Stochastic Localization Sampling For the SK Model

Mark Sellke

IAS CSDM

Collaborators

Ahmed El Alaoui

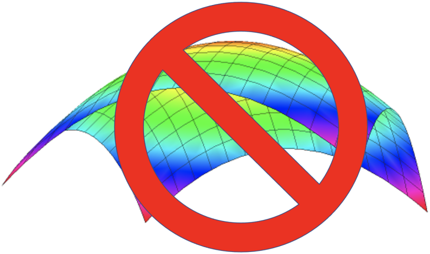
Andrea Montanari

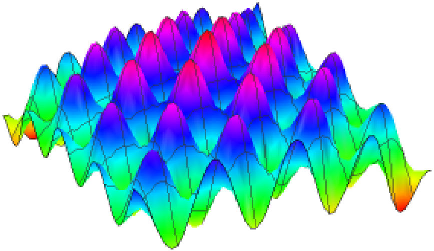
Sampling

Goal: generate

$$x^* \sim \mu(dx) \quad \text{given} \quad \mu \in \mathcal{P}(\mathbb{R}^n).$$

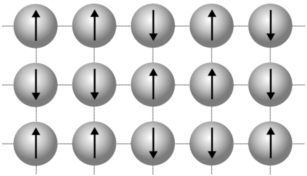
For μ high-dimensional and NOT log-concave.

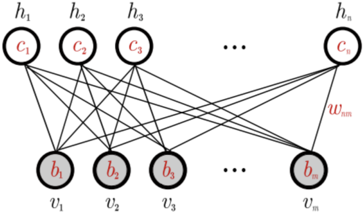




In this talk, focus on Ising models:

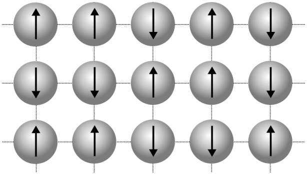
$$\mu_{\boldsymbol{A},\beta}(\boldsymbol{x}) = \frac{1}{Z(\beta)} e^{\beta\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle/2}, \qquad \boldsymbol{x} \in \{-1,+1\}^n.$$

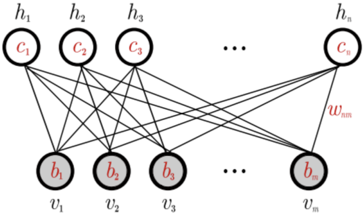




In this talk, focus on Ising models:

$$\mu_{\boldsymbol{A},\beta}(\boldsymbol{x}) = \frac{1}{Z(\beta)} e^{\beta\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle/2}, \qquad \boldsymbol{x} \in \{-1,+1\}^n.$$





Glauber dynamics

- Repeatedly choose $i \in [n]$ and resample x_i given other coordinates.
- Mixes rapidly if $\boldsymbol{A}$ is small. In general, mixing can be very slow.

Sequential Sampling

Given a distribution $\mu \in \mathcal{P}(\{-1,+1\}^n)$, suppose we have a conditional expectation **oracle** to evaluate

$$\boldsymbol{m}^t = \mathbb{E}^{\boldsymbol{x}\sim\mu}[\boldsymbol{x} \mid (x_1 = x_1^*, \dots, x_t = x_t^*)], \quad t \in \{0, 1, \dots, n-1\}.$$

Sequential Sampling

Given a distribution $\mu \in \mathcal{P}(\{-1,+1\}^n)$, suppose we have a conditional expectation **oracle** to evaluate

$$\boldsymbol{m}^t = \mathbb{E}^{\boldsymbol{x}\sim\mu}[\boldsymbol{x} \mid (x_1 = x_1^*, \ldots, x_t = x_t^*)], \quad t \in \{0, 1, \ldots, n-1\}.$$

Then we can directly sample $\boldsymbol{x}$, one coordinate at a time. Namely,

$$\mathbb{P}^t[x_{t+1} = 1 \mid x_1, \ldots, x_t] = \frac{\boldsymbol{m}_{t+1}^t + 1}{2}.$$

This is the foundation for equivalence between counting and sampling.

Downsides of Sequential Sampling

Sequential sampling may be too much to hope for.

- Requires a strong oracle, especially for continuous variables.
- Maybe estimating $\boldsymbol{m}^t$ is no easier than sampling.
- Unclear how to choose a good order for the coordinates.

Downsides of Sequential Sampling

Sequential sampling may be too much to hope for.

- Requires a strong oracle, especially for continuous variables.
- Maybe estimating $\boldsymbol{m}^t$ is no easier than sampling.
- Unclear how to choose a good order for the coordinates.

In sequential sampling, we try to reveal $\boldsymbol{x}^*$ **gradually**.
There are other ways to do this.

Warm-Up: Pólya's Urn

A silly way to sample $p \sim Unif([0,1])$:

- Sample an infinite sequence $(b_1, b_2, \dots)$ of i.i.d. $Ber(p)$ bits, **without knowing p**.
- Use law of large numbers to compute $p = \lim_{t\to\infty} \frac{\sum_{s=1}^t b_s}{t}$.

