Statistics 291: Lecture 19 (April 02, 2024)

Tight Hardness for Optimizing Spherical Spin Glasses via Branching OGP

Instructor: Mark Sellke

Scribe: Kevin Luo

This lecture takes material from [1, 2].

1 Preliminaries

For a pure p-spin model, recall that the following quantity is fundamental:

$$\mathsf{ALG}(p) = E_\infty(p) = 2\sqrt{\frac{p-1}{p}}. \tag{1}$$

We saw in our Kac-Rice analysis of these models that this is the energy level at which local maxima begin to appear, suggesting that it is a limit to optimization algorithms; furthermore, we saw that Subag's Hessian Ascent algorithm achieves exactly this value. For general covariances ξ, Subag's algorithm achieves

$$\mathsf{ALG}(\xi) = \int_0^1 \sqrt{\xi''(t)}\,\mathrm{d}t. \tag{2}$$

As we discussed, sometimes $\mathsf{ALG}(\xi) = \mathsf{OPT}(\xi)$, where $\mathsf{OPT}(\xi)$ denotes the true ground state energy. In fact this is equivalent to being full RSB at zero temperature.

Our goal today will be to prove the following theorem:

Theorem 1.1. *Suppose $\mathscr{A}_n : H_n \mapsto \boldsymbol{\sigma}^{\mathsf{alg}} \in \mathscr{B}_N = \{\boldsymbol{x} \in \mathbb{R}^N : \|\boldsymbol{x}\| \le \sqrt{N}\}$ is L-Lipschitz for all N. Then for any $\epsilon > 0$:*

$$\mathbb{E} H_N(\mathscr{A}_N(H_N))/N \le \mathsf{ALG}(\xi) + \epsilon \tag{3}$$

for N large enough.

Recall that this covers algorithms such as gradient descent as well as Langevin Dynamics, when both are run for $O(1)$ time.

2 Branching OGP

2.1 Tree Structure

Our idea will be similar to the proofs we studied in the previous week. We will show that $\mathscr{A}$ maps a set of correlated Hamiltonians to some forbidden structure. Figure 1 summarizes the types of correlation structures employed in the proofs for Max Independent Set.

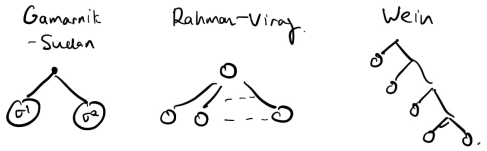

Figure 1: Other structures previously considered for Max Independent Set

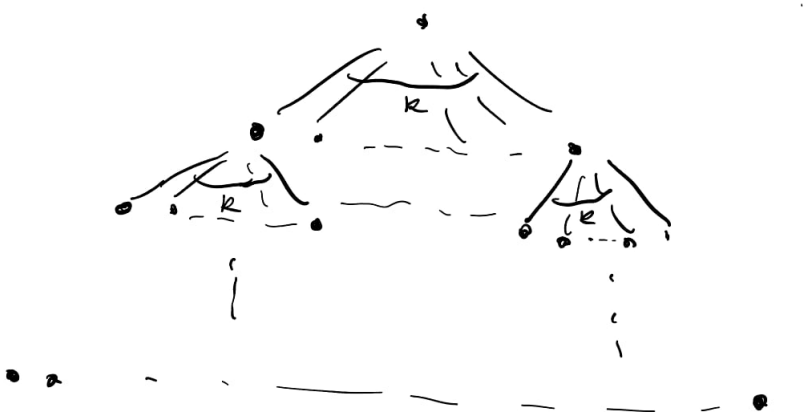

Figure 2: The tree for this lecture.

We will instead use a really big tree. See Figure 2.

Definition 2.1. (H_N, H'_N) are p-correlated, for $0 \le p \le 1$, if their joint law agrees with

$$H_N = \sqrt{p}\widetilde{H_{N,0}} + \sqrt{1-p}\widetilde{H_{N,1}} \tag{4}$$

$$H'_N = \sqrt{p}\widetilde{H_{N,0}} + \sqrt{1-p}\widetilde{H_{N,2}}, \tag{5}$$

where $\widetilde{H_{N,j}}$ are iid spherical models with covariance function ξ.

