p-2)}, \mathbf{f}_r^{\otimes(p-1)}\rangle_N \\ &= \sum_{r\le t}\sum_{q<t}\beta_r(\boldsymbol{G}^{-1}_{p-1,t-1})_{s,r'}\langle \tilde{\boldsymbol{A}}\{\mathbf{f}_q\}, \mathbf{f}_t\rangle_N \langle \mathbf{f}_s, \mathbf{f}_r\rangle_N^{p-2} \overset{p}{\simeq} 0\,, \end{aligned}$$

where in the last step we used $\langle \tilde{\boldsymbol{A}}\{\mathbf{f}_q\}, \mathbf{f}_t\rangle_N \overset{p}{\simeq} 0$, thanks to Lemma A.2.1. □

A.5.3 Proof of (c)

Recall that the process $(U^{p,t})_{t\ge 1}$ is Gaussian by construction, and independent of $U^{p,0}$, Define $C_{r,s} = \mathbb{E}\{U^{p,r}U^{p,s}\}$ and $\boldsymbol{C}_{\le t} = (C_{r,s})_{r,s\le t}$. We then have

$$\mathbb{E}[U^{p,t+1}|U^{p,0},\dots,U^{p,t}] = \sum_{s=1}^{t}\tilde{\alpha}_s U^{p,s}\,, \tag{A.5.15}$$

$$\tilde{\alpha}_s = \sum_{r=1}^{t}(\boldsymbol{C}^{-1}_{\le t})_{s,r}C_{r,t+1}\,. \tag{A.5.16}$$

On the other hand, from point (a), we know that

$$\mathbb{E}[\boldsymbol{q}^{p,t+1}|\mathcal{F}_t] = \sum_{1\le s\le t}\alpha_s \boldsymbol{q}^{s,p}\,, \tag{A.5.17}$$

$$\alpha_s = \sum_{r=1}^{t}(\boldsymbol{G}^{-1}_{p-1,t-1})_{s-1,r-1}(\boldsymbol{G}_{p-1,t})_{r-1,t}\,. \tag{A.5.18}$$

Moreover, by the induction hypothesis we know that, for $r,s\le t$

$$(\boldsymbol{G}_{p-1,t})_{r,s} \overset{p}{\simeq} \mathbb{E}\{f_r(X^0,\dots,X^r)f_t(X^0,\dots,X^s)\}^{p-1}\,,$$

where we recall that $X^t \equiv \sum_{p\le D}U^{p,t}$. Therefore, using the definition of the process $(U^{p,t})_{t\ge 0}$ we obtain $(\boldsymbol{G}_{p-1,t})_{r,s} \overset{p}{\simeq} C_{r+1,s+1}/(pc_p^2)$ for $r,s\le t$, whence $\alpha_s \overset{p}{\simeq} \tilde{\alpha}_s$ (where we used the fact that $\boldsymbol{G}_{p-1,t}$

is well conditioned by assumption). Therefore we also have

$$\begin{aligned}\Big\| \mathbb{E}[\boldsymbol{q}^{p,t+1}|\mathcal{F}_t] - \sum_{s=1}^{t} \tilde{\alpha}_s \boldsymbol{q}^{p,s} \Big\|_N^2 &= \Big\| \sum_{s=1}^{t} (\alpha_s - \tilde{\alpha}_s) \boldsymbol{q}^{p,s} \Big\|_N^2 \\ &= \sum_{s,r=1}^{t} (\alpha_s - \tilde{\alpha}_s)(\alpha_r - \tilde{\alpha}_r) \langle \boldsymbol{q}^{p,s}, \boldsymbol{q}^{p,r} \rangle_N \\ &\overset{p}{\simeq} \sum_{s,r=1}^{t} (\alpha_s - \tilde{\alpha}_s)(\alpha_r - \tilde{\alpha}_r) C_{r,s} \overset{p}{\simeq} 0. \end{aligned} \tag{A.5.19}$$

Moreover, Lemma A.2.1 (point 4) shows that $\mathcal{P}_t^{\perp}(\tilde{\boldsymbol{A}}^{(p)})\{\mathbf{f}_t\} \overset{p}{\simeq} \tilde{\boldsymbol{A}}^{(p)}\{(\mathbf{f}_t^{\otimes p-1})_{\perp}\}$ has entries which are approximately independent Gaussian with variance $\sigma_t^2 \equiv p c_p^2 \|(\mathbf{f}_t^{\otimes p-1})_{\perp}\|^2/N^{p-1}$, even conditionally on $\mathcal{F}_t$. Therefore

$$\boldsymbol{q}^{p,t+1} \overset{\mathrm{d}}{=} \sum_{s=1}^{t} \tilde{\alpha}_s \boldsymbol{q}^{p,s} + \sigma_t \mathbf{g} + \boldsymbol{e}^{p,t+1}\,, \tag{A.5.20}$$

where $\|\boldsymbol{e}\|_N \overset{p}{\simeq} 0$ and $\mathbf{g} \sim \mathsf{N}(\mathbf{0}, \boldsymbol{I}_N)$ is independent of everything else. From here on, the rest of the argument for state evolution for pseudo-Lipschitz functions is exactly the same as in Lemma 5 (b) in [BMN19]. As proved in the previous point, for any $s \le t$,

$$\langle \boldsymbol{q}^{p,t+1}, \boldsymbol{q}^{p,s+1} \rangle_N^2 \overset{p}{\simeq} p c_p^2 \langle \mathbf{f}_t, \mathbf{f}_s \rangle_N^{p-1} \overset{p}{\simeq} \mathbb{E}\{U^{p,t+1} U^{p,s+1}\}\,.$$

Therefore, in order to prove Eq. (A.4.6),it is sufficient to consider $\psi : \mathbb{R}^{D \times t+1} \to \mathbb{R}$ Lipschitz. Using the representation (A.5.20), and focusing for simplicity on a single p, we get

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} \psi(\boldsymbol{q}_i^{p,\le t}, q_i^{p,t+1}) &\overset{p}{\simeq} \frac{1}{N}\sum_{i=1}^{N} \psi\left(\boldsymbol{q}_i^{p,\le t}, \sum_{s=1}^{t} \tilde{\alpha}_s \boldsymbol{q}^{p,s} + \sigma_t g_i\right) \\ &\overset{p}{\simeq} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\,\psi\left(\boldsymbol{q}_i^{p,\le t}, \sum_{s=1}^{t} \tilde{\alpha}_s \boldsymbol{q}^{p,s} + \sigma_t G\right)\,,\end{aligned}$$

where the second equality follows by Gaussian concentration. At this point we apply the induction hypothesis.

A.6 Asymptotic equivalence of Tensor AMP and Tensor LAMP

Here we show that tensor AMP and tensor LAMP produce approximately the same iterates.

Lemma A.6.1. *Let $\{\boldsymbol{W}^{(p)}\}_{p\le D}$ be standard Gaussian tensors, and $\boldsymbol{A}^{(p)} = c_p\boldsymbol{W}^{(p)}$ for $p \ge 2$. Consider the corresponding AMP iterates $\boldsymbol{Z}_t \equiv (\boldsymbol{z}^{p,s})_{p\le D,s\le t}$ and LAMP iterates $\mathbf{Q}_t \equiv (\boldsymbol{q}^{p,s})_{p\le D,s\le t}$, from the same initialization initialization $\boldsymbol{Z}_0 = \mathbf{Q}_0$ satisfying the assumptions of Theorem 39 and Theorem 40.*

Let $\mathbf{f}_t = f_t(\boldsymbol{V}_t)$, $t \ge 0$ be the nonlinearities applied to LAMP iterates and $(\boldsymbol{G}_{p,t}(\boldsymbol{V}))_{r,s} = \langle \mathbf{f}_t, \mathbf{f}_s\rangle^p$ be the corresponding Gram matrices. Further assume that there exist a constant $C < \infty$ such that, for all $t \le T$,

(i) The LAMP Gram matrices $\boldsymbol{G}_{p,t} = \boldsymbol{G}_{p,t}$ are well-conditioned, i.e., $C^{-1} \le \sigma_{\min}(\boldsymbol{G}_{p,t}) \le \sigma_{\max}(\boldsymbol{G}_{p,t}) \le C$ for all $p \le D$, $t \le T$.

(ii) Let the linear operator $\mathcal{T}_{p,t} : \mathbb{R}^{N\times t} \to \mathbb{R}^{N\times t}$ be defined as per Eq. (A.2.7), with $\boldsymbol{G}_{p,t} = \boldsymbol{G}_{p,t}(\boldsymbol{V})$, and $\mathbf{f}_t = f_t(\boldsymbol{V})$, and define $\mathcal{L}_{p,t} = \mathbf{1} + (p-1)\mathcal{T}_{p,t}$. Then $C^{-1} \le \sigma_{\min}(\mathcal{L}_{p,t}) \le \sigma_{\max}(\mathcal{L}_{p,t}) \le C$.

Then, for any $t \le T$, we have

$$\|\boldsymbol{Z}_t - \mathbf{Q}_t\|_N \overset{p}{\simeq} 0\,. \tag{A.6.1}$$

Proof. Throughout the proof we will write $f_t(\boldsymbol{X}_t)$ or $f_t(\boldsymbol{V}_t)$ to distinguish AMP and LAMP iterates, and analogously for $\boldsymbol{G}_{p,t}(\boldsymbol{X}_t)$ or $\boldsymbol{G}_{p,t}(\boldsymbol{V}_t)$. The proof is by induction over the iteration number, so we will assume it to hold at iteration t, and prove it for iteration $t+1$. We prove the induction step by establishing the following two facts:

$$\big\|\mathsf{AMP}_{t+1}(\boldsymbol{Z}_t)_p - \mathsf{AMP}_{t+1}(\mathbf{Q}_t)_p\big\|_N \overset{p}{\simeq} 0\,, \tag{A.6.2}$$

$$\big\|\mathsf{AMP}_{t+1}(\mathbf{Q}_t)_p - \mathsf{LAMP}_{t+1}(\mathbf{Q}_t)_p\big\|_N \overset{p}{\simeq} 0\,. \tag{A.6.3}$$

Let us first consider the claim (A.6.2), and note that

$$\begin{aligned}\mathsf{AMP}_{t+1}(\boldsymbol{Z}_t)_p - \mathsf{AMP}_{t+1}(\mathbf{Q}_t)_p = \boldsymbol{A}^{(p)}\{f_t(\boldsymbol{X}_t)\} - \boldsymbol{A}^{(p)}\{f_t(\boldsymbol{V}_t)\}\\ - \sum_{s\le t} d_{t,s,p}\big(f_{s-1}(\boldsymbol{X}_{s-1}) - f_{s-1}(\boldsymbol{V}_{s-1})\big)\,,\end{aligned}$$

where we wrote $d_{t,s,p}$ for the coefficients of Eq. (A.0.7), with AMP iterates replaced by LAMP iterates. We then have

$$\big\|\mathsf{AMP}_{t+1}(\boldsymbol{Z}_t)_p - \mathsf{AMP}_{t+1}(\mathbf{Q}_t)_p\big\|_N \le D_{1,t} + D_{2,t}\,, \tag{A.6.4}$$

$$D_{1,t} \equiv \big\|\boldsymbol{A}^{(p)}\{f_t(\boldsymbol{X}_t)\} - \boldsymbol{A}^{(p)}\{f_t(\boldsymbol{V}_t)\}\big\|_N\,, \tag{A.6.5}$$

$$D_{2,t} \equiv \sum_{s\le t} |d_{t,s,p}|\big\|f_{s-1}(\boldsymbol{X}_{s-1}) - f_{s-1}(\boldsymbol{V}_{s-1})\big\|_N\,. \tag{A.6.6}$$

Notice that, by the induction assumption (and recalling that f_t is Lipschitz continuous and acts component-wise):

$$\left\|f_t(\boldsymbol{X}_t) - f_t(\boldsymbol{V}_t)\right\|_N \le C_T \sum_{s\le t, p\le D} \|\boldsymbol{x}^{p,s} - \boldsymbol{v}^{p,s}\|_N \overset{p}{\simeq} 0\,. \tag{A.6.7}$$

Further, for any tensor $\boldsymbol{T} \in (\mathbb{R}^N)^{\otimes p}$, and any vectors $\boldsymbol{v}_1, bv_2 \in \mathbb{R}^N$,

$$\|\boldsymbol{T}\{\boldsymbol{v}_1\} - \boldsymbol{T}\{\boldsymbol{v}_2\}\|_N \le (N^{\frac{p-2}{2}}\|\boldsymbol{T}\|_{\text{op}})(\|\boldsymbol{v}_1\|_N + \|\boldsymbol{v}_2\|_N)^{p-2}\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_N \tag{A.6.8}$$

Using Lemma A.2.1, this implies that the following bound holds with high probability for a constant C:

$$\begin{aligned} D_{1,t} &\le C(\|f_t(\boldsymbol{X}_t)\|_N + \|f_t(\boldsymbol{V}_t)\|_N)^{p-2}\|f_t(\boldsymbol{X}_t) - f_t(\boldsymbol{V}_t)\|_N & \text{(A.6.9)}\\ &\le C(2\|f_t(\boldsymbol{V}_t)\|_N + \|f_t(\boldsymbol{X}_t) - f_t(\boldsymbol{V}_t)\|_N)^{p-2}\|f_t(\boldsymbol{X}_t) - f_t(\boldsymbol{V}_t)\|_N \overset{p}{\simeq} 0 & \text{(A.6.10)}\end{aligned}$$

Where the last step follows from Eq. (A.6.7) and Theorem 40, which implies (using the fact that f_t is Lipschitz) $\|f_t(\boldsymbol{V}_t)\|_N \le C$ with high probability. Notice that the same argument implies $\|f_t(\boldsymbol{X}_t)\|_N \le C$ with high probability.

Similarly, $D_{2,t} \overset{p}{\simeq} 0$ follows since $\|f_{s-1}(\boldsymbol{X}_{s-1}) - f_{s-1}(\boldsymbol{V}_{s-1})\|_N \overset{p}{\simeq} 0$ and $|d_{t,s,p}| \le C_T$ by construction, thus yielding the desired claim (A.6.2).

We now turn to proving Eq. (A.6.3). Comparing Eq. (A.0.7) and (A.3.1), and letting $\mathcal{P}_t^{\|} = \mathbf{1} - \mathcal{P}_t^{\perp}$ we obtain

$$\mathsf{AMP}_{t+1}(\mathbf{Q}_t)_p - \mathsf{LAMP}_{t+1}(\mathbf{Q}_t)_p = \mathcal{P}_t^{\|}(\boldsymbol{A}^{(p)})\{f_t(\boldsymbol{V}_t)\} - \mathbf{ons}_{p,t+1} - \sum_{0\le s\le t-1} h_{t,s,p}\boldsymbol{q}^{p,s+1}\,, \tag{A.6.11}$$

$$\mathbf{ons}_{p,t+1} = \sum_{s\le t} d_{t,s,p} f_{s-1}(\boldsymbol{V}_{s-1}) \tag{A.6.12}$$

Note that $\mathcal{P}_t^{\|}(\boldsymbol{A}^{(p)}) = \mathbb{E}\{\boldsymbol{A}^{(p)}|\mathcal{F}_t\}$, where $\mathcal{F}_t$ is the σ-algebra generated by $\{\mathbf{q}^{p,s}\}_{s\le t, p\le D}$. Equivalently, this is the conditional expectation of $\boldsymbol{A}^{(p)}$ given the linear constraints

$$\boldsymbol{A}^{(p)}\{f_s(\boldsymbol{V}_s)\} = \boldsymbol{y}_{p,s+1}\,, \qquad \text{for } s \in \{0,\dots,t-1\}\,, \tag{A.6.13}$$

Also notice that, by the induction hypothesis, and the definition of $\boldsymbol{y}_{p,s}$, Eq. (A.1.1), we have for all $s \le t$,

$$\boldsymbol{y}_{p,s} \overset{p}{\simeq} \boldsymbol{q}^{p,s} + \mathbf{ons}_{p,s}\,. \tag{A.6.14}$$

Lemma A.2.2 implies that $\mathcal{P}_t^{\|}(\boldsymbol{A}^{(p)})$ takes the form of Eq. (A.2.6) for a suitable matrix $\hat{\boldsymbol{Z}}_{p,t} \in \mathbb{R}^{N\times t}$.

The key claim is that

$$\hat{\boldsymbol{Z}}_{p,t} \overset{p}{\simeq} \mathbf{Q}_t \,. \tag{A.6.15}$$

In order to establish this claim, we show that, under the inductive hypothesis,

$$(\mathbf{1} + (p-1)\mathcal{T}_{p,t})\mathbf{Q}_t \overset{p}{\simeq} \boldsymbol{Y}_{p,t}.$$

Since $\mathcal{L}_{p,t} = \mathbf{1} + (p-1)\mathcal{T}_{p,t}$ is well-conditioned by assumption, Eq. (A.2.8) implies $\hat{\boldsymbol{Z}}_{p,t} \overset{p}{\simeq} \mathbf{Q}_t$. Notice that, by Eq. (A.6.14) in order to prove this claim, it is sufficient to show that $(p-1)\mathcal{T}_t\mathbf{Q}_t \overset{p}{\simeq} \mathbf{ONS}_{p,t} := [\mathbf{ons}_{p,1}|\cdots|\mathbf{ons}_{p,t}]$.

In order to prove this claim, we use Theorem 40. Recall $C_{r,s} = \mathbb{E}\{U^{p,r}U^{p,s}\}$, $X^r = \sum_p U^{p,r}$ and $\boldsymbol{C}_{\le t} = (C_{r,s})_{r,s\le t}$. By Theorem 40, $C_{r+1,s+1} \overset{p}{\simeq} \langle \boldsymbol{q}^{p,r+1}, \boldsymbol{q}^{p,s+1}\rangle \overset{p}{\simeq} pc_p^2(\boldsymbol{G}_{p-1,t}(\boldsymbol{V}))_{r,s}$ for $r,s \le t$. This implies for any $0 \le r \le t-1$,

$$\begin{aligned}\sum_{j=0}^{t-1}(\boldsymbol{G}^{-1}_{p-1,t-1})_{rj}\langle \boldsymbol{q}^{p,j+1}, f_{t-1}(\boldsymbol{V}_{t-1})\rangle_N &\overset{p}{\simeq} pc_p^2\sum_{j=0}^{t-1}(\boldsymbol{C}^{-1}_{\le t})_{r+1,j+1}\,\mathbb{E}\{U^{p,j+1}f_{t-1}(X^0,\dots,X^{t-1})\}\\ &= pc_p^2\,\mathbb{E}\left\{\frac{\partial f_{t-1}}{\partial x^{r+1}}(X^0,\dots,X^{t-1})\right\}\mathbf{1}_{r\le t-2}\,,\end{aligned} \tag{A.6.16}$$

where we used Stein's lemma in the second equality. Using this last expression and the definition (A.0.7) allows to check we conclude $(p-1)\mathcal{T}_{p,t}\mathbf{Q}_t \overset{p}{\simeq} \mathbf{ONS}_{p,t}$ as claimed. Indeed we have

$$\begin{aligned}(p-1)\big[\mathcal{T}_{p,t}\mathbf{Q}_t\big]_t &= \sum_{r=0}^{t-1}(\boldsymbol{G}_{p-2,t-1})_{r,t-1}\mathbf{f}_r\Big(\sum_{r'=0}^{t-1}(\boldsymbol{G}^{-1}_{p-1,t-1})_{r,r'}\langle \boldsymbol{q}^{p,r'+1}, \mathbf{f}_{t-1}\rangle\Big)\\ &\overset{p}{\simeq} p(p-1)c_p^2\sum_{r=0}^{t-2}\langle \mathbf{f}_r, \mathbf{f}_{t-1}\rangle_N^{p-2}\mathbf{f}_r\cdot\mathbb{E}\left\{\frac{\partial f_{t-1}}{\partial x^{r+1}}(X^0,\dots,X^{t-1})\right\}\\ &= \mathbf{ons}_{p,t}.\end{aligned}$$

Having established Eq. (A.6.15), we can use the representation of $\mathcal{P}_t^{\|}(\boldsymbol{A}^{(p)}) = \mathbb{E}\{\boldsymbol{A}^{(p)}|\mathcal{F}_t\}$ given in Eq. (A.2.6) to get

$$\mathcal{P}_t^{\|}(\boldsymbol{A}^{(p)})\{\mathbf{f}_t\} \overset{p}{\simeq} \sum_{s\le t}\alpha_s\boldsymbol{q}^{p,s} + (p-1)\sum_{s\le t}\beta_s\mathbf{f}_s\,, \tag{A.6.17}$$

$$\alpha_s = \sum_{0\le r\le t-1}(\boldsymbol{G}^{-1}_{p-1,t-1})_{s,r}\langle f_r(\boldsymbol{V}_r), f_t(\boldsymbol{V}_t)\rangle_N^{p-1}\,, \tag{A.6.18}$$

$$\beta_s = \Big(\sum_{0\le r\le t-1}(\boldsymbol{G}^{-1}_{p-1,t-1})_{s,r}\langle \boldsymbol{q}^{p,r}, \mathbf{f}_t\rangle_N\Big)\langle \mathbf{f}_s, \mathbf{f}_t\rangle_N^{p-2}. \tag{A.6.19}$$

On the other hand, using again Eq. (A.6.16), we obtain

$$(p-1)\sum_{s\le t}\beta_s \mathbf{f}_s \overset{p}{\simeq} \sum_{s\le t-1} d_{t,s,p}\mathbf{f}_{s-1} = \mathbf{ons}_{p,t+1}, \tag{A.6.20}$$

$$\text{and} \quad \sum_{s\le t}\alpha_s \boldsymbol{q}^{p,s} \overset{p}{\simeq} \sum_{0\le s\le t-1} h_{t,s,p}\boldsymbol{q}^{p,s+1}. \tag{A.6.21}$$

We therefore conclude, from Eq. (A.6.11), that $\|\mathsf{AMP}_{t+1}(\mathbf{Q}_t)_p - \mathsf{LAMP}_{t+1}(\mathbf{Q}_t)_p\|_N \overset{p}{\simeq} 0$, and this finishes our proof. □

A.7 Reduction to the well-conditioned case

Theorem 40 and Lemma A.6.1 imply the conclusion of the main statement Theorem 39, under the additional assumptions in points (i) and (ii) of Lemma A.6.1. Here we show how to approximate an arbitrary AMP algorithm with one satisfying those conditions, completing the proof of Theorem 39. This strategy was already employed in [JM13, BMN19], and we refer to these references for further background.