Warm-Up: Pólya's Urn

A silly way to sample $p \sim \mathit{Unif}([0,1])$:

- Sample an infinite sequence $(b_1, b_2, \dots)$ of i.i.d. $\mathit{Ber}(p)$ bits, **without knowing p**.
- Use law of large numbers to compute $p = \lim_{t\to\infty} \frac{\sum_{s=1}^{t} b_s}{t}$.

...and this is really not so bad.

- Given $(b_1, \dots, b_t)$, the posterior expectation for p is given by Laplace's rule of succession:

$$\mathbb{E}^t[p] = \frac{1 + \sum_{s=1}^{t} b_t}{t+2}.$$

Warm-Up: Pólya's Urn

A silly way to sample $p \sim \mathit{Unif}([0,1])$:

- Sample an infinite sequence $(b_1, b_2, \dots)$ of i.i.d. $\mathit{Ber}(p)$ bits, **without knowing p**.
- Use law of large numbers to compute $p = \lim_{t\to\infty} \frac{\sum_{s=1}^t b_s}{t}$.

...and this is really not so bad.

- Given $(b_1, \dots, b_t)$, the posterior expectation for p is given by Laplace's rule of succession:

$$\mathbb{E}^t[p] = \frac{1 + \sum_{s=1}^t b_t}{t+2}.$$

- Hence the sequential rule

$$\mathbb{P}^t[b_{t+1} = 1] = \frac{1 + \sum_{s=1}^t b_t}{t+2},$$

yields an i.i.d. $\mathit{Ber}(p)$ sequence for $p \sim \mathit{Unif}([0,1])$.

Stochastic Localization: Revealing $\boldsymbol{x}^*$ with Gaussian Noise

(A version of) Eldan's Stochastic localization:

$$\boldsymbol{y}_t = t\boldsymbol{x}^* + \boldsymbol{B}_t \quad \sim \mathcal{N}(t\boldsymbol{x}^*, t\boldsymbol{I}_n).$$

$\boldsymbol{x}^* \sim \mu$ is independent of Brownian motion $\boldsymbol{B}_t$.

Stochastic Localization: Revealing $\boldsymbol{x}^*$ with Gaussian Noise

(A version of) Eldan's Stochastic localization:

$$\boldsymbol{y}_t = t\boldsymbol{x}^* + \boldsymbol{B}_t \quad \sim \mathcal{N}(t\boldsymbol{x}^*, t\boldsymbol{I}_n).$$

$\boldsymbol{x}^* \sim \mu$ is independent of Brownian motion $\boldsymbol{B}_t$.

Suggests a sampling algorithm:

1. **Simulate $\boldsymbol{y}_t$** for a long time $t \in [0, T]$ without knowing $\boldsymbol{x}^*$.
2. **Read off**

$$\boldsymbol{x}^* \approx \frac{\boldsymbol{y}_T}{T}.$$

Stochastic Localization: Revealing $\boldsymbol{x}^*$ with Gaussian Noise

(A version of) Eldan's Stochastic localization:

$$\boldsymbol{y}_t = t\boldsymbol{x}^* + \boldsymbol{B}_t \quad \sim \mathcal{N}(t\boldsymbol{x}^*, t\boldsymbol{I}_n).$$

$\boldsymbol{x}^* \sim \mu$ is independent of Brownian motion $\boldsymbol{B}_t$.

Suggests a sampling algorithm:

1. **Simulate $\boldsymbol{y}_t$** for a long time $t \in [0, T]$ without knowing $\boldsymbol{x}^*$.
2. **Read off**
$$\boldsymbol{x}^* \approx \frac{\boldsymbol{y}_T}{T}.$$

Geometric motivation: decompose general μ into posteriors $\mu_t(\mathrm{d}x) \propto e^{\langle \boldsymbol{y}_t, \boldsymbol{x}\rangle - t\|x\|_2^2/2}\mu(\mathrm{d}x)$.

- If μ log-concave, each μ_t is **strongly** log-concave.
- KLS conjecture [Eldan 12, Lee-Vempala 17, Chen 21, Klartag-Lehec 22].

Simulating $\boldsymbol{y}_t$

Similarly to Pólya's urn, we can generate the path

$$\boldsymbol{y}_t = t\boldsymbol{x}^* + \boldsymbol{B}_t, \quad \boldsymbol{x}^* \sim \mu$$

without knowing $\boldsymbol{x}^*$.

Simulating $\boldsymbol{y}_t$

Similarly to Pólya's urn, we can generate the path

$$\boldsymbol{y}_t = t\boldsymbol{x}^* + \boldsymbol{B}_t, \quad \boldsymbol{x}^* \sim \mu$$

without knowing $\boldsymbol{x}^*$. The **annealed** law of $\boldsymbol{y}_t$ is described by

$$\begin{aligned} \mathrm{d}\boldsymbol{y}_t &= \boldsymbol{m}_t \mathrm{d}t + \mathrm{d}\boldsymbol{W}_t; \\ \boldsymbol{m}_t &= \mathbb{E}[\boldsymbol{x}^* \mid \mathcal{F}_t] = \mathbb{E}[\boldsymbol{x}^* \mid \boldsymbol{y}_t] \end{aligned}$$

for W_t another Brownian motion.