Definition 2.2. Define

$$\chi(p) = \chi_{\mathscr{A}}(p) = \mathbb{E}R(\underbrace{\mathscr{A}(H_N)}_{\boldsymbol{\sigma}^{\mathsf{alg}}}, \underbrace{\mathscr{A}(H'_N)}_{(\boldsymbol{\sigma}')^{\mathsf{alg}}}). \tag{6}$$

Proposition 2.3. *Uniformly in p and L-Lipschitz $\mathscr{A}$, we have*

$$\mathbb{P}\left(\left|R(\boldsymbol{\sigma}^{\mathsf{alg}}, (\boldsymbol{\sigma}')^{\mathsf{alg}}) - \chi_{\mathscr{A}}(p)\right| \ge \epsilon\right) \le 2\exp\left(-\Omega(\epsilon^2 N)\right). \tag{7}$$

Proof. This is a consequence of Gaussian Lipschitz concentration. Consider the sequence of maps

$$\begin{pmatrix}\widetilde{H_{N,0}}\\ \widetilde{H_{N,1}}\\ \widetilde{H_{N,2}}\end{pmatrix} \xrightarrow{O(1)} \begin{pmatrix}H_N\\ H'_N\end{pmatrix} \xrightarrow[\mathscr{A}]{L} \begin{pmatrix}\boldsymbol{\sigma}^{\mathsf{alg}}\\ (\boldsymbol{\sigma}')^{\mathsf{alg}}\end{pmatrix} \xrightarrow{O(1/\sqrt{N})} R(\boldsymbol{\sigma}^{\mathsf{alg}}, (\boldsymbol{\sigma}')^{\mathsf{alg}}), \tag{8}$$

where the number above each map denotes the Lipschitz constant. The last map has constant $O(1/\sqrt{N})$ because each vector has norm $\sqrt{N}$, but we divide their inner product by N. The Lipschitz constants multiply and we are done. □

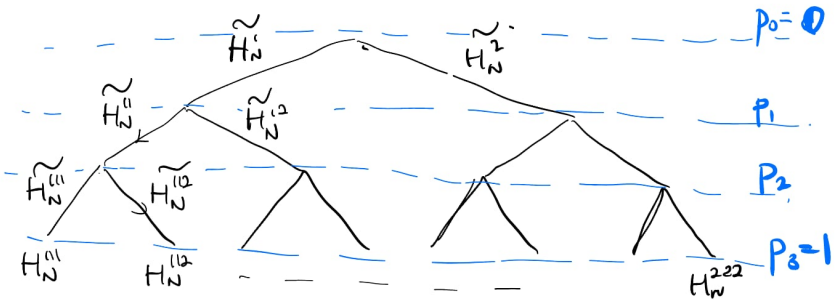

Figure 3: Correlation structure.

The correlation structure we study is as follows. Fix constants $0 = p_0 \le p_1 \le p_2 \le \cdots \le p_D = 1$. The leaf nodes are defined by summing down the path from the root of the tree. For example,

$$H_N^{112} = \sqrt{p_1}\widetilde{H_N^1} + \sqrt{p_2 - p_1}\widetilde{H_N^{11}} + \sqrt{p_3 - p_2}\widetilde{H_N^{112}}, \tag{9}$$

We define the interior points in the same way, e.g.

$$H_N^{21} = \sqrt{p_1}\widetilde{H_N^2} + \sqrt{p_2 - p_1}\widetilde{H_N^{21}}. \tag{10}$$

Note that the interior points do not have Hamiltonians with the correct scale.

In general, we will actually need D levels, where each node has k children. So for two sequences $u, v \in [k]^D$, we define $u \wedge v \in [0, \dots, D]$ be the largest d such that $u_1 = v_1, \dots, u_d = v_d$.

Note that by construction, we now have that (H_N^u, H_N^v) are $p_{u\wedge v}$ correlated. We now want to map all of the leaves of the tree under $\mathscr{A}$ and try to show that if $\mathscr{A}$ achieves energies that are too large, that the resulting structure cannot exist.

2.2 Forbidden Structure

The image of the tree above under $\mathscr{A}$ is a tree in it's own right.

Proposition 2.3 now implies that for any two leaf nodes u and v,

$$R(\boldsymbol{\sigma}^u, \boldsymbol{\sigma}^v) \approx \chi(p_{u\wedge v}) = q_{u\wedge v}. \tag{11}$$

Furthermore, if $\mathbb{E}H_N(\boldsymbol{\sigma}^{\mathsf{alg}}) \ge \mathsf{ALG}(\xi) + \epsilon$, then with high probability, $H_N(\boldsymbol{\sigma}^{\mathsf{alg}}) \ge \mathsf{ALG}(\xi) + \epsilon/2$. We will show that together, these are impossible.