Lemma A.7.1. *Let $(f_t)_{t\ge 0}$, with $f_t : \mathbb{R}^{t+1} \to \mathbb{R}$, be any sequence of Lipschitz functions. Then for any $\varepsilon > 0$ there exists a sequence of smooth functions $\varphi_t : \mathbb{R}^{t+1} \to \mathbb{R}$, with $\|\varphi_t\|_{L^\infty} \le 1$, $\|\nabla\varphi_t\|_{L^\infty} \le 1$, such that the following holds. Defining $f_t^\varepsilon = f_t + \varepsilon\varphi_t$, the sequence of functions $(f_t^\varepsilon)_{t\ge 0}$ satisfies conditions (i) and (ii) of Lemma A.6.1.*

The proof of this lemma is presented in the next two subsections, considering first condition (i), and then condition (ii). Before presenting this proof, we show that this lemma indeed allows to prove Theorem 39.

Proof of Theorem 39. Let $(f_t^\varepsilon)_{t\in\mathbb{N}}$ be a sequence of functions as per Lemma A.7.1, and denote by $\boldsymbol{z}^{\varepsilon,p,t}$ the corresponding iterates, and $\boldsymbol{Z}_t^\varepsilon = (\boldsymbol{z}^{\varepsilon,p,s})_{p\le D, s\le t}$. We instead use $\boldsymbol{Z}_t = (\boldsymbol{z}^{p,s})_{p\le D,s\le t}$ for the unperturbed AMP iteration. Using the same argument as in the proof of Lemma A.6.1 (in particular, the argument to prove Eq. (A.6.2)) we obtain, for every fixed t,

$$\operatorname*{p\text{-}lim}_{\varepsilon\to 0} \limsup_{N\to\infty} \|\boldsymbol{Z}_t - \boldsymbol{Z}_t^\varepsilon\|_N = 0\,. \tag{A.7.1}$$

On the other hand, for any $\varepsilon > 0$, the iterates satisfy the non-degeneracy conditions (i) and (ii) of Lemma A.6.1. We can therefore apply this lemma, and Theorem 40 to conclude that, for any test

pseudo-Lipschitz function $\psi : \mathbb{R}^{D\times T} \to \mathbb{R}$, we have

$$\operatorname*{p-lim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \psi((z_i^{\varepsilon,p,t})_{p\le D,t\le T}) = \mathbb{E}\left\{\psi\big((U^{\varepsilon,p,t})_{p\le D,t\le T}\big)\right\}. \tag{A.7.2}$$

Here $(U^{\varepsilon,p,t})_{p\le D,t\ge 0}$ is the Gaussian process associated to the nonlinearities $(f_t^\varepsilon)_{t\ge 0}$, namely with covariance determined recursively via

$$\mathbb{E}[U^{\varepsilon,p,t+1}U^{\varepsilon,p,s+1}] = pc_p^2\,\mathbb{E}\left\{f_t^\varepsilon\left(X^{\varepsilon,0},\dots,X^{\varepsilon,t}\right) f_s^\varepsilon\left(X^{\varepsilon,0},\dots,X^{\varepsilon,s}\right)\right\}^{p-1}, \tag{A.7.3}$$

$$X^{\varepsilon,t} \equiv \sum_{k=2}^{D} U^{\varepsilon,k,t}. \tag{A.7.4}$$

Recalling that $f_t^\varepsilon = f_t + \varepsilon\varphi_t$ with φ_t bounded, with bounded gradient, it is immediate to show by induction that $\mathbb{E}[U^{\varepsilon,p,t}U^{\varepsilon,p,s}] \to \mathbb{E}[U^{p,t}U^{p,s}]$ as $\varepsilon \to 0$. In particular, it is possible to couple $(U^{\varepsilon,p,t})_{p\le D,t\ge 0}$ and $(U^{p,t})_{p\le D,t\ge 0}$ so that $\mathbb{E}\{(U^{\varepsilon,p,t} - U^{p,t})^2\} \to 0$ for any p, t. We thus conclude that

$$\begin{aligned}\operatorname*{p-lim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \psi((z_i^{p,t})_{p\le D,t\le T}) &\overset{(a)}{=} \lim_{\varepsilon\to 0} \operatorname*{p-lim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \psi((z_i^{\varepsilon,p,t})_{p\le D,t\le T})\\ &\overset{(b)}{=} \lim_{\varepsilon\to 0} \mathbb{E}\left\{\psi\big((U^{\varepsilon,p,t})_{p\le D,t\le T}\big)\right\} \overset{(c)}{=} \mathbb{E}\left\{\psi\big((U^{\varepsilon,p,t})_{p\le D,t\le T}\big)\right\},\end{aligned}$$

where (a) follows from Eq. (A.7.1), (b) from Eq. (A.7.2), and (c) from the remark that $\mathbb{E}\{(U^{\varepsilon,p,t} - U^{p,t})^2\} \to 0$. □

A.7.1 Condition (i): Control of $\boldsymbol{G}_{p,t}$

We begin with condition (i) which requires $C^{-1} \le \sigma_{\min}(\boldsymbol{G}_{p,t}) \le \sigma_{\max}(\boldsymbol{G}_{p.t}) \le C$ with high probability for some constant C independent of N. Note that Lemma A.6.1 requires these bounds to hold for a finite collections of values of p, t. Since this collection is fixed independently of N, it is sufficient to consider a single pair (p, t). By Theorem 40, we know that

$$\operatorname*{p-lim}_{N\to\infty}(\boldsymbol{G}_{p,t})_{r,s} = (\boldsymbol{G}_{p,t}^{\infty})_{r,s} = \big(\mathbb{E}\{f_r(X_0,\dots,X_r)f_s(X_0,\dots,X_s)\}\big)^p. \tag{A.7.5}$$

It is therefore sufficient to prove $\sigma_{\min}(\boldsymbol{G}_{p,t}^{\infty}) > 0$ for all p, t. Note that $\sigma_{\min}(\boldsymbol{G}_{p,t}^{\infty}) < \infty$ is immediate since $\boldsymbol{G}_{p,t}^{\infty}$ has finite entries, and is a matrix of fixed dimensions $t + 1 \times t + 1$.

Recall that Hadamard product preserves positive-semidefinite (PSD) ordering: if $\boldsymbol{A}_1 \succeq \boldsymbol{B}_1 \succeq \boldsymbol{0}$ and $\boldsymbol{A}_2 \succeq \boldsymbol{B}_2 \succeq \boldsymbol{0}$, then $\boldsymbol{A}_1 \odot \boldsymbol{A}_2 \succeq \boldsymbol{B}_1 \odot \boldsymbol{B}_2$. (This follows from decomposing any PSD matrices as a sum of rank-one PSD matrices.) In particular, $\boldsymbol{G}_{1,t}^{\infty} \succeq C\boldsymbol{I}$ implies $\boldsymbol{G}_{p,t}^{\infty} \succeq C^p\boldsymbol{I}$. It is therefore

sufficient to prove $\sigma_{\min}(\boldsymbol{G}_{1,t}^{\infty}) > 0$, which we do in the next lemma

Lemma A.7.2. *Under the assumptions of Lemma A.7.1, there exist functions $\varphi_t : \mathbb{R}^{t+1} \to \mathbb{R}$, with $\|\varphi_t\|_{L^\infty} \le 1$, $\|\nabla\varphi_t\|_{L^\infty} \le 1$, and an $\varepsilon_0 > 0$ such that the following holds. Letting $\boldsymbol{G}_{1,t}^{\infty}$ denote the Gram matrices associated to $(f_t^{\varepsilon})_{t\ge 0}$ we have $\sigma_{\min}(\boldsymbol{G}_{1,t}^{\infty}) > 0$ for $\varepsilon < \varepsilon_0$.*

Proof. We construct φ_t satisfying the claim inductively in t. The base case is clear: $\mathbf{g}_{1,0}^{\infty} = \mathbb{E}\{f_0(X_0)\}^2 > 0$ for f_0 non vanishing. Assuming we have constructed these functions up to φ_{t-1}, we know that the vector $(X_1^{\varepsilon}, X_2^{\varepsilon}, \dots, X_t^{\varepsilon})$ defined by state evolution (for nonlinearities f_t^{ε}) is a non-degenerate Gaussian.

In order to prove our claim, we need to construct φ_t so that the vector $\{f_s^{\varepsilon}(X_0^{\varepsilon}, \dots, X_s^{\varepsilon})\}_{s\le t}$ has non-degenerate covariance. Since we know already that $\{f_s^{\varepsilon}(X_0^{\varepsilon}, \dots, X_s^{\varepsilon})\}_{s\le t-1}$ is non-degenerate, it is sufficient to show that, for any coefficients $(\alpha_s)_{s\le t}$,

$$\mathbb{E}\left\{\left(f_t^{\varepsilon}(X_0^{\varepsilon}, \dots, X_t^{\varepsilon}) - \sum_{s\le t-1} \alpha_s f_s^{\varepsilon}(X_0^{\varepsilon}, \dots, X_s^{\varepsilon})\right)^2\right\} > 0\,. \tag{A.7.6}$$

It is always possible to choose φ_t so that this is the case. Indeed, the space of functions spanned by f_s^{ε} for $s \le t$ has dimension at most t. Therefore, we can take any $t+1$ linearly independent bounded smooth functions of x_t only, and choose φ_t to be a linear combination of these that is outside the span of $(f_s^{\varepsilon})_{s\le t-1}$. Since non-degenerate Gaussians have full support, this implies the non-degeneracy condition (A.7.6) and therefore the induction claim. □

In preparation for the next part, we argue that when the Gram matrices $\boldsymbol{G}_{1,t}^{\infty}$ are non-degenerate, we can perturb the nonlinearities $(f_t)_{t\ge 0}$ to induce any desired small change in $\boldsymbol{G}_{1,t}^{\infty}$. (Below $\mathfrak{S}_m$ denotes the space of $m \times m$ symmetric matrices.)

Lemma A.7.3. *Under the assumptions of Lemma A.7.1, assume the nonlinearities $(f_t)_{t\ge 0}$ are such that $\boldsymbol{G}_{1,t}^{\infty}$ is non-degenerate. Then there exists finite sets of functions $A_s = \{\varphi_{s,1}, \dots, \varphi_{s,n(s)}\}$ of smooth functions $\varphi_{s,j} : \mathbb{R}^s \to \mathbb{R}$, with $\|\varphi_{s,j}\|_{L^\infty} \le 1$, $\|\nabla\varphi_{s,j}\|_{L^\infty} \le 1$, such that the following is true. For $\boldsymbol{\varepsilon} = (\varepsilon_{s,j})_{j\le n(s), s\le t} \in \mathbb{R}^{n_*}$, $n_* := \sum_{s\le t} n(s)$, consider the nonlinearities $(f_s^{\boldsymbol{\varepsilon}})_{s\le t}$ defined by $f_s^{\boldsymbol{\varepsilon}} = f_s + \sum_{j\le n(s)} \varepsilon_{s,j}\varphi_{s,j}$, and let $\boldsymbol{G}_{1,t}^{\infty}(\boldsymbol{\varepsilon})$ to be the corresponding (asymptotic) Gram matrix. If $G_t : \mathbb{R}^{n_*} \to \mathfrak{S}_t$ is the mapping $G_t : \boldsymbol{\varepsilon} \mapsto \boldsymbol{G}_{1,t}^{\infty}(\boldsymbol{\varepsilon})$, then its derivative $DG_t|_{\boldsymbol{\varepsilon}=\mathbf{0}}$ is surjective.*

Proof. Note that $\mathfrak{S}_t \cong \mathbb{R} \times \mathbb{R}^2 \times \cdots \times \mathbb{R}^t$, by identifying $\mathbf{M} \in \mathfrak{S}_t$ which a list of columns M_{11}, $(M_{1,2}, M_{2,2})$, …, $(M_{j,t})_{j\le t}$. Also $\mathbb{R}^{n_*} \cong \mathbb{R}^{n(1)} \times \cdots \times \mathbb{R}^{n(t)}$, by identifying $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_t)$, $\boldsymbol{\varepsilon}_s = (\varepsilon_{s,j})_{j\le n(s)}$. The matrix $DG_t|_{\boldsymbol{\varepsilon}=\mathbf{0}}$ is block-triangular with respect to this decomposition. By an induction argument, it is therefore sufficient to show that A_t can be constructed so that the last diagonal block $DG_t|_{\boldsymbol{\varepsilon}=\mathbf{0}} : \mathbb{R}^{n(t)} \to \mathbb{R}^t$ is surjective.

Note that G_t is the map that takes as input ε_t, and outputs the last column of the asymptotic Gram matrix corresponding to the nonlinearities $f_1, \dots, f_{t-1}$ and $f_t^{\boldsymbol{\varepsilon}} = f_t + \sum_{j \le n(t)} \varepsilon_{t,j} \varphi_{t,j}$. Since by assumption $\boldsymbol{G}_{t,1}$ is non-degenerate, the functions $f_1, \dots, f_t$ are linearly independent (viewed as vectors in the L^2 space associated to the joint distribution of $(\boldsymbol{X}_s)_{s \le t}$). We can therefore construct functions $(\varphi_{t,s})_{s \le t}$ such that $\mathbb{E}\{\varphi_{t,s}(X_0, \dots, X_t) f_r(X_0, \dots, X_r)\} = 0$ if $r \neq s$, and > 0 if $r = s$. It is then immediate to show that the resulting map $DG_t|_{\boldsymbol{\varepsilon}=\boldsymbol{0}}$ is surjective. □

A.7.2 Condition (ii): Control of $\mathcal{L}_{p,t}$

We are left with the task of showing that –after a small perturbation of the nonlinearities $(f_t)_{t \ge 0}$– condition (ii) of Lemma A.6.1 holds, namely $C^{-1} \le \sigma_{\min}(\mathcal{L}_{p,t}) \le \sigma_{\max}(\mathcal{L}_{p,t}) \le C$ for all $p \le D$, $t \le T$, with high probability. Given the results of the previous section A.7.1, we can assume without loss of generality that $C^{-1} \le \sigma_{\min}(\boldsymbol{G}^{\infty}_{p,t}) \le \sigma_{\max}(\boldsymbol{G}^{\infty}_{p,t}) \le C$ for all p, t. Indeed, if this is not the case, we can modify the nonlinearities as described above, as to satisfy this condition. Also, as before, we can consider a single pair (p, t) since we only are interested in a finite (independent of N) collection of such pairs.

Recall that $\mathcal{L}_{p,t} = \mathbf{1} + (p-1)\mathcal{T}_{p,t}$, and, by Eq. (A.2.7),

$$(\mathcal{T}_{p,t})_{is;jr} = \sum_{r'=0}^{t-1} F_{ir'} F_{js} (\boldsymbol{G}^{-1}_{p-1,t-1})_{r',r} (\boldsymbol{G}_{p-2,t-1})_{r',s} \,, \tag{A.7.7}$$

where $F_{is} = (\boldsymbol{F}_{t-1})_{is} = (\mathbf{f}_s)_i$ for $0 \le s \le t-1$, $\boldsymbol{F}_{t-1} \in \mathbb{R}^{N \times t}$ (for consistency, we index the columns of $\boldsymbol{F}_{t-1}$ as $0, \dots, t-1$). This implies that $\mathcal{T}_{p,t}$ has rank at most t^2 since

$$(\mathcal{T}_{p,t})_{is;jr} = \sum_{a,b=0}^{t-1} (\mathcal{U}_{p,t})_{as;br} F_{ir'} F_{js} \,, \tag{A.7.8}$$

$$(\mathcal{U}_{p,t})_{as;br} := (\boldsymbol{G}^{-1}_{p-1,t-1})_{ra} (\boldsymbol{G}_{p-2,t-1})_{sa} \delta_{b,s} \,, \tag{A.7.9}$$

or, in matrix notation

$$\mathcal{T}_{p,t} = (\boldsymbol{I}_t \otimes \boldsymbol{F}_{t-1}) \mathcal{U}_{p,t} (\boldsymbol{I}_t \otimes \boldsymbol{F}^{\mathsf{T}}_{t-1}) \,. \tag{A.7.10}$$

It follows that the $(N-t)t$ singular values of $\mathcal{L}_{p,t}$ are equal to 1, and the other t^2 singular values coincide with the ones of $\tilde{\mathcal{L}}_{p,t} = \mathbf{1}_{t^2} + (p-1)\tilde{\mathcal{T}}_{p,t}$, where

$$\tilde{\mathcal{T}}_{p,t} = (\boldsymbol{I}_t \otimes \boldsymbol{G}^{-1/2}_{1,t-1}) \mathcal{U}_{p,t} (\boldsymbol{I}_t \otimes \boldsymbol{G}^{-1/2}_{1,t-1}) \,. \tag{A.7.11}$$

Indeed $\tilde{\mathcal{T}}_{p,t}$ is unitarily equivalent to $\mathcal{T}_{p,t}$ (when the latter is restricted to its range), using the fact

that $\boldsymbol{F}_{t-1}^{\mathsf{T}}\boldsymbol{F}_{t-1}/N = \boldsymbol{G}_{1,t-1}$.

We now proceed by induction over the iteration number. Assuming the claim to hold up to iteration $t-1$, we need to to show that (for a suitable perturbation of the nonlinearities) $C^{-1} \le \sigma_{\min}(\tilde{\mathcal{L}}_{p,t}) \le \sigma_{\max}(\tilde{\mathcal{L}}_{p,t}) \le C$ with high probability. By using the induction hypothesis Theorem 40 and Lemma A.6.1 we know that $\boldsymbol{G}_{p,t}$ converges in probability to the deterministic limit $\boldsymbol{G}_{p,t}^{\infty}$ which is non-degenerate. Therefore, it is sufficient to prove that (again, for a suitable perturbation of the nonlinearities) $C^{-1} \le \sigma_{\min}(\tilde{\mathcal{L}}_{p,t}^{\infty}) \le \sigma_{\max}(\tilde{\mathcal{L}}_{p,t}^{\infty}) \le C$, where $\tilde{\mathcal{L}}_{p,t}^{\infty} = \mathbf{1}_{t^2} + (p-1)\tilde{\mathcal{T}}_{p,t}^{\infty}$, and $\tilde{\mathcal{T}}_{p,t}^{\infty}$ is obtained from $\tilde{\mathcal{T}}_{p,t}$ by replacing $\boldsymbol{G}_{k,s}$ by its asymptotic version $\boldsymbol{G}_{k,s}^{\infty}$ everywhere. Since the resulting matrix $\tilde{\mathcal{L}}_{p,t}^{\infty}$ is finite (and of dimension independent of N), it is sufficient to prove that $\sigma_{\min}(\tilde{\mathcal{L}}_{p,t}^{\infty}) > 0$. Since $\boldsymbol{G}_{1,t-1}^{\infty}$ is non-degenerate, it is sufficient to prove $\sigma_{\min}(\mathcal{W}_{p,t}^{\infty}) > 0$, where

$$\mathcal{W}_{p,t}^{\infty} := \boldsymbol{I}_t \otimes \boldsymbol{G}_{1,t-1}^{\infty} + (p-1)\mathcal{U}_{p,t}^{\infty}\,, \tag{A.7.12}$$

$$(\mathcal{U}_{p,t}^{\infty})_{as;br} := ((\boldsymbol{G}_{p-1,t-1}^{\infty})^{-1})_{ra}(\boldsymbol{G}_{p-2,t-1}^{\infty})_{sa}\delta_{b,s}\,. \tag{A.7.13}$$

In order to prove the desired non-degeneracy bound for $\mathcal{W}_{p,t}^{\infty}$, it is useful to introduce a piece of terminology.

Definition A.7.4. *We say a subset $S \subseteq \mathbb{R}^d$ is* locally full *if for any open set $U \subseteq R^d$ with $U \cap S \neq \emptyset$ we have $\lambda(U \cap S) > 0$ (with λ denoting the Lebesgue measure on $\mathbb{R}^d$).*

For instance, a full-dimensional convex set is locally full.