Simulating $\boldsymbol{y}_t$

Similarly to Pólya's urn, we can generate the path

$$\boldsymbol{y}_t = t\boldsymbol{x}^* + \boldsymbol{B}_t, \quad \boldsymbol{x}^* \sim \mu$$

without knowing $\boldsymbol{x}^*$. The **annealed** law of $\boldsymbol{y}_t$ is described by

$$\begin{aligned} \mathrm{d}\boldsymbol{y}_t &= \boldsymbol{m}_t \mathrm{d}t + \mathrm{d}\boldsymbol{W}_t; \\ \boldsymbol{m}_t &= \mathbb{E}[\boldsymbol{x}^* \mid \mathcal{F}_t] = \mathbb{E}[\boldsymbol{x}^* \mid \boldsymbol{y}_t] \end{aligned}$$

for W_t another Brownian motion.

Equivalence:

- Quadratic variation is Brownian in either case.
- $\boldsymbol{y}_t - \int_0^t \boldsymbol{m}_t \mathrm{d}t$ is a martingale in either case since $\boldsymbol{m}_t = \mathbb{E}[x_* \mid \mathcal{F}_t]$.
- Now use Lévy's characterization of Brownian motion.

The Resulting Algorithm

$$\mathrm{d}\boldsymbol{y}_t = \boldsymbol{m}_t \mathrm{d}t + \mathrm{d}\boldsymbol{W}_t,$$

A continuous-time stochastic process is not really an algorithm.

Of course, we should discretize.

The Resulting Algorithm

$$\mathrm{d}\boldsymbol{y}_t = \boldsymbol{m}_t \mathrm{d}t + \mathrm{d}\boldsymbol{W}_t,$$

Input: Data: Probability measure μ
Input: Result: Sample $\boldsymbol{x}^* \sim \mu$
for $t \in [0, \delta, \ldots, T - \delta]$ **do**
    Sample $\mathbf{g}_t \sim \mathcal{N}(0, I_n)$
    Set $\boldsymbol{y}_{t+\delta} = \boldsymbol{y}_t + \hat{\boldsymbol{m}}_t(\boldsymbol{y}_t)\delta + \sqrt{\delta}\mathbf{g}_t$
end
Set $\boldsymbol{x}^* = Round(\boldsymbol{y}_T / T) \in \{-1, +1\}^n$
return $\boldsymbol{x}^*$

$$\mathrm{d}\boldsymbol{y}_t = \boldsymbol{m}_t \mathrm{d}t + \mathrm{d}\boldsymbol{W}_t,$$

Input: Data: Probability measure μ
Input: Result: Sample $\boldsymbol{x}^* \sim \mu$
for $t \in [0, \delta, \ldots, T - \delta]$ **do**
  Sample $\mathbf{g}_t \sim \mathcal{N}(0, I_n)$
  Set $\boldsymbol{y}_{t+\delta} = \boldsymbol{y}_t + \hat{\boldsymbol{m}}_t(\boldsymbol{y}_t)\delta + \sqrt{\delta}\mathbf{g}_t$
end
Set $\boldsymbol{x}^* = Round(\boldsymbol{y}_T/T) \in \{-1, +1\}^n$
return $\boldsymbol{x}^*$

Main requirement: a good approximation $\hat{\boldsymbol{m}}_t(\boldsymbol{y}_t) \approx \mathbb{E}[x^* \mid \boldsymbol{y}_t]$.

Where Do We Stand?

So far:

- General sampling procedure.
- Requires estimating $\boldsymbol{m}_t(\boldsymbol{y}_t) \approx \mathbb{E}[x^* \mid \boldsymbol{y}_t]$.

We have replaced the need for one oracle with another...is it any better?

Where Do We Stand?

So far:

- General sampling procedure.
- Requires estimating $\boldsymbol{m}_t(\boldsymbol{y}_t) \approx \mathbb{E}[x^* \mid \boldsymbol{y}_t]$.

We have replaced the need for one oracle with another...is it any better?

Remainder of the talk: example where the answer is **yes**.

- SK model: coupling matrix $\boldsymbol{A}$ is GOE.
- Computing $\boldsymbol{m}_t(\boldsymbol{y}_t)$ falls into the wheelhouse of high-dimensional statistics/optimization.

Sherrington-Kirkpatrick Model

Sherrington-Kirkpatrick Model

Ising model with random couplings:

$$\mu_{\boldsymbol{G},\beta}(\boldsymbol{x}) = \frac{1}{Z_n(\beta)} e^{\beta\langle \boldsymbol{x}, \boldsymbol{G}\boldsymbol{x}\rangle/2}.$$

Random symmetric matrix $\boldsymbol{G} \sim GOE(n)$:

- $\boldsymbol{G} = \boldsymbol{G}^\top$. Entries otherwise independent.
- $G_{i,j} \sim \mathcal{N}(0, 1/n)$ for $i < j$.