We first define $q_0 = 0, q_i = i\delta, q_D = 1$, for $D = 1/\delta$. We then set $p_i = \chi^{-1}(q_i) = \chi^{-1}(\delta \cdot i)$; this is possible because χ is always increasing in p (by Hermite polynomial expansion) and is clearly continuous.

In what follows, we will consider δ a small constant, $\eta \ll \delta$, and $k \gg 1/\eta$.

First, we extend Equation 11 to hold for any two nodes:

In fact, it is not guaranteed that $\chi(0) = 0$. This does not cause serious issues because we can just start q_0 at $\chi(0)$. Then one will end up with the bound $\int_{\chi(0)}^1 \sqrt{\xi''(q)}dq \le \int_0^1 \sqrt{\xi''(q)}dq$. When there is an external field (which is convenient to treat as deterministic rather than random), there is a non-zero optimal value of $\chi(0) = R\big(\mathbb{E}[\sigma^{\mathsf{alg}}], \mathbb{E}[\sigma^{\mathsf{alg}}]\big)$ because $\mathbb{E}[\sigma^{\mathsf{alg}}]$ can correlate with the external field. This results in slightly more involved formulas to optimize $\chi(0)$ but isn't conceptually harder.

In fact it doesn't really matter that χ is increasing, you can just choose p_i to be the smallest solution to $\chi(p_i) = q_i$.

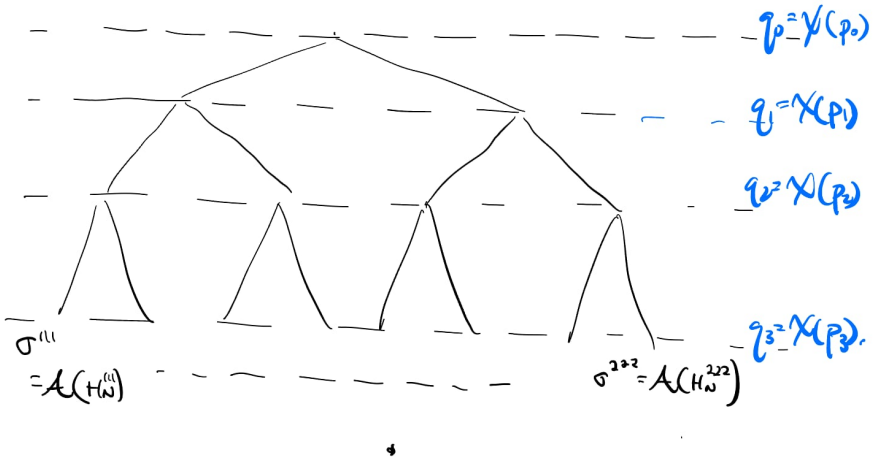

Figure 4: Image under $\mathscr{A}$

Proposition 2.4. *For any δ, η, if k is sufficiently large, then*

$$\left|R(\boldsymbol{\sigma}^u, \boldsymbol{\sigma}^v) - q_{u\wedge v}\right| \le \eta \tag{12}$$

with high probability for all $u \in [k]^{d_1}$, $v \in [k]^{d_2}$, for $0 \le d_1, d_2 \le D$. For interior nodes u, one defines

$$\boldsymbol{\sigma}^u = \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{\sigma}^{ui} \tag{13}$$

for $u \in [k]^{D-1}$, and inductively up the tree.

Proof. If u and v share no children, then this is clear from expanding the covariances. If they do, then let v is a d-th level child of u, and take $w_1, \dots, w_{k^d-1}$ be the others, where we define $w_{k^d} = v$. Now note that

$$R(\boldsymbol{\sigma}^u, \boldsymbol{\sigma}^v) = \frac{1}{k^d}\sum R(\boldsymbol{\sigma}^{w_i}, \boldsymbol{\sigma}^v) = \frac{1}{k^d}\sum R(\boldsymbol{\sigma}^{w_1}, v) + O(1/k^d) = q_{w_1\wedge v} + O(1/k^d) = q_{u\wedge v} + O(1/k^d), \tag{14}$$

so in essence, the distortion is controlled by at most $1/k$, which we took such that $k \gg 1/\eta$. □