Lemma A.7.5. *Let $K \subseteq \mathbb{R}^d$ be locally full and $R : \mathbb{R}^d \to \mathbb{R}$ a rational function which is not identically zero or infinity. For any $\varepsilon > 0$ and $\boldsymbol{x} \in K$ there is $\boldsymbol{x}' \in K$ with $\|\boldsymbol{x} - \boldsymbol{x}'\| \le \varepsilon$ and $R(\boldsymbol{x}') \notin \{0, \pm\infty\}$.*

Proof. Simply recall that any nontrivial polynomial vanishes on a measure zero set. □

We are now in position to show that the nonlinearities $(f_s)_{0\le s\le t}$ can be modified so that the resulting matrix $\mathcal{W}_{p,t}^{\infty}$ has $\sigma_{\min}(\mathcal{W}_{p,t}^{\infty}) > 0$, thus completing the proof.

Lemma A.7.6. *Under the assumptions of Lemma A.7.1, further assume the nonlinearities $(f_s)_{s\ge 0}$ to be such that $\sigma_{\min}(\boldsymbol{G}_{p,t}^{\infty}) > 0$ for all $p \le D$, $t \le T$. Then, for any $\varepsilon > 0$ there exist functions $\varphi_s : \mathbb{R}^{s+1} \to \mathbb{R}$, with $\|\varphi_s\|_{L^\infty} \le 1$, $\|\nabla\varphi_s\|_{L^\infty} \le 1$, such that the following holds.*

Let $\mathcal{W}_{p,t}^{\infty}(\varepsilon)$ the matrix defined in Eqs. (A.7.12), (A.7.13)*, for nonlinearities $f_s^{\varepsilon} = f_s + \varepsilon\varphi_s$, $s \le t$. Then, for any $p \le D$ and $t \le T$, $\sigma_{\min}(\mathcal{W}_{p,t}^{\infty}(\varepsilon)) > 0$.*

Proof. Notice that $\mathcal{W}_{p,t}^{\infty}$ is a function of the matrix $\boldsymbol{G}_{1,t}^{\infty}$ (the matrices $\boldsymbol{G}_{p,t}^{\infty}$ being themselves Hadamard powers of $\boldsymbol{G}_{1,t}$). With a slight abuse of notation, we will write $\mathcal{W}_{p,t}^{\infty} = \mathcal{W}_{p,t}^{\infty}(\boldsymbol{G}_{1,t}^{\infty})$.

Define $R : \mathfrak{S}_{t+1} \to \mathbb{R}$ to be the function that takes as input a $t+1 \times t+1$ symmetric matrix $\boldsymbol{G}$ and outputs

$$R(\boldsymbol{G}) \equiv \prod_{p \le D} \det(\mathcal{W}_{p,t}^{\infty}(\boldsymbol{G}))\,. \tag{A.7.14}$$

By checking Eqs. (A.7.12), (A.7.12), we see that this is a rational function on $\mathfrak{S}_t \cong \mathbb{R}^{\binom{t+1}{2}}$. Further, it is not identically zero or infinity, as it can be checked by computing $\mathcal{W}_{p,t}^{\infty}(\boldsymbol{I})$. Applying Lemma A.7.5 to the set of PSD matrices, which is locally full in $\mathbb{R}^{\binom{t}{2}}$, and the rational function R, we obtain that, for any $\xi > 0$, there exists $\boldsymbol{G}_* \succeq \boldsymbol{0}$, with $\|\boldsymbol{G}_* - \boldsymbol{G}_{p,t}^{\infty}\|_F \le \xi$, and $R(\boldsymbol{G}_*) \notin \{0, \pm\infty\}$, which implies $\sigma_{\min}(\mathcal{W}_{p,t}^{\infty}(\boldsymbol{G}_*)) > 0$ for all $p \le D$.

Finally, using Lemma A.7.3 and the implicit function theorem, we conclude that we can find a perturbation $(\varphi_s)_{s \le t}$, and $\varepsilon_0 > 0$ such that $\boldsymbol{G}_{1,t}(\varepsilon) = \boldsymbol{G}_*$. By taking ξ sufficiently small, we can ensure that ε can also be arbitrarily small. □

A.8 Extension to the case $D = \infty$

Here we extend the state evolution result proved for finite mixtures to the general case where ξ has infinitely many components. The proof proceeds by induction over the number of iterations, and is similar to previous arguments. Let us write $\tilde{\xi}(x) := \sum_{k \le D} c_k^2 x^k$ while $\xi(x) = \sum_{k=2}^{\infty} c_k^2 x^k$. Denote by $(\tilde{X}^0, \cdots, \tilde{X}^\ell)$ the state evolution Gaussian process corresponding to $\tilde{\xi}$, and $(X^0, \cdots, X^\ell)$ the one based on ξ. First, using the fact that f_ℓ is Lipschitz, it is easy to show by induction over ℓ that there exists a coupling such that $\mathbb{E}[(\tilde{X}^\ell - X^\ell)^2] = o_D(1)$ (throughout this section, $o_D(1)$ is a term independent of N that vanishes as $D \to \infty$). We deduce from this that $\tilde{d}_{\ell,j} - d_{\ell,j} = o_D(1)$ for all ℓ, j. (Here, $\tilde{d}_{\ell,j}$ is defined similarly to $d_{\ell,j}$, based on the mixture $\tilde{\xi}$.)

Next we show that the AMP iterates are close. Let $\tilde{\boldsymbol{z}}^0, \cdots \tilde{\boldsymbol{z}}^\ell$ be the AMP iterates based on $\tilde{\xi}$ and $\boldsymbol{z}^0, \cdots \boldsymbol{z}^\ell$ those based on ξ. Let $\tilde{\boldsymbol{z}}^0 = \mathbf{z}^0 = \boldsymbol{0}$ and assume $\lim_{D\to\infty} \text{p-lim}_{N\to\infty} \|\tilde{\boldsymbol{z}}^j - \boldsymbol{z}^j\|_N = 0$ for all $j \le \ell$. Further let $\tilde{\mathbf{f}}_\ell = f_\ell(\tilde{\boldsymbol{z}}^0, \cdots, \tilde{\boldsymbol{z}}^\ell)$. Then

$$\big\|\tilde{\boldsymbol{z}}^{\ell+1} - \boldsymbol{z}^{\ell+1}\big\|_N \le \Big\|\sum_{p=2}^{D} \frac{c_p}{p!} \boldsymbol{W}^{(p)}\{\tilde{\mathbf{f}}_\ell\} - \sum_{p=2}^{\infty} \frac{c_p}{p!} \boldsymbol{W}^{(p)}\{\mathbf{f}_\ell\}\Big\|_N + \Big\|\sum_{j=0}^{\ell} \tilde{d}_{\ell,j}\tilde{\mathbf{f}}_{j-1} - d_{\ell,j}\mathbf{f}_j\Big\|_N \tag{A.8.1}$$

$$=: E_1 + E_2. \tag{A.8.2}$$

We have

$$E_1 \le \sum_{p \ge D+1} \frac{c_p}{p!} \big\|\boldsymbol{W}^{(p)}\{\tilde{\mathbf{f}}_\ell\}\big\|_N + \sum_{p=2}^{\infty} \frac{c_p}{p!} \big\|\boldsymbol{W}^{(p)}\{\mathbf{f}_\ell\} - \boldsymbol{W}^{(p)}\{\tilde{\mathbf{f}}_\ell\}\big\|_N.$$

The first term in the above is bounded by

$$\sum_{p\geq D+1} \frac{c_p}{p!} N^{(p-2)/2} \|\boldsymbol{W}^{(p)}\|_{\mathrm{op}} \cdot \|\tilde{\mathbf{f}}_\ell\|_N^{p-1}.$$

Using Theorem 40, $\|\tilde{\mathbf{f}}_\ell\|_N \leq C$ with high probability. Lemma A.2.1 then implies that the above is $o_D(1)$ with high probability. Next, the second term in E_1 is similarly bounded by

$$\sum_{p=2}^{\infty} \frac{c_p}{p!} N^{(p-2)/2} \|\boldsymbol{W}^{(p)}\|_{\mathrm{op}} \cdot (\|\tilde{\mathbf{f}}_\ell\|_N + \mathbf{f}_\ell\|_N)^{p-2} \|\tilde{\mathbf{f}}_\ell - \mathbf{f}_\ell\|_N.$$

Since f_ℓ is Lipschitz, and using the induction hypothesis, similar considerations show that this term converges to zero in probability as $N \to \infty$. Next,

$$E_2 \leq \sum_{j=0}^{\ell} (\tilde{d}_{\ell,j} - d_{\ell,j}) \|\tilde{\mathbf{f}}_{j-1}\|_N + \sum_{j=0}^{\ell} |d_{\ell,j}| \|\tilde{\mathbf{f}}_{j-1} - \mathbf{f}_{j-1}\|_N \overset{p}{\simeq} o_D(1).$$

This implies

$$\lim_{D\to\infty} \operatorname*{p-lim}_{N\to\infty} \left\|\tilde{\boldsymbol{z}}^{\ell+1} - \boldsymbol{z}^{\ell+1}\right\|_N = 0,$$

which concludes the inductive argument. Finally, for ψ a pseudo-Lipschitz function, we have

$$\begin{aligned}
\operatorname*{p-lim}_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \psi(\boldsymbol{z}^0, \cdots, \boldsymbol{z}^\ell) &= \operatorname*{p-lim}_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \psi(\tilde{\boldsymbol{z}}^0, \cdots, \tilde{\boldsymbol{z}}^\ell) + o_D(1) && \text{(A.8.3)} \\
&= \mathbb{E}[\psi(\tilde{X}^0, \cdots, \tilde{X}^\ell)] + o_D(1) && \text{(A.8.4)} \\
&= \mathbb{E}[\psi(X^0, \cdots, X^\ell)] + o_D(1). && \text{(A.8.5)}
\end{aligned}$$

This concludes our proof of state evolution, Proposition 3.6.1.

Appendix B

Properties of the Parisi PDE and Variational Problem

In this Chapter we prove several useful properties of the extended variational principle $\inf_{\gamma\in\mathscr{L}}\mathsf{P}(\gamma)$. A first set of properties concerns the solution of the Parisi PDE (4.1.1) for $\gamma\in\mathscr{L}$. Most of these are generalizations of results obtained in [JT16] for $\gamma\in\mathscr{U}$ bounded (hence, with finite total variation over $[0,1]$). We will refer to the proofs of [JT16] whenever they can be adapted without significant changes. In several cases, new arguments are required, e.g. in the regularity result of Lemma B.1.3, in the first variation formula of Proposition B.2.1 and elsewhere. The second set of technical results concerns properties of the minimizers. These are of course entirely new because the minimizer is —in general— outside $\mathscr{U}$. Finally in the third section we establish several Lemmas used in Chapter 4.

We consider the function space $\mathscr{L}$ from (4.1.5), endowed with the weighted L^1 distance $\|\gamma_1-\gamma_2\|_{1,\xi''}=\int_0^1\xi''(t)|\gamma_1(t)-\gamma_2(t)|\mathrm{d}t$. We will write $\gamma_n\xrightarrow{L^1_\xi}\gamma$, whenever $\|\gamma_n-\gamma\|_{1,\xi''}\to 0$ as $n\to\infty$. We recall the space of piecewise constant functions

$$\mathsf{SF}_+=\Big\{g=\sum_{i=1}^m a_i\mathbb{I}_{[t_{i-1},t_i)}:\ \ 0=t_0<t_1<\cdots<t_m=1, a_i\geq 0, m\in\mathbb{N}\Big\}. \tag{B.0.1}$$

We study the PDE (4.1.1), with a slightly more general initial condition

$$\begin{aligned}\partial_t\Phi(t,x)+\frac{1}{2}\xi''(t)\Big(\partial_x^2\Phi(t,x)+\gamma(t)(\partial_x\Phi(t,x))^2\Big)=0\,,\\ \Phi(1,x)=f_0(x)\,.\end{aligned} \tag{B.0.2}$$

Throughout we assume f_0 to be convex, continuous, non-negative, with $f_0(-x)=f_0(x)\geq 0$, and

differentiable for $x \neq 0$, with $0 \leq f_0'(x) \leq 1$ for all $x > 0$. We will write $f_0'(x)$ for the weak derivative of f_0 (the right and left derivatives exist but are potentially different at $x = 0$). Associated to the above PDE, we consider the following stochastic differential equation driven by Brownian motion $(B_t)_{t\geq 0}$:

$$\mathrm{d}X_t = \xi''(t)\gamma(t)\partial_x\Phi(t, X_t)\,\mathrm{d}t + \sqrt{\xi''(t)}\,\mathrm{d}B_t\,, \qquad X_0 = h\,. \tag{B.0.3}$$

In the following we will also write Φ_x, Φ_{xx} and so on for the partial derivatives of Φ, and Φ_γ whenever we want to emphasize the dependence of Φ on γ. We write $\partial_t^{\pm}\Phi$ for the left and right derivatives of Φ.

B.1 Existence, Uniqueness, and Regularity

We first collect a few properties of $\Phi(t, x)$ when $\gamma \in \mathsf{SF}_+$.

Proposition B.1.1. *(a) For any $\gamma \in \mathsf{SF}_+$ the solution $\Phi : [0,1] \times \mathbb{R} \to \mathbb{R}$ of Eq. (B.0.2) exists uniquely in the classical sense and is smooth for $t \in [0,1)$. Namely, for any $j > 0$, $\|\partial_x^j\Phi\|_{L^\infty([0,1-\varepsilon)\times\mathbb{R})} \leq C(\gamma,\varepsilon)$, and $\|\partial_t^{\pm}\partial_x^j\Phi\|_{L^\infty([0,1-\varepsilon)\times\mathbb{R})} \leq C(\gamma,\varepsilon)$, with $\partial_t^+\partial_x^j\Phi(t,x) = \partial_t^-\partial_x^j\Phi(t,x)$ whenever t is a continuity point of γ.*

(b) For any $\gamma \in \mathsf{SF}_+$ the solution Φ of Eq. (B.0.2) is such that $x \mapsto \partial_x\Phi(t,\cdot)$ is non-decreasing for all $t \in [0,1]$, with $|\partial_x\Phi(t,x)| \leq 1$ for all $x \in \mathbb{R}$.

(c) If $\gamma_1, \gamma_2 \in \mathsf{SF}_+$ and Φ_{γ_1}, Φ_{γ_2} are the corresponding solutions, then

$$\|\Phi_{\gamma_1} - \Phi_{\gamma_2}\|_\infty \leq \|\gamma_1 - \gamma_2\|_{1,\xi''}\,.$$

Proof. Point (a) follows from the Cole-Hopf representation which allows us to write an explicit form of the solution for $\gamma \in \mathsf{SF}_+$ [Gue01, AC17b]. This solution is C^∞ except (possibly) when $t \in \{t_1, \ldots, t_{m-1}\}$, the set of discontinuity points of γ. As a consequence of point (a), the SDE (B.3.1) is well defined, with unique strong solution on $[0,1]$. Further, Φ satisfies the following representation, for $\gamma \in \mathsf{SF}_+$ [JT16]:

$$\partial_x\Phi(t,x) = \mathbb{E}\left[f_0'(X_1)|X_t = x\right].$$

Since $\|f_0'\|_\infty \leq 1$, this implies $|\partial_x\Phi(t,x)| \leq 1$. The non-decreasing property also follows again by the Cole-Hopf representation.

Finally, point (c) is identical to Lemma 14 in [JT16] (the assumption that γ is non-decreasing is never used there). □

As a consequence of Proposition B.1.1, we can define Φ_γ by continuity for any $\gamma \in \mathscr{L}$. Namely, we construct a sequence $\gamma_n \in \mathsf{SF}_+$, $\gamma_n \xrightarrow{L^1_\xi} \gamma$ and

$$\Phi_\gamma(t,x) = \lim_{n\to\infty} \Phi_{\gamma_n}(t,x)\,.$$

Lemma B.1.2. *For any $\gamma \in \mathscr{L}$, Φ_γ constructed above is such that $\partial_x\Phi_\gamma$ exists in weak sense, is non-decreasing, and $|\partial_x\Phi_\gamma(t,x)| \le 1$ for all $t \in [0,1]$, $x \in \mathbb{R}$. Further, if $\gamma_n \in \mathsf{SF}_+$, $\gamma_n \xrightarrow{L^1_\xi} \gamma$, for any $t \in [0,1]$, we have $\partial_x\Phi_{\gamma_n}(t,x) \to \partial_x\Phi_\gamma(t,x)$ for almost every x.*

Finally, $\Phi = \Phi_\gamma$ is a weak solution of the PDE (B.0.2)*. Namely, for any $g \in C_c^\infty((0,1]\times\mathbb{R})$, we have*

$$0 = \int_{(0,1]}\int_{\mathbb{R}} \Big\{ -\Phi\partial_t h + \frac{1}{2}\xi''(t)\Big(\Phi\partial_x^2 h + \gamma(t)(\partial_x\Phi)^2 h\Big)\Big\}\,\mathrm{d}x\,\mathrm{d}t + \int_{\mathbb{R}} \Phi(1,x)\,f_0(x)\,\mathrm{d}x\,. \tag{B.1.1}$$

Proof. Since $\Phi_\gamma(t,\,\cdot\,)$ is the uniform limit of convex 1-Lipschitz functions, it is also convex 1-Lipschitz. Hence its weak derivative exists, is non-decreasing and is bounded as claimed. The claim $\partial_x\Phi_{\gamma_n}(t,x) \to \partial_x\Phi_\gamma(t,x)$ follows by dominated convergence.

In order to show that Φ is a weak solution, let $\Phi_n = \Phi_{\gamma_n}$ for $\gamma_n \in \mathsf{SF}_+$, $\gamma_n \xrightarrow{L^1_\xi} \gamma$ (hence $\|\Phi_n - \Phi\|_\infty \to 0$). Since Φ_n is a classical solution corresponding to γ_n, we have

$$0 = \int_{(0,1]}\int_{\mathbb{R}} \Big\{ -\Phi_n\partial_t h + \frac{1}{2}\xi''(t)\Big(\Phi_n\partial_x^2 h + \gamma_n(t)(\partial_x\Phi_n)^2 h\Big)\Big\}\,\mathrm{d}x\,\mathrm{d}t + \int_{\mathbb{R}} \Phi_n(1,x)\,f_0(x)\,\mathrm{d}x\,.$$

Letting Δ denote the right-hand side of Eq. (B.1.1), we have (since $\Phi_n(1,x) = \Phi(1,x)$ is independent of n)

$$\begin{aligned}\Delta = &\int_{(0,1]}\int_{\mathbb{R}} \Big\{ (\Phi_n - \Phi)\partial_t h - \frac{1}{2}\xi''(t)(\Phi_n - \Phi)\partial_x^2 h\Big\}\,\mathrm{d}x\,\mathrm{d}t\\ &- \int_{(0,1]}\int_{\mathbb{R}} \frac{1}{2}\xi''(t)\Big(\gamma_n(t)(\partial_x\Phi_n)^2 - \gamma(t)(\partial_x\Phi)^2\Big)h\,\mathrm{d}x\,\mathrm{d}t\,.\end{aligned}$$

The first term vanishes as $n \to \infty$ by dominated convergence. For the second term, by the bound on $\partial_x\Phi$, $\partial_x\Phi_n$, we have

$$|\Delta| \le \frac{1}{2}\int_{(0,1]}\int_{\mathbb{R}} \xi''(t)\,|\gamma_n(t) - \gamma(t)|\,|h|\mathrm{d}x\,\mathrm{d}t + \frac{1}{2}\int_{(0,1]}\int_{\mathbb{R}} \xi''\gamma(t)\,\big|(\partial_x\Phi_n)^2 - (\partial_x\Phi)^2\big|\,|h|\mathrm{d}x\,\mathrm{d}t\,.$$

The first term vanishes as $n \to \infty$ since $\gamma_n \xrightarrow{L^1_\xi} \gamma$, and the second vanishes by dominated convergence, using the fact that $\|\xi''\gamma\|_1 < \infty$. □

Lemma B.1.3. *For $\gamma \in \mathscr{L}$ and any $t \in [0,1)$, the second derivative $\partial_x^2\Phi(t,\,\cdot\,)$ exists in the weak*

sense, with $\sup_{0\le t\le 1-\varepsilon}\|\partial_x^2\Phi(t,\,\cdot\,)\|_{L^2(\mathbb{R})}<\infty$ *for any* $\varepsilon>0$.

Proof. Following [JT16], it is useful to introduce the the smooth time change $\theta(t)=(\xi'(1)-\xi'(t))/2$, and define $u:[0,\theta_M]\times\mathbb{R}$, $\theta_M=\xi'(1)/2$, via $u(\theta(t),x)=\Phi(t,x)$. By a simple change of variables, u is a weak solution of the PDE

$$\partial_\theta u-\Delta u=m(\theta)u_x^2\,,\quad u(0,x)=f_0(x)\,,$$

where $m(s)=\gamma(\theta^{-1}(s))$. The desired claim is implied by showing that the partial derivative $\partial_x^2 u$ exists in weak sense and is bounded uniformly over $\theta>\varepsilon$ (for any $\varepsilon>0$).