Sherrington-Kirkpatrick Model

Ising model with random couplings:

$$\mu_{\boldsymbol{G},\beta}(\boldsymbol{x}) = \frac{1}{Z_n(\beta)} e^{\beta\langle \boldsymbol{x}, \boldsymbol{G}\boldsymbol{x}\rangle/2}.$$

Random symmetric matrix $\boldsymbol{G} \sim GOE(n)$:

- $\boldsymbol{G} = \boldsymbol{G}^\top$. Entries otherwise independent.
- $G_{i,j} \sim \mathcal{N}(0, 1/n)$ for $i < j$.

Goal: given $\boldsymbol{G} \sim GOE(n)$, generate a sample from $\mu_{\boldsymbol{G},\beta}$.

Dobrushin's condition for fast mixing of Glauber works if $\beta \le cn^{-1/2}$.
But we would like β to be constant size.

Brief History of the SK Model

[Ising 1925]: Ising model for ferromagnets.

[Sherrington-Kirkpatrick 1975]: model for **disordered** magnets.

[Parisi 1982]: non-rigorous solution via replica symmetry breaking.

Brief History of the SK Model

[Ising 1925]: Ising model for ferromagnets.

[Sherrington-Kirkpatrick 1975]: model for **disordered** magnets.

[Parisi 1982]: non-rigorous solution via replica symmetry breaking.

[Talagrand 2005] proves the Parisi formula.

- Huge amount of other important work including [Aizenman-Ruelle-Lebowitz 82, Ruelle 87, Chatterjee 09, Panchenko 14, Ding-Sly-Sun 15, Auffinger-Chen 17,...].

Connections

SK model is a prototype for disordered, random probability measures.

- Random MaxCut and K-SAT.
- Coloring random graphs.
- Posteriors in high-dimensional statistics.

Connections

SK model is a prototype for disordered, random probability measures.

- Random MaxCut and K-SAT.
- Coloring random graphs.
- Posteriors in high-dimensional statistics.

E.g. optimal MaxCut in a random sparse graph ([Dembo-Montanari-Sen 17]).
For $\mathrm{G} \sim \mathrm{G}\left(n, \frac{\lambda}{n}\right)$:

$$\mathrm{MaxCut(G)} = n\left(\frac{\lambda}{4} + C_* \sqrt{\frac{\lambda}{4}} + o(\sqrt{\lambda})\right) + o(n).$$

$$\mu_{\boldsymbol{G},\beta}(\boldsymbol{x}) = \frac{1}{Z_n(\beta)} e^{\beta \langle \boldsymbol{x}, \boldsymbol{G}\boldsymbol{x} \rangle / 2}.$$

Expect: efficient sampling possible for $\beta < 1$, impossible for $\beta > 1$.

- Replica symmetric iff $\beta \leq 1$.

Rigorous Results on Sampling

$$\mu_{\boldsymbol{G},\beta}(\boldsymbol{x}) = \frac{1}{Z_n(\beta)} e^{\beta\langle \boldsymbol{x}, \boldsymbol{G}\boldsymbol{x}\rangle/2}.$$

Expect: efficient sampling possible for $\beta < 1$, impossible for $\beta > 1$.

- Replica symmetric iff $\beta \leq 1$.

Recent progress: Glauber mixes in $O(n \log n)$ steps for $\beta < 1/4$.
[Bodineau-Bauerschmidt 20, Eldan-Koehler-Zeitouni 21, Anari-Jain-Koehler-Pham-Vuong 21].
A different method for tensor analogs: [Adhikari-Brennecke-Xu-Yau 22]

Rigorous Results on Sampling

$$\mu_{\boldsymbol{G},\beta}(\boldsymbol{x}) = \frac{1}{Z_n(\beta)} e^{\beta\langle \boldsymbol{x}, \boldsymbol{G}\boldsymbol{x}\rangle/2}.$$

Expect: efficient sampling possible for $\beta < 1$, impossible for $\beta > 1$.

- Replica symmetric iff $\beta \leq 1$.

Recent progress: Glauber mixes in $O(n \log n)$ steps for $\beta < 1/4$.

[Bodineau-Bauerschmidt 20, Eldan-Koehler-Zeitouni 21, Anari-Jain-Koehler-Pham-Vuong 21].

A different method for tensor analogs: [Adhikari-Brennecke-Xu-Yau 22]

Our result: stochastic localization succeeds (in a weaker sense) for $\beta < 1$. (Originally $\beta < 1/2$, improvement by [Celentano 22].)

Normalized Wasserstein Metric

Given $\mu_1, \mu_2 \in \mathcal{P}(\{-1, 1\}^n)$, define the normalized Wasserstein metric

$$W_{1,n}(\mu_1, \mu_2) = \inf_{(x_1,x_2)\sim Coupling(\mu_1,\mu_2)} \frac{\mathbb{E}[\|x_1 - x_2\|_{\ell^1}]}{n}.$$

$W_{1,n}(\mu_1, \mu_2) \leq o(1)$ means that x_1, x_2 differ by $o(n)$ coordinates under an optimal coupling. We will consider such pairs of points to be close.