Corollary 2.5. *If $i \ne j$, then*

$$\left|R(\boldsymbol{\sigma}^{ui} - \boldsymbol{\sigma}^u, \boldsymbol{\sigma}^{uj} - \boldsymbol{\sigma}^u)\right| \le O(\eta) \sim 0 \tag{15}$$

$$\left|R(\boldsymbol{\sigma}^{ui} - \boldsymbol{\sigma}^u, \boldsymbol{\sigma}^u)\right| \le O(\eta) \tag{16}$$

Proof. Use Proposition 2.4 and expand using bilinearity. □

In essence, a walk down the tree consists of walking down orthogonal increments inside $\mathscr{B}_N$; recall that this is exactly what Subag's algorithm does, though it is only for a single Hamiltonian. This suggests that we should try to write the final energy as a telescoping sum. We analyze

$$\frac{1}{k^D}\sum_{u\in[k]^D} H_N^u(\boldsymbol{\sigma}^u) = \underbrace{H_N^{\varnothing}}_{=0} + \sum_{0\le d\le D-1}\frac{1}{k^d}\sum_{u\in[k]^d}\left[\frac{1}{k}\sum_{i=1}^{k}\left(H_N^{ui}(\boldsymbol{\sigma}^{ui}) - H_N^u(\boldsymbol{\sigma}^u)\right)\right] \tag{17}$$

In essence, we sum down the levels of the trees; for each d, we take the sum of the average growth among all of the children of each node at that level. We will now cheat a bit, and assume that all of the Hamiltonians H_N^u at the leaves are actually the same (we denote this as H_N), and that and assume that the increments $\boldsymbol{\sigma}^{ui} - \boldsymbol{\sigma}^u$ are truly orthogonal. Now we will control the terms above via

$$\frac{1}{k}\sum_{i=1}^{k}\left(H_N^{ui}(\boldsymbol{\sigma}^{ui}) - H_N^u(\boldsymbol{\sigma}^u)\right) \le F^k_{q_d,q_{d+1}}(\boldsymbol{\sigma}^u) \tag{18}$$

where

$$F^k_{q_d,q_{d+1}}(\boldsymbol{\sigma}) = \max_{\substack{\boldsymbol{\sigma}^1,\dots,\boldsymbol{\sigma}^k \\ \boldsymbol{\sigma}^i-\boldsymbol{\sigma}\perp\boldsymbol{\sigma}^j-\boldsymbol{\sigma}\perp\boldsymbol{\sigma}\,\forall i,j}} \frac{1}{k}\sum_{i=1}^{k}\left(H_N(\boldsymbol{\sigma}^i) - H_N(\boldsymbol{\sigma})\right). \tag{19}$$

$F^k_{q_d,q_{d+1}}(\boldsymbol{\sigma})$ essentially represents a worst case energy gain over a given level.

Proposition 2.6. *We make two claims:*

(a) *For a fixed $\boldsymbol{\sigma}$, independent of H_N and $q_{d+1} - q_d = \delta$, and k sufficiently large,*

$$\mathbb{E}F^k_{q_d,q_{d+1}}(\boldsymbol{\sigma})/N \le \delta\sqrt{\xi''(q_d)} + O(\delta^{3/2}). \tag{20}$$

(b) *(Uniform concentration). For all δ,η and k sufficiently large,*

$$\mathbb{P}\left[\max_{R(\boldsymbol{\sigma},\boldsymbol{\sigma})=q_d}\left|F^k_{q_d,q_{d+1}}(\boldsymbol{\sigma}) - \mathbb{E}F^k_{q_d,q_{d+1}}(\boldsymbol{\sigma})\right| \ge \eta N\right] \le e^{-cN}. \tag{21}$$

Before we discuss the proof of this, note that if this holds, then the result of Theorem 1.1 is clear (modulo the fact that we assumed all the Hamiltonians are the same). This is because one only gains $\delta\sqrt{\xi''(\delta\cdot i)}$ energy at level i – hence the analysis is the same as the energy gained by Subag's algorithm, yielding the upper bound.