Again, as in [JT16] the fact that u is a weak solution implies the Duhamel principle

$$\begin{aligned}u(\theta)&=G_\theta * f_0+\int_0^\theta m(s)\,G_{\theta-s}*u_x(s)^2\mathrm{d}s\,,\\ G_t(x)&\equiv\frac{1}{\sqrt{4\pi t}}e^{-x^2/4t}\,.\end{aligned}\tag{B.1.2}$$

(Here $*$ denotes convolution and this equation is to be interpreted in weak sense, namely, for any $g\in C_c^\infty(\mathbb{R})$, $\int g(x)u(\theta,x)\,\mathrm{d}x$ is given by the convolution with g of the right hand side.) Note that by Lemma B.1.2, $x\mapsto u_x(s,x)^2$ is bounded between 0 and 1, non-increasing in $(-\infty,0]$, non-decreasing in $[0,\infty)$ and symmetric (the value at $x=0$ is immaterial). Hence, there exists a measure ν_s on $[0,\infty)$, with total mass $\nu_s([0,\infty))\le 1$, such that

$$u_x(s,x)^2=\nu_s([0,x))\,\mathbb{I}_{x>0}+\nu_s([0,-x))\,\mathbb{I}_{x<0}\,.$$

We then obtain, from Eq. (B.1.2)

$$u_{xx}(\theta)=G'_\theta * f'_0+\int_0^\theta m(s)\int_{\mathbb{R}_{\ge 0}}[G'_{\theta-s}(\,\cdot\,-x)+G'_{\theta-s}(\,\cdot\,+x)]\mathrm{d}\nu_s(x)\,\mathrm{d}s\,.\tag{B.1.3}$$

The claim follows by showing that each of the two terms on the right hand side of Eq. (B.1.3) is a well defined function, bounded in $L^2(\mathbb{R})$. For the first term, notice that f'_0 is bounded and non-decreasing. Hence there exists a measure ω_0 on $\mathbb{R}$ with $\omega_0(\mathbb{R})\le 2$, such that $G'_\theta * f'_0=G_\theta *\mathrm{d}\omega_0$, whence

$$\|G'_\theta * f'_0\|_2=\left\|\int G_\theta(\,\cdot\,-x)\,\mathrm{d}\omega_0(x)\right\|_2\le 2\|G_\theta\|_2\le\frac{C}{\theta^{1/4}}\,,$$

where the upper bound follows from Jensen's inequality. The second term on the right-hand side of

(B.1.3) can be treated analogously. Denoting it by $w(\theta)$, we have, again by Jensen with $\theta = \theta(1-\varepsilon)$,

$$\begin{aligned}\|w(\theta)\|_2 &\le \int_0^\theta m(s) \int_{\mathbb{R}_{\ge 0}} \|G'_{\theta-s}(\,\cdot\, - x) + G'_{\theta-s}(\,\cdot\, + x)\|_2 \mathrm{d}\nu_s(x)\,\mathrm{d}s \\ &\le C\int_0^\theta m(s)\,\frac{1}{(\theta-s)^{3/4}}\,\mathrm{d}s \le C'\int_{1-\varepsilon}^1 \frac{\xi''\gamma(s)}{(\xi'(s)-\xi'(1-\varepsilon))^{3/4}}\,\mathrm{d}s\,,\end{aligned}$$

where the second inequality follows by $\|G'_t\|_2 \le C\,t^{-3/4}$. Decomposing the last integral, we get

$$\begin{aligned}\|w(\theta)\|_2 &\le C'\int_{1-\varepsilon}^{1-\varepsilon/2} \frac{\xi''\gamma(s)}{(\xi'(s)-\xi'(1-\varepsilon))^{3/4}}\,\mathrm{d}s + C'\int_{1-\varepsilon/2}^1 \frac{\xi''\gamma(s)}{(\xi'(s)-\xi'(1-\varepsilon))^{3/4}}\,\mathrm{d}s \\ &\le C'\xi''\gamma(1-\varepsilon/2)\int_{1-\varepsilon}^{1-\varepsilon/2} \frac{1}{(\xi'(s)-\xi'(1-\varepsilon))^{3/4}}\,\mathrm{d}s \\ &\quad + \frac{C'}{(\xi'(1-\varepsilon/2)-\xi'(1-\varepsilon))^{3/4}}\int_{1-\varepsilon/2}^1 \xi''\gamma(s)\,\mathrm{d}s \\ &\le C''\|\xi''\gamma\|_{\mathrm{TV}[0,1-\varepsilon/2]} + C''\varepsilon^{-3/4}\,\|\xi''\gamma\|_1\,.\end{aligned}$$

The last expression is bounded by some $C(\varepsilon) < \infty$ since $\gamma \in \mathscr{L}$. □

Lemma B.1.4. *For any $\gamma \in \mathscr{L}$, the solution $\Phi = \Phi_\gamma$ constructed above is continuous on $[0,1]\times\mathbb{R}$, and further satisfies the following regularity properties for any $\varepsilon > 0$*

(*a*) $\partial_x^j \Phi \in L^\infty([0,1-\varepsilon]; L^2(\mathbb{R}) \cap L^\infty(\mathbb{R}))$ *for* $j \ge 2$.

(*b*) $\partial_t \Phi \in L^\infty([0,1]\times\mathbb{R})$ *and* $\partial_t\partial_x^j\Phi \in L^\infty([0,1-\varepsilon]; L^2(\mathbb{R}) \cap L^\infty(\mathbb{R}))$ *for* $j \ge 1$.

Proof. Continuity follows since Φ_γ is the uniform limit of continuous functions. Point (*a*) and (*b*) follows from the same proof as Lemma 10 in [JT16], applied to the PDE (B.0.2) with boundary condition at $t = 1-\varepsilon$, whereby we use Lemma B.1.3 to initiate the bootstrap procedure. □

As a consequence of the stated regularity properties of Φ, we can solve the SDE (B.3.1).

Lemma B.1.5. *For any $\gamma \in \mathscr{L}$, let $\Phi = \Phi_\gamma$ be the PDE solution defined above. Then the stochastic differential equation* (B.3.1) *has unique strong solution on $(X_t)_{t\in[0,1]}$, which is almost surely continuous. Further, for any $t \in [0,1]$*

$$\partial_x\Phi(t, X_t) = \int_0^t \sqrt{\xi''(s)}\,\partial_x^2\Phi(s, X_s)\,\mathrm{d}B_s\,. \tag{B.1.4}$$

Proof. Existence and uniqueness for $t \in [0, 1-\varepsilon)$ follow because $\partial_x\Phi(t,\,\cdot\,)$ is Lipschitz continuous and $\xi''\gamma$ is bounded on such interval (see, e.g., [Oks13, Chapter 5].) By letting $\varepsilon \downarrow 0$, we obtain

existence and uniqueness on $[0,1)$. Further X_t can be extended at $t=1$, letting

$$X_1 = \int_0^1 \xi''(t)\gamma(t)\partial_x\Phi(t,X_t)\mathrm{d}t + \int_0^1 \sqrt{\xi''(t)}\mathrm{d}B_t\,.$$

It is easy to check that this extension is almost surely continuous at $t=1$, since

$$\big|X_1 - X_t\big| \le \int_t^1 \xi''\gamma(s)\mathrm{d}s + \int_t^1 \sqrt{\xi''(t)}\mathrm{d}B_t\,.$$

The first integral vanishes as $t\to 1$ since $\int_0^1 \xi''\gamma(t)\,\mathrm{d}t<\infty$, while the second vanishes by continuity of the Brownian motion.

Next notice that, since $\Phi_x = \partial_x\Phi$ smooth in space and weakly differentiable in time for $t\in[0,1)$ by Lemma B.1.4, it is a weak solution of

$$\partial_t\Phi_x(t,x) + \frac{1}{2}\xi''(t)\Big(\partial_x^2\Phi_x(t,x) + \gamma(t)\partial_x(\Phi_x(t,x))^2\Big) = 0\,.$$

More precisely, for any $x\in\mathbb{R}$ and any $g\in C_c((0,1))$, we have

$$\int\left\{g(t)\partial_t\Phi_x(t,x) + \frac{\xi''(t)}{2}g(t)\left(\partial_x^2\Phi_x(t,x)+\gamma(t)\partial_x(\Phi_x(t,x))^2\right)\right\}\mathrm{d}t = 0\,. \tag{B.1.5}$$

Equation (B.1.4) is then obtained by Itô's formula (see Proposition 22 in [JT16])

$$\begin{aligned}\partial_x\Phi(t,X_t) = &\int_0^t \sqrt{\xi''(s)}\,\partial_x^2\Phi(s,X_s)\,\mathrm{d}B_s\\ &+\int_0^t\left(\partial_s\Phi_x(s,X_s) + \frac{1}{2}\xi''(s)\Big(\partial_x^2\Phi_x(s,X_s)+\gamma(s)\partial_x(\Phi_x(s,X_s))^2\Big)\right\}\mathrm{d}s\,,\end{aligned}$$

The second term vanishes by Eq. (B.1.5). □

Corollary B.1.6. *For any $\gamma\in\mathscr{L}$ and any $t\in[0,1)$,*

$$\mathbb{E}\left[\partial_x\Phi_\gamma(t,X_t)^2\right] = \int_0^t \xi''(s)\;\mathbb{E}\left[\big(\partial_{xx}\Phi_\gamma(s,X_s)\big)^2\right]\,\mathrm{d}s\,.$$

Proof. This follows from Lemma B.1.5, using the regularity properties of Lemma B.1.4. □

Lemma B.1.7. *For any $\gamma\in\mathscr{L}$, the values*

$$\mathbb{E}\left[\partial_x\Phi_\gamma(t,X_t)^2\right], \qquad \mathbb{E}\left[\partial_{xx}\Phi_\gamma(t,X_t)^2\right]$$

are continuous functions of $t\in[0,1)$.

Lemma B.1.8. *The function $\mathsf{P} = \mathsf{P}_{\xi,\mathcal{L}_h}$ is strictly convex on $\mathscr{L}$.*

Proof. The proof is exactly the same as [CHL18, Lemma 5] which shows strict convexity on $\mathscr{U}$. □

B.2 Properties of the Minimizing Order Parameter

We now compute the first variation of the Parisi functional.

Proposition B.2.1. *Let $\gamma \in \mathscr{L}$, and $\delta : [0,1) \to \mathbb{R}$ be such that $\|\xi''\delta\|_{TV[0,t]} < \infty$ for all $t \in [0,1)$, $\|\xi''\delta\|_1 < \infty$, and $\delta(t) = 0$ for $t \in (1-\varepsilon, 1]$, $\varepsilon > 0$. Further assume that $\gamma + s\delta \geq 0$ for all $s \in [0, s_0)$ for some positive s_0. Then*

$$\frac{\mathrm{d}\mathsf{P}}{\mathrm{d}s}(\gamma + s\delta)\Big|_{s=0+} = \frac{1}{2}\int_0^1 \xi''(t)\delta(t)\big(\mathbb{E}\left[\partial_x \Phi_\gamma(t, X_t)^2\right] - t\big)\,\mathrm{d}t\,. \tag{B.2.1}$$

Proof. Let $\gamma^s \equiv \gamma + s\delta$, $s \in [0, \varepsilon)$, and denote by Φ_s the corresponding solution of the Parisi PDE. Following the proof of Lemma 14 in [JT16], we get

$$\Phi_s(0,0) - \Phi_0(0,0) = \frac{s}{2}\int_0^1 \xi''(t)\delta(t)\,\mathbb{E}\{\partial_x \Phi_0(t, Y_t^s)^2\}\,\mathrm{d}t\,, \tag{B.2.2}$$

where Y_t^s is the solution of the SDE

$$\mathrm{d}Y_t^s = \frac{1}{2}\xi''(t)\gamma^s(t)\big[\partial_x \Phi_0(t, Y_t^s) + \partial_x \Phi_s(t, Y_t^s)\big]\,\mathrm{d}t + \sqrt{\xi''(t)}\,\mathrm{d}B_t\,, \qquad Y_0^s = 0\,. \tag{B.2.3}$$

We also obtain (by the same argument as in [JT16, Lemma 14], using Lemma B.1.4, and noting that $\delta(t) = 0$ for $t > 1 - \varepsilon$ and $\xi''\gamma$ is bounded on $[0, 1-\varepsilon)$)

$$\|\partial_x \Phi_s - \partial_x \Phi_0\|_\infty \leq C(\varepsilon, \gamma)\|\xi''\delta\|_1 \cdot s\,. \tag{B.2.4}$$

Taking the difference between this Eqs. (B.2.3) and (B.3.1), we get, for $t \in [0, 1-\varepsilon_0)$

$$\begin{aligned}
|Y_t^s - X_t| \leq & C\int_0^t \xi''(u)|\gamma^s(u) - \gamma(u)|\mathrm{d}u + C\int_0^t \xi''\gamma(u)\big|\partial_x \Phi_0(u, Y_u^s) - \partial_x \Phi_s(u, Y_u^s)\big|\,\mathrm{d}u \\
& + C\int_0^t \xi''\gamma(u)\big|\partial_x \Phi_0(u, X_u) - \partial_x \Phi_0(u, Y_u^s)\big|\,\mathrm{d}u \\
& \leq C\|\xi''(\gamma^s - \gamma^0)\|_1 + C(\varepsilon, \gamma)\|\xi''(\gamma^s - \gamma^0)\|_1\|\xi''\gamma\|_1 \\
& \quad + C(\varepsilon_0)\int_0^t \xi''\gamma(u)\big|Y_u^s - X_u\big|\,\mathrm{d}u\,.
\end{aligned}$$

In the second inequality we used Eq. (B.2.4), and the fact that $\partial_x^2 \Phi$ is bounded for $t \in [0, 1-\varepsilon_0)$,

see Lemma B.1.4. Since $\xi''\gamma(u) \le \|\xi''\gamma\|_{\mathrm{TV}[0,1-\varepsilon_0]}$ for $u \in [0, 1-\varepsilon_0)$, we finally obtain

$$|Y_t^s - X_t| \le C(\gamma,\varepsilon)\, s\|\xi''\delta\|_1 + C(\gamma,\varepsilon_0)\int_0^t \left|Y_u^s - X_u\right| \mathrm{d}u\,.$$

Therefore, we conclude by Gronwall lemma that

$$\sup_{t\le 1-\varepsilon_0} \left|Y_t^s - X_t\right| \le C(\varepsilon,\varepsilon_0,\gamma)\|\xi''\delta\|_1\, s$$

Using this in Eq. (B.2.2), together with the fact that $\partial_x\Phi_0$ is bounded and Lipschitz, and $\delta(t) = 0$ for $t > 1-\varepsilon$, we get

$$\Phi_s(0,0) - \Phi_0(0,0) = \frac{s}{2}\int_0^1 \xi''(t)\delta(t)\,\mathbb{E}\{\partial_x\Phi_0(t,X_t)^2\}\,\mathrm{d}t + O(s^2)\,,$$

whence Eq. (B.2.1) immediately follows. □

For any $\gamma \in \mathscr{L}$, we have $\|\gamma\|_{\mathrm{TV}[0,t]} < \infty$ for any $t \in [0,1)$. We can therefore modify γ in (at most) countably many points to obtain a right-continuous function. Since this modification does not change the solution Φ_γ, by Proposition B.1.1, we will hereafter assume that any $\gamma \in \mathscr{L}$ is right-continuous.

For $\gamma \in \mathscr{L}$, we denote by $S(\gamma) \equiv \{t \in [0,1) : \gamma(t) > 0\}$ the *support* of γ, and by $\overline{S}(\gamma)$ the closure of $S(\gamma)$ in $[0,1)$ (in particular, note that $1 \notin \overline{S}(\gamma)$).

Lemma B.2.2. *The support of $\gamma \in \mathscr{L}_{\underline{q}}$ is a disjoint union of countably many intervals $S(\gamma) = \cup_{\alpha\in A} I_\alpha$, where $I_\alpha = (a_\alpha, b_\alpha)$ or $I_\alpha = [a_\alpha, b_\alpha)$, $\underline{q} \le a_\alpha < b_\alpha \le 1$, and A is countable.*

Proof. If $t_0 \in S(\gamma)$, then by right continuity there exists $\delta > 0$ such that $[t_0, t_0+\delta) \subseteq S(\gamma)$. This implies immediately the claim. □

Corollary B.2.3. *Assume $\gamma_* \in \mathscr{L}$ is such that $\mathsf{P}(\gamma_*) = \inf_{\gamma\in\mathscr{L}} \mathsf{P}(\gamma)$. Then*

$$t \in \overline{S}(\gamma_*) \quad \Rightarrow \quad \mathbb{E}\{\partial_x\Phi_{\gamma_*}(t,X_t)^2\} = t\,, \tag{B.2.5}$$

$$t \in [0,1)\setminus \overline{S}(\gamma_*) \quad \Rightarrow \quad \mathbb{E}\{\partial_x\Phi_{\gamma_*}(t,X_t)^2\} \ge t\,. \tag{B.2.6}$$

Proof. First consider Eq. (B.2.7). For any $0 \le t_1 < t_2 < 1$, set $\delta(t) = \gamma_*(t)\mathbb{I}(t \in [t_1,t_2))$. Clearly $\gamma_* + s\delta \in \mathscr{L}$ for $s \in (-1,1)$. By the optimality of γ_*, and using Proposition B.2.1, we have

$$0 = \frac{\mathrm{d}\mathsf{P}}{\mathrm{d}s}(\gamma_* + s\delta)\Big|_{s=0} = \frac{1}{2}\int_{t_1}^{t_2} \xi''(t)\gamma_*(t)\big(\,\mathbb{E}\{\partial_x\Phi_{\gamma_*}(t,X_t)^2\} - t\big)\,\mathrm{d}t$$

Since t_1, t_2 are arbitrary, and $\xi''(t) > 0$ for $t \in (0,1)$ this implies $\gamma_*(t)(\mathbb{E}\{\partial_x \Phi_{\gamma_*}(t, X_t)^2\} - t) = 0$ for almost every $t \in [0,1)$. Since $\gamma_*(t)$ is right-continuous and $\mathbb{E}\{\partial_x \Phi_{\gamma_*}(t, X_t)^2\}$ is continuous (see Corollary B.1.6), it follows that $\gamma_*(t)(\mathbb{E}\{\partial_x \Phi_{\gamma_*}(t, X_t)^2\} - t) = 0$ for every $t \in [0,1)$. This in turns implies $\mathbb{E}\{\partial_x \Phi_{\gamma_*}(t, X_t)^2\} = t$ for every $t \in S(\gamma_*)$. This can be extended to $t \in \overline{S}(\gamma_*)$ again by continuity of $t \mapsto \mathbb{E}\{\partial_x \Phi_{\gamma_*}(t, X_t)^2\}$.

Next consider Eq. (B.2.8). Notice that, by Lemma B.2.2, $[0,1) \setminus \overline{S}(\gamma_*)$ is a disjoint union of open intervals. Let J be such an interval, and consider any $[t_1, t_2] \subseteq J$. Set $\delta(t) = \mathbb{I}(t \in (t_1, t_2])$, and notice that $\gamma_* + s\delta \in \mathscr{L}$ for $s \geq 0$. By Proposition B.2.1, we have

$$0 \leq \frac{\mathrm{d}\mathsf{P}}{\mathrm{d}s}(\gamma + s\delta)\Big|_{s=0} = \frac{1}{2}\int_{t_1}^{t_2} \xi''(t)\big(\,\mathbb{E}\{\partial_x \Phi(t, X_t)^2\} \,-\, t\big)\,\mathrm{d}t\,.$$

Since t_1, t_2 are arbitrary, $\xi''(t) > 0$ for $t \in (0,1)$ and $t \mapsto \mathbb{E}\{\partial_x \Phi(t, X_t)^2\}$ is continuous, this implies $\mathbb{E}\{\partial_x \Phi(t, X_t)^2\} \geq t$ for all $t \in J$, and hence all $t \in [0,1) \setminus \overline{S}(\gamma_*)$. □

Corollary B.2.4. *Assume $\gamma_* \in \mathscr{L}$ is such that $\mathsf{P}(\gamma_*) = \inf_{\gamma \in \mathscr{L}} \mathsf{P}(\gamma)$. Then*

$$t \in \overline{S}(\gamma_*) \quad \Rightarrow \quad \xi''(t)\,\mathbb{E}\{\partial_x^2 \Phi_{\gamma_*}(t, X_t)^2\} = 1\,.$$

Proof. Set $\Phi(t,x) = \Phi_{\gamma_*}(t,x)$. By Lemma B.2.2, $\overline{S}(\gamma_*)$ is a disjoint union of closed intervals with non-empty interior. Let K be one such intervals. Then, for any $[t_1, t_2] \in K$, we have, by Lemma B.2.3

$$t_2 - t_1 = \mathbb{E}\{\partial_x \Phi(t_2, X_{t_2})^2\} - \mathbb{E}\{\partial_x \Phi(t_1, X_{t_1})^2\} = \int_{t_1}^{t_2} \xi''(t)\,\mathbb{E}\{\partial_x^2 \Phi(t, X_t)^2\}\mathrm{d}t\,.$$

Since t_1, t_2 are arbitrary, we get $\xi''(t)\,\mathbb{E}\{\partial_x^2 \Phi(t, X_t)^2\} = 1$ for almost every $t \in K$. Using Lemma B.1.7 we get $\xi''(t)\,\mathbb{E}\{\partial_x^2 \Phi(t, X_t)^2\} = 1$ for every $t \in \overline{S}(\gamma_*)$. □

Throughout this section we let $\gamma_*^{\mathscr{L}_{\underline{q}}}$ be the minimizer of P over $\mathscr{L}_{\underline{q}}$, assuming it exists. Note that we will eventually show in Lemma 4.1.3 that $\gamma_*^{\mathscr{L}_{\underline{q}}} = \gamma_*^{\mathscr{L}}$ if either minimizer exists.