Main Result

Theorem (El Alaoui-Montanari-S 22, Celentano 22)

For any $\beta < 1$ and $\varepsilon > 0$, there exists a randomized algorithm with complexity $O(n^2)$ which given $\boldsymbol{G}$ outputs $\boldsymbol{x} \sim \mu^{\mathrm{alg}}_{\boldsymbol{G},\beta}$ such that

$$\mathbb{E}[W_{1,n}(\mu^{\mathrm{alg}}_{\boldsymbol{G},\beta}, \mu_{\boldsymbol{G},\beta})] \leq \varepsilon.$$

Algorithmic Stability

Our algorithm is **stable** with respect to $(\boldsymbol{G}, \beta)$: just uses $O_{\beta,\varepsilon}(1)$ matrix-vector products, and some 1-dimensional non-linearities.

Concretely, from i.i.d. $\boldsymbol{G} = \boldsymbol{G}_0$ and $\boldsymbol{G}_1$, consider perturbation path

$$\boldsymbol{G}_s = \sqrt{1-s^2}\boldsymbol{G}_0 + s\boldsymbol{G}_1.$$

Stability of the algorithm means:

$$\lim_{s\to 0}\lim_{n\to\infty}\mathbb{E}[W_{1,n}(\mu^{\mathrm{alg}}_{\boldsymbol{G}_0,\beta}, \mu^{\mathrm{alg}}_{\boldsymbol{G}_s,\beta})] = 0.$$

Algorithmic Stability

Our algorithm is **stable** with respect to $(\boldsymbol{G}, \beta)$: just uses $O_{\beta,\varepsilon}(1)$ matrix-vector products, and some 1-dimensional non-linearities.

Concretely, from i.i.d. $\boldsymbol{G} = \boldsymbol{G}_0$ and $\boldsymbol{G}_1$, consider perturbation path

$$\boldsymbol{G}_s = \sqrt{1-s^2}\,\boldsymbol{G}_0 + s\boldsymbol{G}_1.$$

Stability of the algorithm means:

$$\lim_{s\to 0}\lim_{n\to\infty} \mathbb{E}[W_{1,n}(\mu^{\mathrm{alg}}_{\boldsymbol{G}_0,\beta}, \mu^{\mathrm{alg}}_{\boldsymbol{G}_s,\beta})] = 0.$$

A purely structural consequence with an algorithmic proof:

Theorem (El Alaoui-Montanari-S 22; Celentano 22)

*The **true** SK Gibbs measures are stable when $\beta < 1$:*

$$\lim_{s\to 0}\lim_{n\to\infty} \mathbb{E}[W_{1,n}(\mu_{\boldsymbol{G}_0,\beta}, \mu_{\boldsymbol{G}_s,\beta})] = 0.$$

Similar stability holds for small pertubations in β.

Hardness via Chaos

The stability property

$$\lim_{s\to 0}\lim_{n\to\infty}\mathbb{E}[W_{1,n}(\mu_{\boldsymbol{G}_0,\beta},\mu_{\boldsymbol{G}_s,\beta})]=0.$$

for the true Gibbs measure is **false** for $\beta > 1$. Combination of:

Hardness via Chaos

The stability property

$$\lim_{s\to 0}\lim_{n\to\infty}\mathbb{E}[W_{1,n}(\mu_{\boldsymbol{G}_0,\beta},\mu_{\boldsymbol{G}_s,\beta})] = 0.$$

for the true Gibbs measure is **false** for $\beta > 1$. Combination of:

Theorem (Chatterjee 09; Disorder Chaos)

Let $(\boldsymbol{x}_0, \boldsymbol{x}_s) \sim \mu_{\boldsymbol{G}_0,\beta} \times \mu_{\boldsymbol{G}_s,\beta}$. *For all* $\beta \in \mathbb{R}$ *and* $s > 0$,

$$\lim_{n\to\infty}\mathbb{E}[|\langle \boldsymbol{x}_0, \boldsymbol{x}_s\rangle|/n] = 0.$$

Hardness via Chaos

The stability property

$$\lim_{s\to 0}\lim_{n\to\infty}\mathbb{E}[W_{1,n}(\mu_{\boldsymbol{G}_0,\beta},\mu_{\boldsymbol{G}_s,\beta})]=0.$$

for the true Gibbs measure is **false** for $\beta > 1$. Combination of:

Theorem (Chatterjee 09; Disorder Chaos)

Let $(\boldsymbol{x}_0,\boldsymbol{x}_s)\sim\mu_{\boldsymbol{G}_0,\beta}\times\mu_{\boldsymbol{G}_s,\beta}$. *For all* $\beta\in\mathbb{R}$ *and* $s>0$,

$$\lim_{n\to\infty}\mathbb{E}[|\langle\boldsymbol{x}_0,\boldsymbol{x}_s\rangle|/n]=0.$$