Proof of Proposition 2.6. For (a), the idea will be to Taylor expand around $\boldsymbol{\sigma}$.

$$\frac{1}{k}\sum_{i=1}^{k}\left(H_N(\boldsymbol{\sigma}^i) - H_N(\boldsymbol{\sigma})\right) = \left\langle \nabla H_N(\boldsymbol{\sigma}), \frac{1}{k}\sum_{i=1}^{k}(\boldsymbol{\sigma}^i - \boldsymbol{\sigma})\right\rangle + \left\langle \nabla^2_{\tan}H_N(\boldsymbol{\sigma}), \frac{1}{k}\sum_{i=1}^{k}(\boldsymbol{\sigma}^i - \boldsymbol{\sigma})^{\otimes 2}\right\rangle + O(\delta^{3/2}) \tag{22}$$

The first term is approximately zero, because the increments are orthogonal to each other. Actually, recall that before we made the assumption they were all exactly orthogonal, $\boldsymbol{\sigma}$ was exactly the mean of the $\boldsymbol{\sigma}^i$'s.

For the Hessian term, note that it is at most $\lambda_1(\nabla^2_{\tan}H_N(\boldsymbol{\sigma}))\cdot\delta N$, which, as we mentioned in the lecture on Subag's algorithm, for general models with covariance function ξ concentrates well and has expectation $(\sqrt{\xi''(q_d)} + o(1))\delta N$.

For the second, the idea will be to use Borell-TIS to control $F^k_{q_d,q_{d+1}}$ at every point, and then union bound over an ϵ-net on the sphere. Recall the statement of Borell-TIS:

Proposition 2.7 (Borell-TIS)**.** *If $g_1,\dots,g_m$ are possibly dependent scalar Gaussians, with $\max_i \mathsf{Var}(g_i) \le c$, then their maximum is sub-Gaussian with variance proxy c:*

$$\mathbb{P}\left[\left|\max_i g_i - \mathbb{E}\max_i g_i\right| \ge \epsilon\right] \le 2e^{-\Omega(\epsilon^2/c)}. \tag{23}$$

Recall then that

$$F^k_{q_d,q_{d+1}}(\boldsymbol{\sigma}) = \max_{\substack{\boldsymbol{\sigma}^1,\dots,\boldsymbol{\sigma}^k \\ \boldsymbol{\sigma}^i-\boldsymbol{\sigma}\perp\boldsymbol{\sigma}^j-\boldsymbol{\sigma}\perp\boldsymbol{\sigma}\,\forall i,j}} \frac{1}{k}\sum_{i=1}^{k}\left(H_N(\boldsymbol{\sigma}^i) - H_N(\boldsymbol{\sigma})\right) \tag{24}$$

which is exactly of this form. Furthermore, letting $g_{\boldsymbol{\sigma}^1,\dots,\boldsymbol{\sigma}^k}$ denote the objective at $\boldsymbol{\sigma}^1,\dots,\boldsymbol{\sigma}^k$, we have

$$\mathsf{Var}(g_{\boldsymbol{\sigma}^1,\dots,\boldsymbol{\sigma}^k}) \le O(N/k) \tag{25}$$

Indeed due to the orthogonality of the increments, for each i in the sum comprising $F^k_{q_d,q_{d+1}}$, one has

$$\mathbb{E}H_N(\boldsymbol{\sigma})H_N(\boldsymbol{\sigma}) = \mathbb{E}H_N(\boldsymbol{\sigma})H_N(\boldsymbol{\sigma}^i) = \mathbb{E}H_N(\boldsymbol{\sigma}^i)H_N(\boldsymbol{\sigma}^j) = N\xi(q_d). \tag{26}$$

(This is just the usual covariance structure of our Hamiltonian, for any fixed inputs.) It follows that if $\boldsymbol{\sigma}^i - \boldsymbol{\sigma} \perp \boldsymbol{\sigma}^j - \boldsymbol{\sigma} \perp \boldsymbol{\sigma}$, then

$$\mathbb{E}\Big[\big(H_N(\boldsymbol{\sigma}^i) - H_N(\boldsymbol{\sigma})\big)\big(H_N(\boldsymbol{\sigma}^j) - H_N(\boldsymbol{\sigma})\big)\Big] = 0.$$

In other words, the k terms in the sum are uncorrelated, so the variance of their average scales as $1/k$. (Moreover their average is still a centered Gaussian.) Therefore (24) is indeed the maximum of a Gaussian process with maximal variance $O(N/k)$, so Borell-TIS applies as claimed.