Lemma B.2.5. *Assume $\gamma_*^{\mathscr{L}_{\underline{q}}}$ exists. Then*

$$t \in \mathrm{supp}(\gamma_*^{\mathscr{L}_{\underline{q}}}) \quad \Rightarrow \quad \mathbb{E}[\partial_x \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, X_t)^2] = t\,, \tag{B.2.7}$$

$$t \geq \underline{q} \quad \Rightarrow \quad \mathbb{E}[\partial_x \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, X_t)^2] \geq t\,. \tag{B.2.8}$$

Proof. We first show Equation (B.2.7). For $\underline{q} \leq t_1 < t_2 < 1$ we take $\delta(t) = \gamma_*^{\mathscr{L}_{\underline{q}}}(t) 1_{t \in [t_1, t_2)}$. Clearly $\gamma_*^{\mathscr{L}_{\underline{q}}} + s\delta \in \mathscr{L}_{\underline{q}}$. Since $\gamma_*^{\mathscr{L}_{\underline{q}}}$ minimizes $\mathsf{P}(\cdot)$ over $\mathscr{L}_{\underline{q}}$,

$$0 \le \frac{\mathrm{d}\mathsf{P}}{\mathrm{d}s}(\gamma_*^{\mathscr{L}_{\underline{q}}} + s\delta)\Big|_{s=0} = \frac{1}{2}\int_{t_1}^{t_2} \xi''(t)\gamma_*^{\mathscr{L}_{\underline{q}}}(t)\big(\mathbb{E}\left[\partial_x \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, X_t)^2\right] - t\big)\,\mathrm{d}t$$

Since t_1, t_2 are arbitrary, and $\xi''(t) > 0$ for $t \in (0,1)$ this implies $\gamma_*^{\mathscr{L}_{\underline{q}}}(t)(\mathbb{E}\left[\partial_x \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, X_t)^2\right] - t) = 0$ for almost every $t \in [\underline{q}, 1)$. Since $\gamma_*^{\mathscr{L}_{\underline{q}}}(t)$ is right-continuous and $\mathbb{E}\left[\partial_x \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, X_t)^2\right]$ is continuous by Lemma B.1.7, it follows that $\gamma_*^{\mathscr{L}_{\underline{q}}}(t)(\mathbb{E}\left[\partial_x \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, X_t)^2\right] - t) = 0$ for every $t \in [\underline{q}, 1)$. This in turns implies $\mathbb{E}\left[\partial_x \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, X_t)^2\right] = t$ for every $t \in S(\gamma_*^{\mathscr{L}_{\underline{q}}})$ by right-continuity of $\gamma_*^{\mathscr{L}_{\underline{q}}}$. This can be extended to all $t \in \mathrm{supp}(\gamma_*^{\mathscr{L}_{\underline{q}}})$ by again using continuity of $t \mapsto \mathbb{E}\left[\partial_x \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, X_t)^2\right]$.

Next consider Eq. (B.2.8), where it suffices now to consider $t \in [\underline{q}, 1) \setminus \mathrm{supp}(\gamma_*^{\mathscr{L}_{\underline{q}}})$. By Lemma B.2.2, $[\underline{q}, 1) \setminus \mathrm{supp}(\gamma_*^{\mathscr{L}_{\underline{q}}})$ is a disjoint union of open intervals. Let J be such an interval, and consider any $[t_1, t_2] \subseteq J$. Set $\delta(t) = \mathbb{I}(t \in (t_1, t_2])$, and notice that $\gamma_*^{\mathscr{L}_{\underline{q}}} + s\delta \in \mathscr{L}_{\underline{q}}$ for $s \ge 0$. By Proposition B.2.1, we have

$$0 \le \frac{\mathrm{d}\mathsf{P}}{\mathrm{d}s}(\gamma + s\delta)\Big|_{s=0} = \frac{1}{2}\int_{t_1}^{t_2} \xi''(t)\big(\mathbb{E}\left[\partial_x \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, X_t)^2\right] - t\big)\,\mathrm{d}t\,.$$

Since t_1, t_2 are arbitrary, $\xi''(t) > 0$ for $t \in (0,1)$ and $t \mapsto \mathbb{E}\left[\partial_x \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, X_t)^2\right]$ is continuous, this implies $\mathbb{E}\left[\partial_x \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, X_t)^2\right] \ge t$ for all $t \in J$, and hence all $t \in [\underline{q}, 1) \setminus \mathrm{supp}(\gamma_*^{\mathscr{L}_{\underline{q}}})$. □

Corollary B.2.6. *Assume $\gamma_*^{\mathscr{L}_{\underline{q}}}$ exists. Then*

$$t \in \mathrm{supp}(\gamma_*^{\mathscr{L}_{\underline{q}}}) \quad \Rightarrow \quad \xi''(t)\,\mathbb{E}\left[\partial_{xx} \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, X_t)^2\right] = 1\,.$$

Proof. By Lemma B.2.2, $\mathrm{supp}(\gamma_*^{\mathscr{L}_{\underline{q}}})$ is a disjoint union of closed intervals with non-empty interior. Let K be one such interval. Then, for any $[t_1, t_2] \in K$, Corollary B.1.6 and Lemma B.2.5 imply

$$t_2 - t_1 = \mathbb{E}\left[\partial_x \Phi(t_2, X_{t_2})^2\right] - \mathbb{E}\left[\partial_x \Phi(t_1, X_{t_1})^2\right] = \int_{t_1}^{t_2} \xi''(t)\,\mathbb{E}\left[\partial_{xx} \Phi(t, X_t)^2\right]\mathrm{d}t\,.$$

Since t_1, t_2 are arbitrary, $\xi''(t)\,\mathbb{E}\left[\partial_{xx} \Phi(t, X_t)^2\right] = 1$ for almost every $t \in K$. By Lemma B.1.7 it follows that $\xi''(t)\,\mathbb{E}\left[\partial_{xx} \Phi(t, X_t)^2\right] = 1$ for all $t \in \mathrm{supp}(\gamma_*^{\mathscr{L}_{\underline{q}}})$. □

Lemma B.2.7. *Let $\gamma \in \mathscr{L}$ satisfy $\gamma(t) = 0$ for all $t \in (t_1, 1)$, where $t_1 < 1$. Then, for any $t_* \in (t_1, 1)$, the probability distribution of X_{t_*} has a density p_{t_*} with respect to the Lebesgue measure. Further, for any $t_* \in (t_1, 1)$ and any $M \in \mathbb{R}_{\ge 0}$, there exists $\varepsilon(t_*, M, \gamma) > 0$ such that*

$$\inf_{|x| \le M, t \in [t_*, 1]} p_t(x) \ge \varepsilon(t_*, M, \gamma)\,.$$

Proof. Since the SDE (B.3.1) has strong solutions, X_{t_1} is a well defined random variable taking values in $\mathbb{R}$. Therefore, there exists $C_1 = C_1(\gamma) < \infty$ such that $\mathbb{P}(|X_{t_1}| \le C_1) \ge 1/2$. For $t \in (t_1, 1)$, X_t satisfies $\mathrm{d}X_t = \sqrt{\xi''(t)}\,\mathrm{d}B_t$ and therefore the law of X_t is the convolution of a Gaussian (with variance $\theta(t)^2 \equiv \xi'(t) - \xi(t_1) > 0$) with the law of X_{t_1}, and therefore has a density. To prove the desired lower bound on the density, let $f_G(x) = \exp(-x^2/2)/\sqrt{2\pi}$ denote the standard Gaussian density. Note that, for any $|x| \le M$,

$$\begin{aligned} p_t(x) &= \mathbb{E}\Big\{\frac{1}{\theta(t)} f_G\Big(\frac{x - X_{t_1}}{\theta(t)}\Big)\Big\} \\ &\ge \mathbb{E}\Big\{\frac{1}{\theta(t)} f_G\Big(\frac{x - X_{t_1}}{\theta(t)}\Big)\mathbb{I}_{|X_{t_2}| \le C_1}\Big\} \\ &\ge \frac{1}{\theta(t)} f_G\Big(\frac{M + C_1}{\theta(t)}\Big)\,\mathbb{P}(|X_{t_1}| \le C_1) \ge \frac{1}{2\theta(t)} f_G\Big(\frac{M + C_1}{\theta(t)}\Big)\,. \end{aligned}$$

The latter expression is lower bounded by $\varepsilon(t_*, M, \gamma) > 0$ for any $t \in [t_*, 1]$, as claimed. □

Proposition B.2.8. *For any $\gamma \in \mathscr{L}$, the following identities hold:*

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}\left[\Phi_\gamma(t, X_t)\right] = \frac{1}{2}\,\xi''(t)\gamma(t)\,\mathbb{E}\left[\partial_x \Phi_\gamma(t, X_t)^2\right] \tag{B.2.9}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbb{E}\left[X_t \partial_x \Phi_\gamma(t, X_t)\right] = \xi''(t)\gamma(t)\,\mathbb{E}\left[\partial_x \Phi_\gamma(t, X_t)^2\right] + \xi''(t)\,\mathbb{E}\left[\partial_{xx} \Phi_\gamma(t, X_t)\right]\,. \tag{B.2.10}$$

Proof. We will write $\Phi_t = \partial_t \Phi$, $\Phi_x = \partial_x \Phi$ and $\Phi_{xx} = \partial_x^2 \Phi$. For the first identity, using the regularity properties of Lemma B.1.4 and Itô's formula, we get

$$\begin{aligned} \mathrm{d}\Phi(t, X_t) =& \Phi_t(t, X_t)\,\mathrm{d}t + \xi''(t)\gamma(t)\Phi_x(t, X_t)^2 \mathrm{d}t + \sqrt{\xi''(t)}\,\Phi_x(t, X_t)\,\mathrm{d}B_t + \frac{1}{2}\Phi_{xx}(t, X_t)\,\xi''(t)\mathrm{d}t \\ =& \frac{1}{2}\xi''(t)\gamma(t)\Phi_x(t, X_t)^2 \mathrm{d}t + \sqrt{\xi''(t)}\,\Phi_x(t, X_t)\,\mathrm{d}B_t\,, \end{aligned}$$

where the equalities hold after integrating over a test function $g \in C_c^\infty([0, 1))$ and in the second step we used the fact that Φ is a weak solution of Eq. (B.0.2). The claim (B.2.9) follows by taking expectations.

We proceed analogously for the second identity. Using Lemma B.1.5, and the fact that the $(X_t)_{t \in [0,1)}$ solved the SDE (B.3.1), we get

$$\begin{aligned} \mathrm{d}\big(X_t \Phi_x(t, X_t)\big) =& \Phi_x(t, X_t)\mathrm{d}X_t + X_t \mathrm{d}\big(\Phi_x(t, X_t)\big) + \xi''(t)\Phi_{xx}(t, X_t)\,\mathrm{d}t \\ =& \xi''(t)\gamma(t)\Phi_x(t, X_t)^2 \mathrm{d}t + \sqrt{\xi''(t)}\Phi_x(t, X_t)\mathrm{d}B_t + \sqrt{\xi''(t)}X_t\Phi_{xx}(t, X_t)\mathrm{d}B_t \\ &+ \xi''(t)\Phi_{xx}(t, X_t)\,\mathrm{d}t\,. \end{aligned}$$

The claim (B.2.9) follows again by taking expectations. □

We now show that any minimizer γ_* of the Parisi functional over the extended space $\mathscr{L}$ has support given by an interval containing 1. Note that this is unrelated to the no-overlap gap property which concerns solutions γ_* that are non-decreasing, and concerns the points of increase of γ_*.

Lemma B.2.9. *Assuming $\gamma_*^{\mathscr{L}_{\underline{q}}}$ exists, we have* $\operatorname{supp}(\gamma_*^{\mathscr{L}_{\underline{q}}}) = [\underline{q}, 1)$.

Proof. By Lemma B.2.2, $[\underline{q}, 1) \setminus \operatorname{supp}(\gamma_*^{\mathscr{L}_{\underline{q}}})$ is a countable union of disjoint intervals, open in $[\underline{q}, 1)$. First assume that at least one of these intervals is of the form (t_1, t_2) with $\underline{q} \le t_1 < t_2 < 1$. By Lemma B.2.5 and Corollary B.2.6 we know that

$$\mathbb{E}\left[\partial_x \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t_i, X_{t_i})^2\right] = t_i\,, \quad \xi''(t_i)\,\mathbb{E}\left[\partial_{xx} \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t_i, X_{t_i})^2\right] = 1\,, \quad i \in \{1, 2\}\,, \tag{B.2.11}$$

$$\mathbb{E}\left[\partial_x \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, X_t)^2\right] \ge t \qquad \forall t \in (t_1, t_2)\,. \tag{B.2.12}$$

Further, for $t \in (t_1, t_2)$, $\Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}$ solves the PDE

$$\partial_t \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, x) + \frac{\xi''(t)}{2} \partial_{xx} \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, x) = 0$$

which is simply the heat equation up to a time change. We therefore obtain

$$\Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, x) = \mathop{\mathbb{E}}^{Z \sim \mathsf{N}(0,1)}\left[\Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t_2, x + \sqrt{\xi'(t_2) - \xi'(t)}\, Z)\right], \quad \forall t \in (t_1, t_2].$$

Differentiating this equation and using dominated convergence (recall that $\partial_{xx} \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t_2, x)$ is bounded by Proposition 4.2.5), we obtain $\partial_{xx} \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, x) = \mathbb{E}^{Z \sim \mathsf{N}(0,1)}\left[\partial_{xx} \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t_2, x + \sqrt{\xi'(t_2) - \xi'(t)}\, Z)\right]$. Because $\mathrm{d}X_t = \sqrt{\xi''(t)}\, \mathrm{d}B_t$, we can rewrite the last equation as

$$\partial_{xx} \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, X_t) = \mathbb{E}\left[\partial_{xx} \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t_2, X_{t_2}) | X_t\right].$$

By Jensen's inequality,

$$\mathbb{E}\left[\partial_{xx} \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, X_t)^2\right] \le \mathbb{E}\left[\partial_{xx} \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t_2, X_{t_2})^2\right] = \frac{1}{\xi''(t_2)}\,, \tag{B.2.13}$$

where in the last step we used Eq. (B.2.11). Using Corollary B.1.6 we get, for $t \in [t_1, t_2]$

$$\begin{aligned}\mathbb{E}\left[\partial_x \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, X_t)^2\right] &= \mathbb{E}\left[\partial_x \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t_1, X_{t_1})^2\right] + \int_{t_1}^t \xi''(s)\, \mathbb{E}\left[\partial_{xx} \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(s, X_s)^2\right] \mathrm{d}s \\ &\le t_1 + \int_{t_1}^t \frac{\xi''(s)}{\xi''(t_2)}\, \mathrm{d}s < t\,,\end{aligned}$$

where in the last step we used the fact that $t \mapsto \xi''(t)$ is increasing. The last equation is in contradiction with Eq. (B.2.12), and therefore $[\underline{q}, 1) \setminus \operatorname{supp}(\gamma_*^{\mathscr{L}_{\underline{q}}})$ is either empty or consists of a

single interval $(t_1, 1)$.

In order to complete the proof, we need to rule out the case $[\underline{q}, 1)\backslash\mathrm{supp}(\gamma_*^{\mathscr{L}_{\underline{q}}}) = (t_1, 1)$. Assume for sake of contradiction that indeed $[\underline{q}, 1)\backslash\mathrm{supp}(\gamma_*^{\mathscr{L}_{\underline{q}}}) = (t_1, 1)$. For $t \in (t_1, 1)$, let $r = r(t) = \xi'(1) - \xi'(t)$, and notice that $r(t)$ is decreasing with $r(t) = \xi''(1)(1-t) + O((1-t)^2)$ as $t \to 1$. By solving the Parisi PDE in the interval $(t_1, 1)$, we find that for all $t \in (t_1, 1)$,

$$\partial_x \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, x) = \mathop{\mathbb{E}}^{Z\sim\mathsf{N}(0,1)}\left[\mathrm{sign}\left(Z + \frac{x}{\sqrt{r(t)}}\right)\right]$$

and therefore

$$1 - \mathbb{E}\left[\partial_x \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, X_t)^2\right] = \mathbb{E}\left[Q\left(\frac{X_t}{\sqrt{r(t)}}\right)\right],$$
$$Q(x) \equiv 1 - \mathop{\mathbb{E}}^{Z\sim\mathsf{N}(0,1)}\left[\mathrm{sign}(x + Z)\right]^2.$$

Note that $0 \le Q(x) \le 1$ is continuous, with $Q(0) = 1$. Hence, there exists a numerical constant $\delta_0 \in (0, 1)$ such that $Q(x) \ge 1/2$ for $|x| \le \delta_0$. Therefore, fixing $t_* \in (t_1, 1)$, for any $t \in (t_*, 1)$

$$1 - \mathbb{E}\left[\partial_x \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, X_t)^2\right] \ge \frac{1}{2}\mathbb{P}\left[|X_t| \le \delta_0\sqrt{r(t)}\right]$$
$$\overset{(a)}{\ge} \delta_0 \varepsilon(t_*, 1, \gamma)\sqrt{r(t)} \overset{(b)}{\ge} C\sqrt{1-t},$$

where (a) follows by Lemma B.2.7 and (b) holds for some $C = C(\gamma) > 0$. We therefore obtain $\mathbb{E}\left[\partial_x \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, X_t)^2\right] \le 1 - C\sqrt{1-t}$, which contradicts Lemma B.2.5 for t close enough to 1. □

In the next lemma, we show that minimization of P over $\mathscr{L}$ subsumes minimization over $\mathscr{L}_{\underline{q}}$. A priori, one might expect that tuning the value of $\underline{q}$ could lead to many different minima.

Lemma B.2.10. *Suppose $\gamma_*^{\mathscr{L}_{\underline{q}}}$ exists. Then $\gamma_*^{\mathscr{L}} = \gamma_*^{\mathscr{L}_{\underline{q}}}$.*

Proof. Let $f(t) = \mathbb{E}[\partial_x \Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, X_t)^2]$. First by Proposition B.2.1, it suffices to show that $f(t) \ge t$ for all $0 \le t \le \underline{q}$. Indeed this combined with Corollary B.1.6 and Lemma B.2.5 would imply that $\frac{d\mathsf{P}}{ds}((1-s)\gamma_*^{\mathscr{L}_{\underline{q}}} + s\gamma)|_{s=0+} \ge 0$ for any $\gamma \in \mathscr{L}$. This suffices as P is convex by Lemma B.1.8.

To show $f(t) \ge t$ for all $0 \le t \le \underline{q}$, we first recall that X_t is simply a time-changed Brownian motion on $0 \le t \le \underline{q}$ while $\Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}$ solves the time-changed heat equation on the same time interval, therefore $\partial_{xx}\Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t, X_t) = \mathbb{E}^t[\partial_{xx}\Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(\underline{q}, X_{\underline{q}})]$. By Jensen's inequality, it follows that for $0 \le t \le \underline{q}$, we have

$$\begin{aligned}\mathbb{E}[\partial_{xx}\Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(t,X_t)^2] &\le \mathbb{E}[\partial_{xx}\Phi_{\gamma_*^{\mathscr{L}_{\underline{q}}}}(\underline{q},X_{\underline{q}})^2]\\ &= \frac{1}{\xi''(\underline{q})}\\ &\le \frac{1}{\xi''(t)}.\end{aligned}$$

In the last line we used that ξ'' is increasing as ξ is a power series with non-negative coefficients. Next, from Lemma B.2.5 and Lemma B.2.9 it follows that $f(\underline{q}) = \underline{q}$. In light of Corollary B.1.6, we showed just above that $f'(t) \le 1$ for $t \le \underline{q}$. It now follows that $f(t) \ge t$ for all $0 \le t \le \underline{q}$ which completes the proof. □

B.3 Proofs of Lemmas 4.1.3, 4.1.4, and 4.2.8

We first restate and prove Lemmas 4.1.3 and 4.1.4.

Lemma 4.1.3. *For $\gamma_* \in \mathscr{L}$ and $\underline{q} = \inf(\mathrm{supp}(\gamma_*))$, the following are equivalent:*

1. *γ_* is optimizable.*
2. *$\mathsf{P}(\gamma_*) = \inf_{\gamma\in\mathscr{L}} \mathsf{P}(\gamma)$.*
3. *$\mathsf{P}(\gamma_*) = \inf_{\gamma\in\mathscr{L}_{\underline{q}}} \mathsf{P}(\gamma)$.*

Moreover if a minimizer exists in either variational problem just above, then it is unique.