Theorem (Replica Symmetry Breaking)

Let $\boldsymbol{x}_0,\boldsymbol{x}_0'\sim\mu_{\boldsymbol{G}_0,\beta}$ *be independent. For all* $\beta>1$,

$$\liminf_{n\to\infty}\mathbb{E}[|\langle\boldsymbol{x}_0,\boldsymbol{x}_0'\rangle|/n]\geq c(\beta)>0.$$

Hardness via Chaos

The previous results show that $\mu_{\boldsymbol{G}_0,\beta}$ and $\mu_{\boldsymbol{G}_s,\beta}$ must be significantly different. Therefore:

Theorem (El Alaoui-Montanari-S 22)

Let $\mu^{\mathrm{alg}}_{\boldsymbol{G},\beta}$ be the law of $\mathrm{ALG}_n(\boldsymbol{G},\beta,\omega)$ conditional on $\boldsymbol{G}$. If ALG_n is ***stable****, then for all $\beta > 1$,*

$$\liminf_{n\to\infty} \mathbb{E}[W_{1,n}(\mu^{\mathrm{alg}}_{\boldsymbol{G},\beta}, \mu_{\boldsymbol{G},\beta})] > c(\beta) > 0.$$

Stability holds for gradient-based methods such as Langevin dynamics and AMP, at least on **dimension-independent** time-scales.

Back to the Main Story...

Back to the Main Story...

To sample for $\beta < 1$, our main requirement is to estimate $\boldsymbol{m}_t = \mathbb{E}[\boldsymbol{x}^* \mid \boldsymbol{y}_t]$ for

$$\boldsymbol{y}_t = t\boldsymbol{x}^* + \boldsymbol{B}_t.$$

The solution goes through several ideas in high-dimensional statistics and optimization.

Back to the Main Story...

To sample for $\beta < 1$, our main requirement is to estimate $\boldsymbol{m}_t = \mathbb{E}[\boldsymbol{x}^* \mid \boldsymbol{y}_t]$ for

$$\boldsymbol{y}_t = t\boldsymbol{x}^* + \boldsymbol{B}_t.$$

The solution goes through several ideas in high-dimensional statistics and optimization.

Two phase procedure:

- Rough estimate for $\boldsymbol{m}_t$ using approximate message passing.
- High-accuracy estimate for $\boldsymbol{m}_t$ using gradient descent on a well-chosen potential.

Step 1: Rough Estimate of $\boldsymbol{m}_t$

Self-consistent "naive mean-field" equation for $\boldsymbol{m}_t = \mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y}_t]$:

$$\boldsymbol{m}_t \approx \tanh\left(\beta \boldsymbol{G} \boldsymbol{m}_t + \boldsymbol{y}_t\right)$$

- Intuitively, $\left(\beta \boldsymbol{G} \boldsymbol{m}_t + \boldsymbol{y}_t\right)_i$ is the effective field on $\boldsymbol{x}_i$.
- $\tanh(\cdot)$ converts from field on $\{-1, +1\}$ to probabilities

Step 1: Rough Estimate of $\boldsymbol{m}_t$

Self-consistent "naive mean-field" equation for $\boldsymbol{m}_t = \mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y}_t]$:

$$\boldsymbol{m}_t \approx \tanh\left(\beta \boldsymbol{G}\boldsymbol{m}_t + \boldsymbol{y}_t\right)$$

- Intuitively, $\left(\beta \boldsymbol{G}\boldsymbol{m}_t + \boldsymbol{y}_t\right)_i$ is the effective field on $\boldsymbol{x}_i$.
- $\tanh(\cdot)$ converts from field on $\{-1, +1\}$ to probabilities
- Not quite right. It actually should be

$$\boldsymbol{m}_t = \mathbb{E}^t[\tanh\left(\beta \boldsymbol{G}\boldsymbol{x} + \boldsymbol{y}_t\right)].$$

  $\tanh(\cdot)$ is non-linear and although $\mathbb{E}^t[\boldsymbol{G}x] = \boldsymbol{G}\boldsymbol{m}_t$ there is nontrivial conditional randomness left.

Step 1: Rough Estimate of $\boldsymbol{m}_t$

Self-consistent "naive mean-field" equation for $\boldsymbol{m}_t = \mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y}_t]$:

$$\boldsymbol{m}_t \approx \tanh\left(\beta \boldsymbol{G} \boldsymbol{m}_t + \boldsymbol{y}_t\right)$$

- Intuitively, $\left(\beta \boldsymbol{G} \boldsymbol{m}_t + \boldsymbol{y}_t\right)_i$ is the effective field on $\boldsymbol{x}_i$.
- $\tanh(\cdot)$ converts from field on $\{-1, +1\}$ to probabilities
- Not quite right. It actually should be

  $$\boldsymbol{m}_t = \mathbb{E}^t[\tanh\left(\beta \boldsymbol{G} \boldsymbol{x} + \boldsymbol{y}_t\right)].$$

  $\tanh(\cdot)$ is non-linear and although $\mathbb{E}^t[\boldsymbol{G} x] = \boldsymbol{G} \boldsymbol{m}_t$ there is nontrivial conditional randomness left.