One now union bounds over an $\eta\sqrt{N}$-net of the sphere and extends by the fact that H_N has controlled derivative. This is enough, since the failure probability within the net is controlled as

$$\left(\frac{10}{\eta}\right)^N e^{-\Omega(\eta^2 kN)} \ll 1 \tag{27}$$

for k large enough depending on η. □

What is left is to deal with the H^u_N not actually all being equal. Here, the ideas are about the same but the formulas become a bit more complicated. The expansion in Claim 1 changes to something along the lines of

$$\frac{1}{k}\sum_{i=1}^k \left[H^{ui}_N(\boldsymbol{\sigma}^{ui}) - H^u_N(\boldsymbol{\sigma}^u)\right] \approx \left\langle \nabla H^u_N, \frac{1}{k}\sum_{i=1}^k (\boldsymbol{\sigma}^{ui} - \boldsymbol{\sigma}^u)\right\rangle + \frac{1}{k}\sum_{i=1}^k \sqrt{p_{d+1}-p_d}\left\langle \nabla \widetilde{H^{ui}_N}, \boldsymbol{\sigma}^{ui} - \boldsymbol{\sigma}^u\right\rangle \tag{28}$$

$$+ \frac{1}{k}\sum_{i=1}^k \left\langle \nabla^2_{\tan} H^u_N, (\boldsymbol{\sigma}^{ui} - \boldsymbol{\sigma}^u)^{\otimes 2}\right\rangle. \tag{29}$$

The first term is again approximately zero due to orthogonality, but now it seems we may be in some trouble, since the second term can now contribute additional energy. However, recall that the third term is now scaled down by a $\sqrt{p_d}$ factor, since not all the Hamiltonians are the same anymore. The way we analyze this is by noting that the second and third terms comprise a quadratic spin glass, which we saw in the homework how to optimize in a simple setting.

For more general covariances, the limiting energy is as follows. Take α to be any left inverse of χ, so that $\alpha(\chi(p_d)) = p_d$, $\alpha(q_d) = p_d$, and α is piecewise linear. Then for such a quadratic spin glass, the limiting ground state is

$$\int_0^1 \sqrt{\alpha'(t)\xi'(t) + \alpha(t)\xi''(t)}\,\mathrm{d}t. \tag{30}$$

The reason for this formula is that for quadratic-plus-linear Hamiltonians, the limiting ground state is just

$$(\mathsf{GS}^2_{\mathsf{lin}} + \mathsf{GS}^2_{\mathsf{quad}})^{1/2}. \tag{31}$$

(This is essentially equivalent to the BBP formula for outlier eigenvalues of spiked GOE matrices, as explained in Homework 3.) In any case when $\xi'(0) = 0$ so there is no external field, the maximum energy is achieved (over settings of q_d) by taking $\alpha(t) = 1$, from which we now recover the previous setting, since now all the Hamiltonians are perfectly correlated. The idea for showing $\alpha = 1$ is optimal is to use Karamata's inequality on the square root function.

3 Necessity

One can ask whether such a large tree was necessary to the analysis.

Proposition 3.1. *Suppose $\mathbb{T}$ is a rooted tree and contains no complete binary subtree of depth d. An example of a deep tree with no deep binary subtree is in Figure 5. Then any such tree appears in the superlevel sets*

$$\{\boldsymbol{x} \in \mathscr{S}_N : H_N(\boldsymbol{x})/N \geq \mathsf{ALG}(\xi) + \epsilon_d\} \tag{32}$$

unless $\mathsf{ALG}(\xi) = \mathsf{OPT}(\xi)$.

Figure 5: Deep tree with no deep binary subtree

What this says is that if $\mathsf{ALG} \neq \mathsf{OPT}$, if we want to find a structure actually forbidden, then it needs to contain a deep tree – if you don't, you cannot reach this algorithmic threshold. Instead your proof will find some threshold strictly between ALG and OPT.

Note that the above setting basically generalizes to trees where each node has k children. Roughly, one can take a binary tree and think of "skipping" intermediate levels, so that each node now has 2^ℓ children for some ℓ.

References

[1] Brice Huang and Mark Sellke. Tight lipschitz hardness for optimizing mean field spin glasses, 2022. 1

[2] Brice Huang and Mark Sellke. Algorithmic threshold for multi-species spherical spin glasses, 2023. 1