Proof. Lemma B.1.8 immediately implies uniqueness of minimizers. The second statement immediately implies the third, while Lemma B.2.10 provides the converse result. To show that the first statement implies the third, we observe that Proposition B.2.1 immediately yields

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathsf{P}((1-s)\gamma_* + s\gamma)|_{s=0+} = 0$$

for any $\gamma \in \mathscr{L}_{\underline{q}}$ when γ_* is optimizable; this implies the third statement by again invoking Lemma B.1.8. It only remains to show that if $\mathsf{P}(\gamma_*) = \inf_{\gamma\in\mathcal{L}} \mathsf{P}(\gamma)$, then γ_* is $\underline{q}$-optimizable, which follows from Lemmas B.2.5 and B.2.9. □

Lemma 4.1.4. *If $\gamma_*^{\mathscr{U}}$ strictly increases on $[\underline{q}, 1)$ for $\underline{q} = \inf(\mathrm{supp}(\gamma_*^{\mathscr{U}}))$, then no overlap gap holds, i.e. $\gamma_*^{\mathscr{U}}$ is optimizable.*

Proof. Fix $\underline{q} < t_1 < t_2 < 1$ and define $\delta(t) = [\gamma_*^{\mathscr{U}}(t_1) - \gamma_*^{\mathscr{U}}(t)]\mathbb{I}_{(t_1,t_2)}(t)$. It is easy to see that this satisfies the assumptions of Proposition B.2.1 with $s_0 = 1$. Letting $\gamma^s = \gamma_*^{\mathscr{U}} + s\delta$,

$$\frac{\mathrm{d}\mathsf{P}}{\mathrm{d}s}(\gamma^s)\bigg|_{s=0+} = -\frac{1}{2}\int_{t_1}^{t_2} \xi''(t)\big(\gamma_*^{\mathscr{U}}(t) - \gamma_*^{\mathscr{U}}(t_1)\big)\left(\mathbb{E}\left[\partial_x \Phi_{\gamma_*^{\mathscr{U}}}(t, X_t)^2\right] - t\right)\mathrm{d}t.$$

On the other hand, $\gamma^s \in \mathscr{U}$ for $s \in [0,1]$ (since $\gamma_*^{\mathscr{U}}$ is strictly increasing on $[\underline{q}, 1)$), so

$$\int_{t_1}^{t_2} \xi''(t)\big(\gamma_*^{\mathscr{U}}(t) - \gamma_*^{\mathscr{U}}(t_1)\big)\left(\mathbb{E}[\partial_x \Phi_{\gamma_*^{\mathscr{U}}}(t, X_t)^2] - t\right)\mathrm{d}t \le 0\,.$$

for all $t_1 < t_2$. Since $\gamma_*^{\mathscr{U}}(t) - \gamma_*^{\mathscr{U}}(t_1) > 0$ strictly for all $t > t_1$, this implies

$$\mathbb{E}[\partial_x \Phi_{\gamma_*^{\mathscr{U}}}(t, X_t)^2] \le t$$

for almost every t, and therefore for every t. The inequality

$$\mathbb{E}[\partial_x \Phi_{\gamma_*^{\mathscr{U}}}(t, X_t)^2] \ge t$$

is proved in the same way using $\delta(t) = [\gamma_*^{\mathscr{U}}(t_2) - \gamma_*^{\mathscr{U}}(t)]\mathbb{I}_{(t_1,t_2)}(t)$. □

Finally we turn to Lemma 4.2.8, for which we recall some technical results on solutions to SDEs with non-Lipschitz coefficients. We say the 1-dimensional SDE

$$\mathrm{d}X_t = b(t, X_t)\mathrm{d}t + \sigma(t, X_t)\mathrm{d}B_t \tag{B.3.1}$$

satisfies the *Yamada-Watanabe* criteria if $|\sigma(s,x) - \sigma(s,y)|^2 \le \frac{1}{|x-y|}$ and $b(t,x)$ is uniformly Lipschitz in x on compact time-sets.

Proposition B.3.1 ([RY13, Chapter IX, Theorem 3.5 part (ii)])**.** *If the solution X_t to Equation* (B.3.1) *satisfies the Yamada-Watanabe criteria, then X_t also satisfies pathwise uniqueness.*

Proposition B.3.2 ([IW77, Theorem 1.1])**.** *Let the measurable functions $\sigma_1(t,x), \sigma_2(t,x), b_1(t,x), b_2(t,x)$ be such that $(\sigma_1, b_1), (\sigma_2, b_2)$ satisfy the Yamada-Watanabe criteria. Further suppose that*

$$b_1(t,x) \le b_2(t,x)$$

holds for all t, x. Let X_t^1, X_t^2 solve

$$\mathrm{d}X_t^i = b_i(t, X_t)\mathrm{d}t + \sigma_i(t, X_t)\mathrm{d}B_t, \qquad i \in \{1,2\}.$$

If $X_0^1 = X_0^2$, then almost surely $X_t^1 \le X_t^2$ for all $t \ge 0$.

Lemma 4.2.8. *If $\gamma_* \in \mathscr{L}$ is $\underline{q}$-optimizable then it satisfies:*

$$\mathbb{E}[\partial_{xx}\Phi_{\gamma_*}(t, X_t)^2] = \frac{1}{\xi''(t)}, \quad t \geq \underline{q}, \tag{4.2.7}$$

$$\mathbb{E}[\partial_{xx}\Phi_{\gamma_*}(t, X_t)] = \int_t^1 \gamma_*(s)\mathrm{d}s, \quad t \in [0, 1]. \tag{4.2.8}$$

Proof. First, (4.2.7) is clear given Corollary B.1.6. To establish (4.2.8), we first show that

$$d(\partial_{xx}\Phi_{\gamma_*}(t, X_t)) = -\xi''(t)\gamma_*(t)(\partial_{xx}\Phi_{\gamma_*}(t, X_t))^2\mathrm{d}t + \partial_{xxx}\Phi_{\gamma_*}(t, X_t)\sqrt{\xi''(t)}\mathrm{d}B_t. \tag{B.3.2}$$

Indeed (B.3.2) follows by using Ito's formula to derive

$$\begin{aligned}\mathrm{d}(\partial_{xx}\Phi_{\gamma_*}(t, X_t)) = &\left(\partial_{txx}\Phi_{\gamma_*}(t, X_t) + \partial_x\Phi_{\gamma_*}(t, X_t)\partial_{xxx}\Phi_{\gamma_*}(t, X_t)\xi''(t)\gamma_*(t) + \frac{\xi''(t)\partial_{xxxx}\Phi_{\gamma_*}(t, X_t)}{2}\right)\mathrm{d}t \\ &+ \partial_{xxx}\Phi_{\gamma_*}(t, X_t)\sqrt{\xi''(t)}\mathrm{d}B_t\end{aligned}$$

and taking the second derivative with respect to x of the Parisi PDE to obtain

$$\begin{aligned}0 &= \partial_{xx}\left(\partial_t\Phi_{\gamma_*}(t, x) + \frac{\xi''(t)}{2}\left(\partial_{xx}\Phi_{\gamma_*}(t, x) + \gamma_*(t)(\partial_x\Phi_{\gamma_*}(t, x))^2\right)\right) \\ &= \partial_{txx}\Phi_{\gamma_*}(t, x) + \frac{\xi''(t)\partial_{xxxx}\Phi_{\gamma_*}(t, x)}{2} + \xi''(t)\gamma_*(t)\left((\partial_{xx}\Phi_{\gamma_*}(t, x))^2 + \partial_x\Phi_{\gamma_*}(t, x)\partial_{xxx}\Phi_{\gamma_*}(t, x)\right).\end{aligned}$$

In particular (B.3.2) implies that for all $t \in [0, 1)$,

$$\frac{d}{dt}\mathbb{E}[\partial_{xx}\Phi_{\gamma_*}(t, x)] = -\xi''(t)\gamma_*(t)\mathbb{E}[(\partial_{xx}\Phi_{\gamma_*}(t, x))^2] \tag{B.3.3}$$

$$= -\gamma_*(t). \tag{B.3.4}$$

Therefore to show (4.2.8) it suffices to show $\lim_{t\to 1}\mathbb{E}[\partial_{xx}\Phi_{\gamma_*}(t, X_t)] = 0$. Recalling that $\mathbb{E}[(\partial_{xx}\Phi_{\gamma_*}(t, X_t))^2] = \frac{1}{\xi''(t)}$ is bounded on $t \in [\underline{q}, 1]$, we use the general inequality

$$\begin{aligned}\mathbb{E}[Y] &\leq \mathbb{E}[Y \cdot \mathbb{I}_{Y\geq\varepsilon}] + \varepsilon \\ &\leq \sqrt{\mathbb{E}[Y^2] \cdot \mathbb{P}[Y \geq \varepsilon]} + \varepsilon\end{aligned}$$

which holds for any random variable Y. Taking $Y = \partial_{xx}\Phi_{\gamma_*}(t, X_t))$ and noting that $\mathbb{E}[Y^2]$ is uniformly bounded (independent of t), it suffices to show p-$\lim_{t\to 1}\partial_{xx}\Phi_{\gamma_*}(t, X_t) = 0$. To this end we recall that $\Phi_{\gamma_*}(t, x)$ is continuous on $[0, 1] \times \mathbb{R}$ and is convex in x, with $\Phi_{\gamma_*}(1, x) = |x|$. It follows

that for any $\varepsilon > 0$,

$$\lim_{t\to 1} \sup_{|x|\geq\varepsilon} \partial_{xx}\Phi_{\gamma_*}(t,x) = 0.$$

Therefore to establish p-$\lim_{t\to 1} \partial_{xx}\Phi_{\gamma_*}(t, X_t) = 0$ it suffices to show

$$\lim_{\varepsilon\to 0} \lim_{t\to 1} \mathbb{P}[|X_t| \leq \varepsilon] = 0. \tag{B.3.5}$$

To do this we will use Proposition B.3.2 to show that $|X_t|$ is stochastically larger than $|Z_t|$ for $Z_t = \int_0^t \sqrt{\xi''(t)}\mathrm{d}B_t \sim \mathsf{N}(0, \xi'(t))$, which implies (B.3.5). Applying Ito's formula to $(X_t)^2, (Z_t)^2$ gives

$$\begin{aligned}
\mathrm{d}(X_t)^2 &= (\xi''(t)\gamma_*(t)X_t\partial_x\Phi_{\gamma_*}(t, X_t) + \xi''(t))\mathrm{d}t + 2X_t\sqrt{\xi''(t)}\mathrm{d}B_t,\\
\mathrm{d}(Z_t)^2 &= \xi''(t)\mathrm{d}t + 2Z_t\sqrt{\xi''(t)}\mathrm{d}B_t.
\end{aligned}$$

We now define Brownian motions B_t^1, B_t^2 via $B_t^1 = \mathrm{sign}(X_t)B_t$ and $B_t^2 = \mathrm{sign}(Z_t)\mathrm{d}B_t$. It is easy to see by symmetry of the above SDEs that these are both Brownian motions. Then taking $Y_t = (X_t)^2, W_t = (Z_t)^2$,

$$\begin{aligned}
\mathrm{d}Y_t &= (\xi''(t)\gamma_*(t)\sqrt{Y_t}\partial_x\Phi_{\gamma_*}(t, \sqrt{Y_t}) + \xi''(t))\mathrm{d}t + 2\sqrt{Y_t\xi''(t)}\mathrm{d}B_t^1,\\
\mathrm{d}W_t &= \xi''(t)\mathrm{d}t + 2\sqrt{W_t\xi''(t)}\mathrm{d}B_t^2.
\end{aligned}$$

Here we use the fact that $x\partial_x\Phi_{\gamma_*}(t, x)$ is an even function (because $\Phi_{\gamma_*}(t, x)$ is even for any t) to obtain the first equation. Proposition B.3.1 applies to both SDEs, implying pathwise uniqueness and hence uniqueness in law for Y_t, W_t. Moreover $x\partial_x\Phi_{\gamma_*}(t, x) \geq 0$ holds for all (t, x) because $\Phi_{\gamma_*}(t, x)$ is convex and even in x. Hence Proposition B.3.2 applies to the above pair of SDEs, ensuring that

$$Y_t \geq W_t \geq 0$$

holds pathwise if $B_t^1 = B_t^2$. (Here we treat B_t^1, B_t^2 as unrelated Brownian motions which can be coupled together, forgetting their definitions based on B_t.) Uniqueness in law now implies that Y_t is stochastically larger than W_t, hence $|X_t|$ is stochastically larger than $|Z_t|$. We conclude that (B.3.5) holds, completing the proof of Equation (4.2.8) when γ_* is optimizable. □

Appendix C

Deferred Proofs from Chapter 5

C.1 Overlap Concentration of Standard Optimization Algorithms

In this section we prove using Gaussian concentration of measure and Kirszbraun's theorem that approximately τ-Lipschitz functions $\mathcal{A} : \mathcal{H}_N \to B_N$ are $(\lambda, e^{-c_{\lambda,\tau} N})$ overlap concentrated. We also show that common optimization algorithms such as gradient descent, AMP, and Langevin dynamics are approximately Lipschitz.

C.1.1 Overlap Concentration of Approximately Lipschitz Algorithms

Recall that we identify each Hamiltonian H_N with its disorder coefficients $(\mathbf{G}^{(p)})_{p \in 2\mathbb{N}}$, which we concatenate into an infinite vector $\mathbf{g} = \mathbf{g}(H_N)$. We can define a (possibly infinite) distance on these Hamiltonians by

$$\|H_N - H'_N\|_N = \frac{1}{\sqrt{N}} \|\mathbf{g}(H_N) - \mathbf{g}(H'_N)\|_2. \tag{C.1.1}$$

We consider algorithms $\mathcal{A} : \mathcal{H}_N \to B_N$ that are τ-Lipschitz with respect to the $\|\cdot\|_N$ norms, i.e. $\mathcal{A}$ satisfying

$$\|\mathcal{A}(H_N) - \mathcal{A}(H'_N)\|_N \le \tau \|H_N - H'_N\|_N. \tag{C.1.2}$$

for all $H_N, H'_N \in \mathcal{H}_N$. This is the same notion of Lipschitz as in Theorem 22, though the current scaling with $\|\cdot\|_N$ norms will be more convenient for proofs.

We will show overlap concentration for the following class of algorithms that relax the Lipschitz condition to a high probability set of inputs.

Definition C.1.1. *Let $\tau, \nu > 0$. An algorithm $\mathcal{A} : \mathcal{H}_N \to B_N$ is (τ, ν)-approximately Lipschitz if there exists a τ-Lipschitz $\mathcal{A}' : \mathcal{H}_N \to B_N$ with*

$$\mathbb{P}\left[\mathcal{A}(H_N) = \mathcal{A}'(H_N)\right] \geq 1 - \nu. \tag{C.1.3}$$

Proposition C.1.2. *If $\mathcal{A} : \mathcal{H}_N \to B_N$ is τ-Lipschitz, then for all $\lambda > 0$ it is $\left(\lambda, \exp\left(-\frac{\lambda^2}{8\tau^2}N\right)\right)$ overlap concentrated.*

Proof. We write $\mathcal{A}(\mathbf{g})$ to mean $\mathcal{A}(H_N)$ for the Hamiltonian H_N with disorder coefficients $\mathbf{g} = \mathbf{g}(H_N)$. Let $\mathcal{A}_i(\mathbf{g})$ denote the i-th coordinate of $\mathcal{A}(\mathbf{g})$, so $\mathcal{A}(\mathbf{g}) = (\mathcal{A}_1(\mathbf{g}), \ldots, \mathcal{A}_N(\mathbf{g}))$. Define the gradient matrix $\nabla\mathcal{A}(\mathbf{g}) \in \mathbb{R}^{\mathbb{N}\times N}$ by

$$\nabla\mathcal{A}(\mathbf{g}) = \begin{bmatrix}\nabla\mathcal{A}_1(\mathbf{g}) & \nabla\mathcal{A}_2(\mathbf{g}) & \cdots & \nabla\mathcal{A}_N(\mathbf{g})\end{bmatrix}.$$

Because $\mathcal{A}$ is τ-Lipschitz, we have for all $\mathbf{g}, \mathbf{g}' \in \mathbb{R}^{\mathbb{N}}$ that

$$\lambda \geq \frac{\|\mathcal{A}(\mathbf{g}) - \mathcal{A}(\mathbf{g}')\|_N}{\|\mathbf{g} - \mathbf{g}'\|_N}.$$

By taking the limit $\mathbf{g}' \to \mathbf{g}$ from the best direction, we conclude that for all $\mathbf{g} \in \mathbb{R}^{\mathbb{N}}$,

$$\lambda \geq s_{\max}(\nabla\mathcal{A}(\mathbf{g})), \tag{C.1.4}$$

where $s_{\max}$ denotes the largest singular value.

Consider any $p \in [0, 1]$. We can generate p-correlated $\mathbf{g}^{(1)}, \mathbf{g}^{(2)} \in \mathbb{R}^{\mathbb{N}}$ by generating i.i.d. $\mathbf{g}^{[0]}, \mathbf{g}^{[1]}, \mathbf{g}^{[2]} \in \mathbb{R}^{\mathbb{N}}$, each with i.i.d. standard Gaussian entries, and setting, for $i = 1, 2$,

$$\mathbf{g}^{(i)} = \sqrt{p}\mathbf{g}^{[0]} + \sqrt{1-p}\mathbf{g}^{[i]}.$$

We will apply Gaussian concentration to the function

$$F(\mathbf{g}^{[0]}, \mathbf{g}^{[1]}, \mathbf{g}^{[2]}) = R\left(\mathcal{A}(\mathbf{g}^{(0)}), \mathcal{A}(\mathbf{g}^{(1)})\right),$$

which is a function of i.i.d. standard Gaussians. For each $i \in \mathbb{N}$, let $\nabla\mathcal{A}_{\cdot,i}(\mathbf{g})$ denote the i-th row of $\nabla\mathcal{A}(\mathbf{g})$, i.e.

$$\nabla\mathcal{A}_{\cdot,i}(\mathbf{g}) = \begin{bmatrix}\frac{\partial\mathcal{A}_1}{\partial\mathbf{g}_i}(\mathbf{g}) & \frac{\partial\mathcal{A}_2}{\partial\mathbf{g}_i}(\mathbf{g}) & \cdots & \frac{\partial\mathcal{A}_N}{\partial\mathbf{g}_i}(\mathbf{g})\end{bmatrix}.$$

We can compute that

$$\frac{\partial F}{\partial \mathbf{g}_i^{[0]}}(\mathbf{g}^{[0]},\mathbf{g}^{[1]},\mathbf{g}^{[2]}) = \frac{\sqrt{p}}{N}\left[\nabla \mathcal{A}_{\cdot,i}(\mathbf{g}^{(1)})\mathcal{A}(\mathbf{g}^{(2)}) + \nabla \mathcal{A}_{\cdot,i}(\mathbf{g}^{(2)})\mathcal{A}(\mathbf{g}^{(1)})\right], \tag{C.1.5}$$

$$\frac{\partial F}{\partial \mathbf{g}_i^{[1]}}(\mathbf{g}^{[0]},\mathbf{g}^{[1]},\mathbf{g}^{[2]}) = \frac{\sqrt{1-p}}{N}\nabla \mathcal{A}_{\cdot,i}(\mathbf{g}^{(1)})\mathcal{A}(\mathbf{g}^{(2)}), \tag{C.1.6}$$

$$\frac{\partial F}{\partial \mathbf{g}_i^{[2]}}(\mathbf{g}^{[0]},\mathbf{g}^{[1]},\mathbf{g}^{[2]}) = \frac{\sqrt{1-p}}{N}\nabla \mathcal{A}_{\cdot,i}(\mathbf{g}^{(2)})\mathcal{A}(\mathbf{g}^{(1)}). \tag{C.1.7}$$

By the inequality $(x+y) \le 2x^2 + 2y^2$, (C.1.5) implies

$$\frac{\partial F}{\partial \mathbf{g}_i^{[0]}}(\mathbf{g}^{[0]},\mathbf{g}^{[1]},\mathbf{g}^{[2]})^2 \le \frac{2p}{N^2}\left[\left(\nabla \mathcal{A}_{\cdot,i}(\mathbf{g}^{(1)})\mathcal{A}(\mathbf{g}^{(2)})\right)^2 + \left(\nabla \mathcal{A}_{\cdot,i}(\mathbf{g}^{(2)})\mathcal{A}(\mathbf{g}^{(1)})\right)^2\right].$$

Similarly, (C.1.6) and (C.1.7) imply

$$\frac{\partial F}{\partial \mathbf{g}_i^{[1]}}(\mathbf{g}^{[0]},\mathbf{g}^{[1]},\mathbf{g}^{[2]})^2 \le \frac{2(1-p)}{N^2}\left(\nabla \mathcal{A}_{\cdot,i}(\mathbf{g}^{(1)})\mathcal{A}(\mathbf{g}^{(2)})\right)^2,$$
$$\frac{\partial F}{\partial \mathbf{g}_i^{[2]}}(\mathbf{g}^{[0]},\mathbf{g}^{[1]},\mathbf{g}^{[2]})^2 \le \frac{2(1-p)}{N^2}\left(\nabla \mathcal{A}_{\cdot,i}(\mathbf{g}^{(2)})\mathcal{A}(\mathbf{g}^{(1)})\right)^2.$$

Summing over the last three inequalities and over $i \in \mathbb{N}$ gives

$$\begin{aligned}\left\|\nabla F(\mathbf{g}^{[0]},\mathbf{g}^{[1]},\mathbf{g}^{[2]})\right\|_2^2 &\le \frac{2}{N^2}\sum_{i\in\mathbb{N}}\left(\nabla \mathcal{A}_{\cdot,i}(\mathbf{g}^{(1)})\mathcal{A}(\mathbf{g}^{(2)})\right)^2 + \frac{2}{N^2}\sum_{i\in\mathbb{N}}\left(\nabla \mathcal{A}_{\cdot,i}(\mathbf{g}^{(2)})\mathcal{A}(\mathbf{g}^{(1)})\right)^2 \\ &= \frac{2}{N^2}\left\|\nabla \mathcal{A}(\mathbf{g}^{(1)})\mathcal{A}(\mathbf{g}^{(2)})\right\|_2^2 + \frac{2}{N^2}\left\|\nabla \mathcal{A}(\mathbf{g}^{(2)})\mathcal{A}(\mathbf{g}^{(1)})\right\|_2^2.\end{aligned}$$

Since $\mathcal{A}(\mathbf{g}^{(1)}), \mathcal{A}(\mathbf{g}^{(2)}) \in B_N$, this implies

$$\left\|\nabla F(\mathbf{g}^{[0]},\mathbf{g}^{[1]},\mathbf{g}^{[2]})\right\|_2^2 \le \frac{2}{N}s_{\max}\left(\nabla \mathcal{A}(\mathbf{g}^{(1)})\right)^2 + \frac{2}{N}s_{\max}\left(\nabla \mathcal{A}(\mathbf{g}^{(2)})\right)^2 \le \frac{4\tau^2}{N}$$

for all $\mathbf{g}^{[0]},\mathbf{g}^{[1]},\mathbf{g}^{[2]} \in \mathbb{R}^{\mathbb{N}}$. The last inequality uses (C.1.4). By Gaussian concentration,

$$\mathbb{P}\left[\left|F(\mathbf{g}^{[0]},\mathbf{g}^{[1]},\mathbf{g}^{[2]}) - \mathbb{E}\,F(\mathbf{g}^{[0]},\mathbf{g}^{[1]},\mathbf{g}^{[2]})\right| \ge \lambda\right] \le \exp\left(-\frac{\lambda^2}{8\tau^2}N\right).$$

Note that Gaussian concentration of measure applies in infinite-dimensional abstract Weiner spaces as explained just before [Led01, Theorem 2.7] regarding Equation (2.10) therein. Alternatively if one wishes to avoid infinite-dimensional Gaussian measures, it suffices to prove the present proposition for the (still τ-Lipschitz) conditional expectations

$$\mathcal{A}_p(H_N) = \mathbb{E}[\mathcal{A}(H_N)|\mathbf{G}^{(2)},\ldots,\mathbf{G}^{(p)}]$$

and observe that $\lim_{p\to\infty} \mathcal{A}_p(H_N) = \mathcal{A}(H_N)$ holds almost surely and in L^1. □

Proposition C.1.3. *Suppose $\mathcal{A} : \mathcal{H}_N \to B_N$ is (τ, ν)-approximately Lipschitz. Then, for any $\lambda > 0$, it is $\left(\lambda, \exp\left(-\frac{(\lambda-4\nu)_+^2}{8\tau^2}N\right) + 2\nu\right)$ overlap concentrated.*

Proof. If $\lambda \le 4\nu$ the result is trivial, so suppose $\lambda > 4\nu$. Let $\mathcal{A}'$ be such that (C.1.3) holds.