Revised Thouless-Anderson-Palmer (TAP) equation:

$$\boldsymbol{m}_t \approx \tanh\left(\beta \boldsymbol{G} \boldsymbol{m}_t + \boldsymbol{y}_t - \beta^2 \left(1 - \frac{\|\boldsymbol{m}_t\|_2^2}{n}\right) \boldsymbol{m}_t\right).$$

Step 1: Rough Estimate of $\boldsymbol{m}_t$

Turn the TAP equation into a **recursion** and repeat until convergence to an approximate **fixed point**:

$$\hat{\boldsymbol{m}}_t^{(k+1)} = \tanh\left(\beta \boldsymbol{G} \hat{\boldsymbol{m}}_t^{(k)} + \boldsymbol{y}_t - b_k \hat{\boldsymbol{m}}_t^{(k-1)}\right),$$

$$b_k = \beta^2 \left(1 - \frac{\|\boldsymbol{m}_t^{(k)}\|_2^2}{n}\right).$$

Step 1: Rough Estimate of $\boldsymbol{m}_t$

Turn the TAP equation into a **recursion** and repeat until convergence to an approximate **fixed point**:

$$\hat{\boldsymbol{m}}_t^{(k+1)} = \tanh\left(\beta \boldsymbol{G}\hat{\boldsymbol{m}}_t^{(k)} + \boldsymbol{y}_t - b_k \hat{\boldsymbol{m}}_t^{(k-1)}\right),$$

$$b_k = \beta^2\left(1 - \frac{\|\boldsymbol{m}_t^{(k)}\|_2^2}{n}\right).$$

- This is an **approximate message passing** algorithm. Generalizes belief propagation to dense matrices $\boldsymbol{G}$.
  - Onsager term $b_k \hat{\boldsymbol{m}}_t^{(k-1)}$ cancels "backtracking" paths.

Step 1: Rough Estimate of $\boldsymbol{m}_t$

Turn the TAP equation into a **recursion** and repeat until convergence to an approximate **fixed point**:

$$\hat{\boldsymbol{m}}_t^{(k+1)} = \tanh\left(\beta \boldsymbol{G} \hat{\boldsymbol{m}}_t^{(k)} + \boldsymbol{y}_t - b_k \hat{\boldsymbol{m}}_t^{(k-1)}\right),$$

$$b_k = \beta^2 \left(1 - \frac{\|\boldsymbol{m}_t^{(k)}\|_2^2}{n}\right).$$

- This is an **approximate message passing** algorithm. Generalizes belief propagation to dense matrices $\boldsymbol{G}$.
  - Onsager term $b_k \hat{\boldsymbol{m}}_t^{(k-1)}$ cancels "backtracking" paths.
  - By now, a major tool in high-dimensional statistics.

    [Bolthausen 14, Donoho-Maleki-Montanari 09, Bayati-Montanari 11, Javanmard-Montanari 12, Rush-Venkataramanan 18, Chen-Lam 20, Fan 20, Dudeja-Lu-Sen 22]
- In our case, the AMP state evolution is unclear. $\boldsymbol{y}_t = t\boldsymbol{x}^* + B_t$ for $\boldsymbol{x}^* \sim \mu_{\boldsymbol{G},\beta}$ has a complicated distribution.

Contiguity with a Simpler Spiked Model

To analyze the AMP recursion, we consider a **spiked** joint distribution $\mathbb{Q}$ over $(\boldsymbol{G}, \boldsymbol{x}^*, \boldsymbol{y}_t)$. Under $\mathbb{Q}$:

$$\boldsymbol{x}^* \sim Unif\left(\{-1,1\}^n\right), \qquad \boldsymbol{y}_t = t\boldsymbol{x}^* + B_t,$$

$$\boldsymbol{G} \sim GOE(n) + \frac{\beta \boldsymbol{x}\boldsymbol{x}^\top}{n}.$$

Contiguity with a Simpler Spiked Model

To analyze the AMP recursion, we consider a **spiked** joint distribution $\mathbb{Q}$ over $(\boldsymbol{G}, \boldsymbol{x}^*, \boldsymbol{y}_t)$. Under $\mathbb{Q}$:

$$\boldsymbol{x}^* \sim Unif\left(\{-1, 1\}^n\right), \qquad \boldsymbol{y}_t = t\boldsymbol{x}^* + B_t,$$

$$\boldsymbol{G} \sim GOE(n) + \frac{\beta \boldsymbol{x}\boldsymbol{x}^\top}{n}.$$

The resulting conditional law $\mathbb{Q}[\boldsymbol{G} \mid \boldsymbol{x}^*]$ looks similar to $\mathbb{P}[\boldsymbol{x}^* \mid \boldsymbol{G}]$ for the SK model:

$$\mathbb{Q}[\boldsymbol{G} \mid \boldsymbol{x}^*] \propto e^{\beta\langle \boldsymbol{x}^*, \boldsymbol{G}\boldsymbol{x}^* \rangle/2}.$$