Let $p \in [0,1]$, and let $H_N^{(1)}, H_N^{(2)}$ be p-correlated. We have

$$\left|R\left(\mathcal{A}(H_N^{(1)}), \mathcal{A}(H_N^{(2)})\right) - R\left(\mathcal{A}'(H_N^{(1)}), \mathcal{A}'(H_N^{(2)})\right)\right| \le 2$$

pointwise. Furthermore, (C.1.3) implies that

$$\mathcal{A}(H_N^{(1)}) = \mathcal{A}'(H_N^{(1)}) \qquad \text{and} \qquad \mathcal{A}(H_N^{(2)}) = \mathcal{A}'(H_N^{(2)}) \tag{C.1.8}$$

with probability at least $1 - 2\nu$. So,

$$\left|\mathbb{E}\, R\left(\mathcal{A}(H_N^{(1)}), \mathcal{A}(H_N^{(2)})\right) - \mathbb{E}\, R\left(\mathcal{A}'(H_N^{(1)}), \mathcal{A}'(H_N^{(2)})\right)\right| \le 4\nu. \tag{C.1.9}$$

By Proposition C.1.3, we have

$$\left|R\left(\mathcal{A}'(H_N^{(1)}), \mathcal{A}'(H_N^{(2)})\right) - \mathbb{E}\, R\left(\mathcal{A}'(H_N^{(1)}), \mathcal{A}'(H_N^{(2)})\right)\right| \le \lambda - 4\nu \tag{C.1.10}$$

with probability at least $1 - \exp\left(-\frac{(\lambda-4\nu)^2}{8\tau^2}N\right)$. The events (C.1.8) and (C.1.10) occur simultaneously with probability at least $1 - \exp\left(-\frac{(\lambda-4\nu)^2}{8\tau^2}N\right) - 2\nu$. On this event, (C.1.9) and (C.1.10) imply

$$\left|R\left(\mathcal{A}(H_N^{(1)}), \mathcal{A}(H_N^{(2)})\right) - \mathbb{E}\, R\left(\mathcal{A}(H_N^{(1)}), \mathcal{A}(H_N^{(2)})\right)\right| \le \lambda,$$

as desired. □

C.1.2 Standard Deterministic Optimization Algorithms are Approximately Lipschitz

Fix constants $T_0, T, k_0 \in \mathbb{N}$ and $r \in [1, \sqrt{2})$. We take as initialization a sequence $(\boldsymbol{x}^{-T_0}, \dots, \boldsymbol{x}^{-1})$ of vectors in B_N, which is independent of the Hamiltonian H_N. We consider rather general k_0-th order optimization algorithms which compute

$$\boldsymbol{x}^{t+1} = f_t\left(\left(\boldsymbol{x}^s\right)_{-T_0 \le s \le t}, \left(\nabla^k H_N(\boldsymbol{x}^s)\right)_{1 \le k \le k_0, -T_0 \le s \le t}\right), \quad 0 \le t \le T-1 \tag{C.1.11}$$

and output $\boldsymbol{x}^T$. Here, $(f_0, f_1, \dots, f_{T-1})$ is a deterministic sequence of functions such that $f_0, \dots, f_{T-2}$ have codomain rB_N, f_{T-1} has codomain B_N, and these functions are all Lipschitz in the sense that there exist constants $c_0, \dots, c_{T-1} > 0$ such that

$$\begin{aligned}&\left\| f_t\left((\boldsymbol{x}^s)_{-T_0 \le s \le t}, (A_k^s)_{1 \le k \le k_0, -T_0 \le s \le t} \right) - f_t\left((\boldsymbol{y}^s)_{-T_0 \le s \le t}, (B_k^s)_{1 \le k \le k_0, -T_0 \le s \le t} \right) \right\|_N \\ &\le c_t \left[\sum_{s=-T_0}^{t} \|\boldsymbol{x}^s - \boldsymbol{y}^s\|_N + \sum_{k=1}^{k_0} \sum_{s=-T_0}^{t} \|A_k^s - B_k^s\|_{\mathrm{op}} \right]. \end{aligned} \tag{C.1.12}$$

As we review below, the majority of standard convex optimization algorithms fall into this class. However we remark that some optimization algorithms for highly smooth and convex functions, such as Newton's method and the recent advances [GDG+19, Nes21], do not fall into this class. This is because they require inverting a Hessian matrix or solving another inverse problem each iteration.

Example C.1.1. Projected gradient descent is of the form in (C.1.11) via

$$f_k = \rho\left(\boldsymbol{x}^k - \eta_k \nabla H_N(\boldsymbol{x}^k)\right).$$

Here ρ is the projection map onto either B_N or C_N and the learning rate parameters $(\eta_1, \dots,)$ are arbitrary constants. Other variants such as accelerated gradient descent, ISTA, and FISTA (see e.g. [Bub15]) can similarly be expressed in the form (C.1.11).

Example C.1.2. Approximate message passing (AMP) with arbitrary Lipschitz non-linearities can be expressed in the form of (C.1.11). Given a deterministic sequence of Lipschitz functions $f_t : \mathbb{R}^{t+1} \to \mathbb{R}$ for each $t \ge 0$, the AMP iterates are given by

$$\boldsymbol{x}^{t+1} = \nabla \widetilde{H}_N(f_t(\boldsymbol{x}^0, \dots, \boldsymbol{x}^t)) - \sum_{s=1}^{t} d_{t,s} f_{s-1}(\boldsymbol{x}^0, \dots, \boldsymbol{x}^{s-1}), \tag{C.1.13}$$

$$d_{t,s} = \xi''\left(R\left(f_t(\boldsymbol{x}^0, \dots, \boldsymbol{x}^t), f_{s-1}(\boldsymbol{x}^0, \dots, \boldsymbol{x}^{s-1})\right)\right) \cdot \mathbb{E}\left[\frac{\partial f_t}{\partial X^s}(X^0, \dots, X^t)\right]. \tag{C.1.14}$$

Here $X^0 \sim p_0$ is a uniformly bounded random variable, and $\boldsymbol{x}^0$ has i.i.d. coordinates generated from the same law. The non-linearities f_t are applied entry-wise as functions $f_t : \mathbb{R}^{N \times (t+1)} \to \mathbb{R}^N$. The sequence $(X^t)_{t \ge 1}$ is an independent centered Gaussian process with covariance $Q_{t,s} = \mathbb{E}[X^t X^s]$ defined recursively by

$$Q_{t+1,s+1} = \xi'\left(\mathbb{E}\left[f_t\left(X^0, \dots, X^t\right) f_s\left(X^0, \dots, X^s\right)\right]\right), \quad t, s \ge 0. \tag{C.1.15}$$

It is not difficult to see that the iteration (C.1.13) is captured by (C.1.11), by defining the non-linearities $f_t(\boldsymbol{x}^0, \dots, \boldsymbol{x}^t)$ as additional iterates $\boldsymbol{x}^\ell$ so that their gradients can be evaluated.

Theorem 41. *For any functions $f_0, \dots, f_{T-1}$ as above and any initialization $(\boldsymbol{x}^{-T_0}, \dots, \boldsymbol{x}^{-1})$ of*

vectors in B_N, there exist constants τ, c such that the map $H_N \to \boldsymbol{x}^T$ defined by the iteration (C.1.11) *is (τ, ν)-approximately Lipschitz with $\nu = e^{-cN}$.*

Proof. We will first show the existence of τ such that the map $H_N \to \boldsymbol{x}^T$, with domain restricted to K_N (recall Proposition C.2.1), is τ-Lipschitz with respect to the $\|\cdot\|_N$ norms. Consider running the iteration (C.1.11) on two Hamiltonians $H_N, H'_N \in K_N$ with the same initializaton $(\boldsymbol{x}^{-T_0}, \dots, \boldsymbol{x}^{-1})$; call the respective iterates $\boldsymbol{x}^0, \dots, \boldsymbol{x}^T$ and $\boldsymbol{y}^0, \dots, \boldsymbol{y}^T$. A straightforward induction using Proposition C.2.2 and (C.1.12) gives constants $C_0, \dots, C_T$ such that for $0 \le t \le T$,

$$\left\|\boldsymbol{x}^t - \boldsymbol{y}^t\right\|_N \le C_t \|H_N - H'_N\|_N.$$

In particular, we may take $\tau = C_T$.

By Kirszbraun's theorem, there exists a τ-Lipschitz $\mathcal{A}'$ such that $\mathcal{A}(H_N) = \mathcal{A}'(H_N)$ for $H_N \in K_N$. By Proposition C.2.1, there exists c such that $\mathbb{P}(H_N \in K_N) \ge 1 - e^{-cN}$. Therefore $\mathcal{A}$ is (τ, ν)-approximately Lipschitz for $\nu = e^{-cN}$. □

The following corollary follows immediately from Theorem 41 and Proposition C.1.3.

Corollary C.1.4. *For any functions $f_0, \dots, f_{T-1}$ as above and any initialization $(\boldsymbol{x}^{-T_0}, \dots, \boldsymbol{x}^{-1})$ of vectors in B_N, for every $\lambda > 0$ there exists a constant c_λ such that for sufficiently large N, the map $H_N \to \boldsymbol{x}^T$ defined by the iteration* (C.1.11) *is $(\lambda, e^{-c_\lambda N})$ overlap concentrated.*

C.1.3 Reflected Langevin Dynamics are Approximately Lipschitz

Here we show that a natural version of Langevin dynamics, run for bounded time, is approximately Lipschitz for almost any realization of the driving Brownian motion and hence falls into the scope of our main results. The Langevin dynamics for a Hamiltonian H_N are given by the diffusion

$$\mathrm{d}X_t = \frac{\beta}{2}\nabla H_N \mathrm{d}t + \mathrm{d}B_t.$$

When X_t can range over all of space, the SDE above may explode to infinity in finite time. We therefore modify the naïve dynamics above by enforcing an inward-normal reflecting boundary for the convex body $\mathcal{K} = rB_N$ or $\mathcal{K} = rC_N$. We refer the reader to [Pil14] for the relevant definitions. In short, the result is a stochastic differential equation of the form

$$\mathrm{d}X_t = \frac{\beta}{2}\nabla H_N(X_t)\mathrm{d}t + \mathrm{d}B_t - v_t \mathrm{d}\ell_t. \tag{C.1.16}$$

Here ℓ_t is non-decreasing and only increases at times when $X_t \in \partial\mathcal{K}$. Meanwhile $v_t \in \mathbb{R}^N$ is contained in the outward normal cone of $X_t \in \partial\mathcal{K}$ for all t. Note that there may be several inequivalent choices

for such a reflected process; our results apply to any of these choices. The Langevin dynamics we consider consists of solving (C.1.16) for a constant time T starting from X_0 which is independent of H_N, and then projecting X_T onto B_N or C_N.

The corresponding Skorokhod problem was shown to have a Lipschitz solution for convex polyhedra such as rC_N in [DI91, Proposition 2.2]. In this case, solving (C.1.16) reduces to solving an SDE with Lipschitz coefficients as explained in [Pil14, Section 2.2]. As a result, the solutions to (C.1.16) from different starting points X_0 (but with a shared Brownian motion) can be coupled together to give a continuous stochastic flow (see [RW94, Chapter 5, Section 13]). In the case of a smooth boundary such as B_N, although the Skorokhod problem does not have a Lipschitz solution, the results of [LS84] imply the existence of a stochastic flow as explained in [Bur09].

Lemma C.1.5. *Let X_t, Y_t solve* (C.1.16) *inside a convex body $\mathcal{K}$ with the same Brownian motion. Then*

$$\int_0^t \left\langle X_t - Y_t, v_t^X \mathrm{d}\ell_t^X - v_t^Y \mathrm{d}\ell_t^Y \right\rangle \geq 0.$$

Here (v_t^X, ℓ_t^X) denote the reflecting boundary terms for X_t and similarly for Y_t.

Proof. Recall that ℓ_t^X, ℓ_t^Y are increasing. Moreover $\langle X_t - Y_t, v_t^X \rangle \geq 0$ whenever $X_t \in \partial\mathcal{K}$ by the definition of the normal cone, and similarly $\langle Y_t - X_t, v_t^Y \rangle \geq 0$ whenever $Y_t \in \partial\mathcal{K}$. The result follows. □

Theorem 42. *Both variants of Langevin dynamics above define, for any initialization $X_0 \in B_N$ and for almost every path $(B_t)_{t\in[0,T]}$, a (τ, ν) approximately Lipschitz map $\mathcal{A} : \mathcal{H}_N \to B_N$ with $\tau = O_{\xi,h,T}(1)$ and $\nu \leq e^{-\Omega(N)}$.*

Proof. Fix Hamiltonians

$$H_N^X, H_N^Y \in K_N \subseteq \mathcal{H}_N$$

satisfying $\left\|H_N^X - H_N^Y\right\|_N = \Delta$. Let X_t, Y_t be the solutions to (C.1.16) driven by a shared Brownian motion with H_N^X and H_N^Y for H_N respectively, and with shared initial condition $X_0 = Y_0$. We will show that

$$\|X_T - Y_T\|_N \leq C\Delta$$

holds almost surely for some constant $C = C(\xi, h, T)$. This suffices to imply the result. (Note that $\mathcal{A}$ might not be defined on all of $\mathcal{H}_N$, but it suffices for it to be well-defined and Lipschitz on K_N.)

First observe that $X_t - Y_t$ is a finite variation process, i.e. it has no Brownian component. With

ℓ^X and ℓ^Y the corresponding finite variation processes in (C.1.16), Ito's formula gives

$$\begin{aligned}\frac{1}{2}\mathrm{d}\|X_t - Y_t\|_2^2 &= \langle X_t - Y_t, \mathrm{d}X_t - \mathrm{d}Y_t\rangle\,\mathrm{d}t \\ &= \langle X_t - Y_t, -v_t^X \mathrm{d}\ell_t^X + v_t^Y \mathrm{d}\ell_t^Y\rangle \mathrm{d}t + \beta\langle X_t - Y_t, \nabla H_N^X(X_t) - \nabla H_N^Y(Y_t)\rangle \mathrm{d}t.\end{aligned}$$

Integrating and using Lemma C.1.5, we find

$$\begin{aligned}\|X_t - Y_t\|_2^2 &\le \int_0^t \langle X_s - Y_s, -v_s^X \mathrm{d}\ell_s^X + v_s^Y \mathrm{d}\ell_s^Y\rangle \mathrm{d}s + \beta\int_0^t \langle X_s - Y_s, \nabla H_N^X(X_s) - \nabla H_N^Y(Y_s)\rangle \mathrm{d}s \\ &\le \beta\int_0^t \langle X_s - Y_s, \nabla H_N^X(X_s) - \nabla H_N^Y(Y_s)\rangle \mathrm{d}s.\end{aligned}$$

By Proposition C.2.2 with $C = C_1'$,

$$\left\|\nabla H_N^X(X_t) - \nabla H_N^Y(Y_t)\right\|_N \le C(\Delta + \|X_t - Y_t\|_N).$$

Using AM-GM and rescaling, we obtain for each $t \in [0, T]$ the self-bounding inequality

$$\begin{aligned}\|X_t - Y_t\|_N^2 &\le C\int_0^t \Delta\|X_s - Y_s\|_N + \|X_s - Y_s\|_N^2 \mathrm{d}t \\ &\le 2C\int_0^t \Delta^2 + \|X_s - Y_s\|_N^2 \mathrm{d}t \\ &\le 2C\Delta^2 T + 2C\int_0^t \|X_s - Y_s\|_N^2 \mathrm{d}t.\end{aligned}$$

Grönwall's inequality now implies $\|X_T - Y_T\|_N^2 \le 2C\Delta^2 T e^{2CT}$. This concludes the proof. □

C.2 Bounds on Hamiltonian Derivatives

In this section we will prove high-probability bounds on the derivatives of H_N, including Proposition 5.2.3. We write $H_N(\boldsymbol{\sigma}) = \langle \boldsymbol{h}, \boldsymbol{\sigma}\rangle + \widetilde{H}_N(\boldsymbol{\sigma})$ for

$$\widetilde{H}_N(\boldsymbol{\sigma}) = \sum_{p \in 2\mathbb{N}} \gamma_p H_{N,p}(\boldsymbol{\sigma}),$$

where the p-tensor component is

$$H_{N,p}(\boldsymbol{\sigma}) = \frac{1}{N^{(p-1)/2}}\langle \mathbf{G}^{(p)}, \boldsymbol{\sigma}^{\otimes p}\rangle.$$

By slight abuse of notation, we also denote $\frac{1}{N^{(p-1)/2}}\mathbf{G}^{(p)} = H_{N,p}$.

Proposition C.2.1. *There exists universal constants $c, C > 0$ such that for all sufficiently large N,*

$$\|H_{N,p}\|_{\mathrm{op}} \le C\sqrt{p}$$

for all $p \in 2\mathbb{N}$ with probability at least $1 - e^{-cN}$

Proof. By [BASZ20, Equation B.6] with $k = p$, we have for some universal constant K and all $p \in 2\mathbb{N}$,

$$\mathbb{P}\left[\|H_{N,p}\|_{\mathrm{op}} \ge 2K\sqrt{p}\right] \le e^{-K^2 pN/2}.$$

Take $C = 2K$. The result follows by a union bound over $p \in 2\mathbb{N}$. □

Proof of Proposition 5.2.3. Let $K_N \subseteq \mathcal{H}_N$ be the set of Hamiltonians H_N satisfying the conclusion of Proposition C.2.1. We will take

$$C_k = C \sum_{p \in 2\mathbb{N}, p \ge k} \gamma_p r^{p-k} p^{\underline{k}} \sqrt{p} + h\mathbb{I}\{k = 1\},$$

where C is given by Proposition C.2.1 and $p^{\underline{k}} = p(p-1)\cdots(p-k+1)$ denotes the k-th falling power of p. This is finite because $r < \sqrt{2}$ and $\sum_{p \in 2\mathbb{N}} \gamma_p^2 2^p < \infty$ implies $\limsup_{p\to\infty} \frac{\gamma_{p+2}}{\gamma_p} \le \frac{1}{2}$.

If $H_N \in K_N$, for each $\boldsymbol{\sigma}^1, \dots, \boldsymbol{\sigma}^k \in S_N$ we have

$$\begin{aligned}
\frac{1}{N}\left\langle \nabla^k \widetilde{H}_N(\boldsymbol{x}), \boldsymbol{\sigma}^1 \otimes \cdots \otimes \boldsymbol{\sigma}^k \right\rangle &= \sum_{p \in 2\mathbb{N}, p \ge k} \frac{\gamma_p}{N} \left\langle \nabla^k H_{N,p}(\boldsymbol{x}), \boldsymbol{\sigma}^1 \otimes \cdots \otimes \boldsymbol{\sigma}^k \right\rangle \\
&= \sum_{p \in 2\mathbb{N}, p \ge k} \frac{\gamma_p p^{\underline{k}}}{N} \left\langle H_{N,p}, \boldsymbol{x}^{\otimes(p-k)} \otimes \boldsymbol{\sigma}^1 \otimes \cdots \otimes \boldsymbol{\sigma}^k \right\rangle \\
&\le \sum_{p \in 2\mathbb{N}, p \ge k} \gamma_p r^{p-k} p^{\underline{k}} \|H_{N,p}\|_{\mathrm{op}} \\
&\le C \sum_{p \in 2\mathbb{N}, p \ge k} \gamma_p r^{p-k} p^{\underline{k}} \sqrt{p},
\end{aligned}$$

by Proposition C.2.1. Thus

$$\left\|\nabla^k \widetilde{H}_N(\boldsymbol{x})\right\|_{\mathrm{op}} \le C \sum_{p \in 2\mathbb{N}, p \ge k} \gamma_p r^{p-k} p^{\underline{k}} \sqrt{p}.$$

For $k \ge 2$, $\nabla^k H_N(\boldsymbol{x}) = \nabla^k \widetilde{H}_N(\boldsymbol{x})$, and for $k = 1$, $\|\nabla H_N(\boldsymbol{x})\|_{\mathrm{op}} \le \left\|\nabla \widetilde{H}_N(\boldsymbol{x})\right\|_{\mathrm{op}} + h$. This proves

the first claim. Similarly,

$$\begin{aligned}
&\frac{1}{N}\left\langle \nabla^k H_N(\boldsymbol{x}) - \nabla^k H_N(\boldsymbol{y}), \boldsymbol{\sigma}^1 \otimes \cdots \otimes \boldsymbol{\sigma}^k \right\rangle \\
&= \sum_{p\in 2\mathbb{N}, p\geq k} \frac{\gamma_p}{N}\left\langle \nabla^k H_{N,p}(\boldsymbol{x}) - \nabla^k H_{N,p}(\boldsymbol{y}), \boldsymbol{\sigma}^1 \otimes \cdots \otimes \boldsymbol{\sigma}^k \right\rangle \\
&= \sum_{p\in 2\mathbb{N}, p\geq k} \frac{\gamma_p p^{\underline{k}}}{N}\left\langle H_{N,p}, \left(\boldsymbol{x}^{\otimes(p-k)} - \boldsymbol{y}^{\otimes(p-k)}\right) \otimes \boldsymbol{\sigma}^1 \otimes \cdots \otimes \boldsymbol{\sigma}^k \right\rangle \\
&= \sum_{p\in 2\mathbb{N}, p\geq k} \frac{\gamma_p p^{\underline{k}}}{N} \sum_{j=0}^{p-k-1} \left\langle H_{N,p}, (\boldsymbol{x}-\boldsymbol{y}) \otimes \boldsymbol{x}^{\otimes(p-k-1-j)} \otimes \boldsymbol{y}^{\otimes j} \otimes \boldsymbol{\sigma}^1 \otimes \cdots \otimes \boldsymbol{\sigma}^k \right\rangle \\
&\leq \sum_{p\in 2\mathbb{N}, p\geq k} \gamma_p r^{p-k-1} p^{\underline{k}} (p-k) \|\boldsymbol{x}-\boldsymbol{y}\|_N \|H_{N,p}\|_{\mathrm{op}} \\
&\leq C_{k+1}\|\boldsymbol{x}-\boldsymbol{y}\|_N,
\end{aligned}$$

so $\left\|\nabla^k H_N(\boldsymbol{x}) - \nabla^k H_N(\boldsymbol{y})\right\|_{\mathrm{op}} \leq C_{k+1}\|\boldsymbol{x}-\boldsymbol{y}\|_N$, proving the second claim. □

Proposition C.2.2. *Fix a model (ξ, h) and a constant $r \in [1, \sqrt{2})$. Let K_N be given by Proposition 5.2.3. There exists a sequence of constants $(C'_k)_{k\geq 1}$ independent of N such that for all $H_N, H'_N \in K_N$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$ with $\|\boldsymbol{x}\|_N, \|\boldsymbol{y}\|_N \leq r$,*

$$\left\|\nabla^k H_N(\boldsymbol{x}) - \nabla^k H'_N(\boldsymbol{y})\right\|_{\mathrm{op}} \leq C'_k\left(\|\boldsymbol{x}-\boldsymbol{y}\|_N + \|H_N - H'_N\|_N\right),$$

where $\|H_N - H'_N\|_N$ is defined by (C.1.1).

Note that when $\|H_N - H'_N\|_N$ is infinite, this proposition is vacuously true.

Proof. We have that

$$\left\|\nabla^k H_N(\boldsymbol{x}) - \nabla^k H'_N(\boldsymbol{y})\right\|_{\mathrm{op}} \leq \left\|\nabla^k H_N(\boldsymbol{x}) - \nabla^k H'_N(\boldsymbol{x})\right\|_{\mathrm{op}} + \left\|\nabla^k H'_N(\boldsymbol{x}) - \nabla^k H'_N(\boldsymbol{y})\right\|_{\mathrm{op}},$$

and by (5.2.8),

$$\left\|\nabla^k H'_N(\boldsymbol{x}) - \nabla^k H'_N(\boldsymbol{y})\right\|_{\mathrm{op}} \leq C_{k+1}\|\boldsymbol{x}-\boldsymbol{y}\|_N.$$

For all $\boldsymbol{\sigma}^1, \ldots, \boldsymbol{\sigma}^k \in S_N$, we have

$$\begin{aligned}
\frac{1}{N}\left\langle \nabla^k H_N(\boldsymbol{x}) - \nabla^k H'_N(\boldsymbol{x}), \boldsymbol{\sigma}^1 \otimes \cdots \otimes \boldsymbol{\sigma}^k \right\rangle &= \sum_{p\in 2\mathbb{N}, p\geq k} \frac{\gamma_p}{N}\left\langle \nabla^k H_{N,p}(\boldsymbol{x}) - \nabla^k H'_{N,p}(\boldsymbol{x}), \boldsymbol{\sigma}^1 \otimes \cdots \otimes \boldsymbol{\sigma}^k \right\rangle \\
&= \sum_{p\in 2\mathbb{N}, p\geq k} \frac{\gamma_p p^{\underline{k}}}{N}\left\langle H_{N,p} - H'_{N,p}, \boldsymbol{x}^{\otimes(p-k)} \otimes \boldsymbol{\sigma}^1 \otimes \cdots \otimes \boldsymbol{\sigma}^k \right\rangle \\
&\leq \sum_{p\in 2\mathbb{N}, p\geq k} \gamma_p r^{p-k} p^{\underline{k}} \left\|H_{N,p} - H'_{N,p}\right\|_{\mathrm{op}}.
\end{aligned}$$

Moreover,

$$\begin{aligned}\left\|H_{N,p} - H'_{N,p}\right\|_{\mathrm{op}} &= \frac{1}{N^{(p+1)/2}} \max_{\boldsymbol{\sigma}^1,\dots,\boldsymbol{\sigma}^p \in S_N} \langle \mathbf{G}^{(p)} - \mathbf{G}'^{(p)}, \boldsymbol{\sigma}^1 \otimes \cdots \otimes \boldsymbol{\sigma}^p \rangle \\ &\le \frac{1}{\sqrt{N}} \left\| \mathbf{G}^{(p)} - \mathbf{G}'^{(p)} \right\|_2 \\ &\le \|H_N - H'_N\|_N.\end{aligned}$$

Thus we have

$$\frac{1}{N} \left\langle \nabla^k H_N(\boldsymbol{x}) - \nabla^k H'_N(\boldsymbol{x}), \boldsymbol{\sigma}^1 \otimes \cdots \otimes \boldsymbol{\sigma}^k \right\rangle \le \sum_{p \in 2\mathbb{N}, p \ge k} \gamma_p r^{p-k} p^{\underline{k}} \cdot \|H_N - H'_N\|_N.$$

Because this holds for all $\boldsymbol{\sigma}^1, \dots, \boldsymbol{\sigma}^k \in S_N$, we have

$$\left\| \nabla^k H_N(\boldsymbol{x}) - \nabla^k H'_N(\boldsymbol{x}) \right\|_{\mathrm{op}} \le \sum_{p \in 2\mathbb{N}, p \ge k} \gamma_p r^{p-k} p^{\underline{k}} \cdot \|H_N - H'_N\|_N.$$

The result follows by taking C'_k to be the larger of C_{k+1} and $\sum_{p \in 2\mathbb{N}, p \ge k} \gamma_p r^{p-k} p^{\underline{k}}$. □

C.3 Explicit Formula for the Spherical Algorithmic Threshold

In this section, we will prove Proposition 5.2.2, which gives an explicit formula for $\mathsf{ALG}^{\mathrm{Sp}}_{\xi,h}$.

We first remark that the $\widehat{q}$ defined in the second case of Proposition 5.2.2 exists and is unique. Define $f(q) = q\xi''(q) - \xi'(q) = \sum_{p \in 2\mathbb{N}} p(p-2)\gamma_p^2 q^{p-1}$. If we are in the second case of the proposition, then $h^2 + \xi'(1) < \xi''(1)$, so $f(1) > h^2$. Since $f(0) = 0 \le h^2$, existence of $\widehat{q}$ follows from the Intermediate Value Theorem. Moreover, $f(1) > h^2 \ge 0$ implies $\gamma_p > 0$ for some $p > 2$, so $f(q)$ is strictly increasing for $q \in [0,1]$. This implies uniqueness.

Recall that the spherical Parisi functional P^{Sp} (5.2.2) is defined in terms of a function $B_\zeta(t) = B - \int_t^1 \xi''(q)\zeta(q)\,\mathrm{d}q$. As (B,ζ) ranges over $\mathscr{K}(\xi)$, $B_\zeta(t)$ ranges over all continuous, nondecreasing functions from $[0,1]$ to $\mathbb{R}_{>0}$. We can thus reparametrize the minimizaton (5.2.5) as one over continuous and nondecreasing $B : [0,1] \to \mathbb{R}_{>0}$. By slight abuse of notation, for continuous and nondecreasing $B : [0,1] \to \mathbb{R}_{>0}$ define

$$\mathsf{P}^{\mathrm{Sp}}(B) = \mathsf{P}^{\mathrm{Sp}}_{\xi,h}(B) = \frac{1}{2}\left[\frac{h^2}{B(0)} + \int_0^1 \left(\frac{\xi''(q)}{B(q)} + B(q)\right)\,\mathrm{d}q\right]$$

Proof of Proposition 5.2.2. We first handle the case $h = 0$. By AM-GM,

$$\mathsf{P}^{\mathrm{Sp}}(B) = \frac{1}{2}\int_0^1 \left(\frac{\xi''(q)}{B(q)} + B(q)\right)\,\mathrm{d}q \geq \int_0^1 \xi''(q)^{1/2}\,\mathrm{d}q.$$

Equality holds when $B(q) = \xi''(q)^{1/2}$ for all $q \in [0,1]$. However, this requires $B(0) = 0$, so this objective is not attained, though approximations to this B get arbitrarily close. Thus $\mathsf{ALG}^{\mathrm{Sp}} = \int_0^1 \xi''(q)^{1/2}\,\mathrm{d}q$. We will show this $\mathsf{ALG}^{\mathrm{Sp}}$ equals the value claimed. If $\gamma_p > 0$ for some $p > 2$, then $\xi'(1) < \xi''(1)$, so we are in the second case of the proposition. Since $\widehat{q} = 0$, we are done. Otherwise, $\gamma_p = 0$ for all $p > 2$, and $\xi'(1) = \xi''(1)$. Then $\xi''(q)$ is constant, so $\mathsf{ALG}^{\mathrm{Sp}} = \xi''(1)^{1/2} = \xi'(1)^{1/2}$ as claimed.

Otherwise, $h > 0$. We extend the definition of $\widehat{q}$ to

$$\widehat{q} = \sup\left\{q \in [0,1] : h^2 + \xi'(q) \geq q\xi''(q)\right\}.$$

This gives $\widehat{q} = 1$ in the first case of the proposition, and matches the definition of $\widehat{q}$ in the second case. Note that $\widehat{q} > 0$. Define

$$\widehat{B} = \left(\frac{h^2 + \xi'(\widehat{q})}{\widehat{q}}\right)^{1/2}.$$

We will prove both cases simultaneously by showing that for any continuous and nondecreasing $B : [0,1] \to \mathbb{R}_{>0}$, we have

$$\mathsf{P}^{\mathrm{Sp}}(B) \geq \widehat{q}^{1/2}\left(h^2 + \xi'(\widehat{q})\right)^{1/2} + \int_{\widehat{q}}^1 \xi''(q)^{1/2}\,\mathrm{d}q,$$

with equality if and only if

$$B(q) = \begin{cases} \widehat{B} & q \leq \widehat{q}, \\ \xi''(q)^{1/2} & q > \widehat{q}. \end{cases}$$

It is easy to check that this B is continuous and nondecreasing (i.e. if $\widehat{q} < 1$, then $\widehat{B} = \xi''(\widehat{q})^{1/2}$) and that it corresponds to the equality cases claimed in the proposition. By AM-GM,

$$\frac{1}{2}\int_{\widehat{q}}^1 \left(\frac{\xi''(q)}{B(q)} + B(q)\right)\,\mathrm{d}q \geq \int_{\widehat{q}}^1 \xi''(q)^{1/2}\,\mathrm{d}q, \tag{C.3.1}$$

with equality if and only if $B(q) = \xi''(q)^{1/2}$ on $(\widehat{q}, 1]$. Define the truncated Parisi operator

$$\mathsf{P}^{\mathrm{Sp},\widehat{q}}(B) = \frac{1}{2}\left[\frac{h^2}{B(0)} + \int_0^{\widehat{q}} \left(\frac{\xi''(q)}{B(q)} + B(q)\right)\,\mathrm{d}q\right].$$

Let $\widetilde{B} : [0, \widehat{q}] \to \mathbb{R}_{>0}$ be given by $\widetilde{B}(q) = \widehat{B}$, and note that $\mathsf{P}^{\mathrm{Sp},\widehat{q}}(\widetilde{B}) = \widehat{q}^{1/2}\left(h^2 + \xi'(\widehat{q})\right)^{1/2}$. We will

show that for continuous and nondecreasing $B : [0, \widehat{q}] \to \mathbb{R}_{>0}$, we have $\mathsf{P}^{\mathrm{Sp},\widehat{q}}(B) \geq \mathsf{P}^{\mathrm{Sp},\widehat{q}}(\widetilde{B})$, with equality if and only if $B \equiv \widetilde{B}$ on $[0, \widehat{q}]$. Along with (C.3.1), this implies the conclusion. We consider two cases.

Case 1: $B(0) < \widehat{B}$**.** Define

$$\widetilde{q} = \sup\left\{q \in [0, \widehat{q}] : B(q) \leq \widehat{B}\right\}.$$

It is possible that $\widetilde{q} = \widehat{q}$. For $q \in [\widetilde{q}, \widehat{q}]$, we have $B(q) \geq \widehat{B}$, so

$$\int_{\widetilde{q}}^{\widehat{q}} \left(\frac{\xi''(q)}{B(q)} + B(q)\right) - \int_{\widetilde{q}}^{\widehat{q}} \left(\frac{\xi''(q)}{\widehat{B}} + \widehat{B}\right) = \int_{\widetilde{q}}^{\widehat{q}} \left(\frac{1}{\widehat{B}} - \frac{1}{B(q)}\right)\left(B(q)\widehat{B} - \xi''(q)\right)\,\mathrm{d}q.$$

Because

$$B(q)\widehat{B} \geq \widehat{B}^2 \geq \frac{h^2 + \xi'(\widehat{q})}{\widehat{q}} \geq \xi''(\widehat{q}) \geq \xi''(q),$$

we have

$$\int_{\widetilde{q}}^{\widehat{q}} \left(\frac{\xi''(q)}{B(q)} + B(q)\right) \geq \int_{\widetilde{q}}^{\widehat{q}} \left(\frac{\xi''(q)}{\widehat{B}} + \widehat{B}\right). \tag{C.3.2}$$

Moreover, for $q \in [0, \widetilde{q}]$, we have $B(q) \leq \widehat{B}$, so

$$\begin{aligned}
2\left(\mathsf{P}^{\mathrm{Sp},\widetilde{q}}(B) - \mathsf{P}^{\mathrm{Sp},\widetilde{q}}(\widetilde{B})\right) &= h^2\left(\frac{1}{B(0)} - \frac{1}{\widehat{B}}\right) - \int_0^{\widetilde{q}} \left(B(q)\widehat{B} - \xi''(q)\right)\left(\frac{1}{B(q)} - \frac{1}{\widehat{B}}\right)\,\mathrm{d}q \\
&\geq h^2\left(\frac{1}{B(0)} - \frac{1}{\widehat{B}}\right) - \int_0^{\widetilde{q}} \left(\widehat{B}^2 - \xi''(q)\right)\left(\frac{1}{B(q)} - \frac{1}{\widehat{B}}\right)\,\mathrm{d}q \\
&= h^2\left(\frac{1}{B(0)} - \frac{1}{\widehat{B}}\right) - \int_0^{\widetilde{q}} \left(\frac{h^2 + \xi'(\widehat{q})}{\widehat{q}} - \xi''(q)\right)\left(\frac{1}{B(q)} - \frac{1}{\widehat{B}}\right)\,\mathrm{d}q \\
&\geq h^2\left(\frac{1}{B(0)} - \frac{1}{\widehat{B}}\right) - \int_0^{\widetilde{q}} \left(\frac{h^2 + \xi'(\widehat{q})}{\widehat{q}} - \xi''(q)\right)\left(\frac{1}{B(0)} - \frac{1}{\widehat{B}}\right)\,\mathrm{d}q \\
&= \left(\frac{1}{B(0)} - \frac{1}{\widehat{B}}\right)\left[h^2 - \int_0^{\widetilde{q}} \left(\frac{h^2 + \xi'(\widehat{q})}{\widehat{q}} - \xi''(q)\right)\,\mathrm{d}q\right] \\
&\geq \left(\frac{1}{B(0)} - \frac{1}{\widehat{B}}\right)\left[h^2 - \int_0^{\widehat{q}} \left(\frac{h^2 + \xi'(\widehat{q})}{\widehat{q}} - \xi''(q)\right)\,\mathrm{d}q\right] \\
&= 0.
\end{aligned}$$

Thus $\mathsf{P}^{\mathrm{Sp},\widetilde{q}}(B) \geq \mathsf{P}^{\mathrm{Sp},\widetilde{q}}(\widetilde{B})$, with equality only if $\widetilde{q} = \widehat{q}$ and $B(q) = \widehat{B}$ for all $q \in [0, \widetilde{q}]$. Combining this with (C.3.2) gives that $\mathsf{P}^{\mathrm{Sp},\widehat{q}}(B) \geq \mathsf{P}^{\mathrm{Sp},\widehat{q}}(\widetilde{B})$, with equality only if $B \equiv \widetilde{B}$ on $[0, \widehat{q}]$.

Case 2: $B(0) \geq \widehat{B}$**.** In this case, $B(q) \geq \widehat{B}$ for all $q \in [0, \widehat{q}]$. So,

$$\begin{aligned}
2\left(\mathsf{P}^{\mathrm{Sp},\widehat{q}}(B) - \mathsf{P}^{\mathrm{Sp},\widehat{q}}(\widetilde{B})\right) &= -h^2\left(\frac{1}{\widehat{B}} - \frac{1}{B(0)}\right) + \int_0^{\widehat{q}} \left(B(q)\widehat{B} - \xi''(q)\right)\left(\frac{1}{\widehat{B}} - \frac{1}{B(q)}\right)\,\mathrm{d}q \\
&\geq -h^2\left(\frac{1}{\widehat{B}} - \frac{1}{B(0)}\right) + \int_0^{\widehat{q}} \left(\widehat{B}^2 - \xi''(q)\right)\left(\frac{1}{\widehat{B}} - \frac{1}{B(q)}\right)\,\mathrm{d}q \\
&= -h^2\left(\frac{1}{\widehat{B}} - \frac{1}{B(0)}\right) + \int_0^{\widehat{q}} \left(\frac{h^2 + \xi'(\widehat{q})}{\widehat{q}} - \xi''(q)\right)\left(\frac{1}{\widehat{B}} - \frac{1}{B(q)}\right)\,\mathrm{d}q \\
&\geq -h^2\left(\frac{1}{\widehat{B}} - \frac{1}{B(0)}\right) + \int_0^{\widehat{q}} \left(\frac{h^2 + \xi'(\widehat{q})}{\widehat{q}} - \xi''(q)\right)\left(\frac{1}{\widehat{B}} - \frac{1}{B(0)}\right)\,\mathrm{d}q \\
&= \left(\frac{1}{\widehat{B}} - \frac{1}{B(0)}\right)\left[-h^2 + \int_0^{\widehat{q}} \left(\frac{h^2 + \xi'(\widehat{q})}{\widehat{q}} - \xi''(q)\right)\,\mathrm{d}q\right] \\
&= 0.
\end{aligned}$$

For equality to hold, we must have $B(q) = \widetilde{B}$ for all $q \in [0, \widehat{q}]$, so $B \equiv \widetilde{B}$ on $[0, \widehat{q}]$. □