Contiguity with a Simpler Spiked Model

To analyze the AMP recursion, we consider a **spiked** joint distribution $\mathbb{Q}$ over $(\boldsymbol{G}, \boldsymbol{x}^*, \boldsymbol{y}_t)$. Under $\mathbb{Q}$:

$$\boldsymbol{x}^* \sim \mathit{Unif}\left(\{-1,1\}^n\right), \qquad \boldsymbol{y}_t = t\boldsymbol{x}^* + B_t,$$

$$\boldsymbol{G} \sim \mathit{GOE}(n) + \frac{\beta \boldsymbol{x}\boldsymbol{x}^\top}{n}.$$

The resulting conditional law $\mathbb{Q}[\boldsymbol{G} \mid \boldsymbol{x}^*]$ looks similar to $\mathbb{P}[\boldsymbol{x}^* \mid \boldsymbol{G}]$ for the SK model:

$$\mathbb{Q}[\boldsymbol{G} \mid \boldsymbol{x}^*] \propto e^{\beta \langle \boldsymbol{x}^*, \boldsymbol{G}\boldsymbol{x}^* \rangle / 2}.$$

Swapping the order distorts probabilities by a partition function factor

$$Z_{SK}(\boldsymbol{G}) = \sum_{\boldsymbol{v} \in \{-1,+1\}^n} e^{\beta \langle \boldsymbol{v}, \boldsymbol{G}\boldsymbol{v} \rangle / 2}.$$

- $Z_{SK}(\boldsymbol{G})$ fluctuates **mildly** for $\beta < 1$ [Aizenman-Ruelle-Lebowitz 82]. The spiked model is **contiguous** with the original.

State Evolution for AMP

$$\hat{\boldsymbol{m}}_t^{(k+1)} = \tanh\left(\beta \boldsymbol{G}\hat{\boldsymbol{m}}_t^{(k)} + \boldsymbol{y}_t - b_k \hat{\boldsymbol{m}}_t^{(k-1)}\right)$$

Idea of AMP: for fixed $\boldsymbol{v}, \boldsymbol{w}$, the vectors

$$(\boldsymbol{G}\boldsymbol{v}, \boldsymbol{G}\boldsymbol{w})$$

each have i.i.d. Gaussian coordinates. Covariance between $(\boldsymbol{G}\boldsymbol{v})_i$ and $(\boldsymbol{G}\boldsymbol{w})_i$ equals $\langle \boldsymbol{v}, \boldsymbol{w} \rangle$.

State Evolution for AMP

$$\hat{\boldsymbol{m}}_t^{(k+1)} = \tanh\left(\beta \boldsymbol{G} \hat{\boldsymbol{m}}_t^{(k)} + \boldsymbol{y}_t - b_k \hat{\boldsymbol{m}}_t^{(k-1)}\right)$$

Idea of AMP: for fixed $\boldsymbol{v}, \boldsymbol{w}$, the vectors

$$(\boldsymbol{G}\boldsymbol{v}, \boldsymbol{G}\boldsymbol{w})$$

each have i.i.d. Gaussian coordinates. Covariance between $(\boldsymbol{G}\boldsymbol{v})_i$ and $(\boldsymbol{G}\boldsymbol{w})_i$ equals $\langle \boldsymbol{v}, \boldsymbol{w} \rangle$.

- Onsager term lets us apply this recursively to each $\hat{\boldsymbol{m}}_t^{(k+1)}$, despite accumulating dependence on $\boldsymbol{G}$.
- In spiked model, correlation with x_i also enters the recursion.

State Evolution for AMP

$$\hat{\boldsymbol{m}}_t^{(k+1)} = \tanh\left(\beta \boldsymbol{G}\hat{\boldsymbol{m}}_t^{(k)} + \boldsymbol{y}_t - b_k \hat{\boldsymbol{m}}_t^{(k-1)}\right)$$

Idea of AMP: for fixed $\boldsymbol{v}, \boldsymbol{w}$, the vectors

$$(\boldsymbol{G}\boldsymbol{v}, \boldsymbol{G}\boldsymbol{w})$$

each have i.i.d. Gaussian coordinates. Covariance between $(\boldsymbol{G}\boldsymbol{v})_i$ and $(\boldsymbol{G}\boldsymbol{w})_i$ equals $\langle \boldsymbol{v}, \boldsymbol{w}\rangle$.

- Onsager term lets us apply this recursively to each $\hat{\boldsymbol{m}}_t^{(k+1)}$, despite accumulating dependence on $\boldsymbol{G}$.
- In spiked model, correlation with x_i also enters the recursion.

State evolution: i-th coordinate of $\hat{\boldsymbol{m}}_t^{(k)}$ behaves like

$$\tanh(a_t^{(k)} x_i + b_t^{(k)} Z), \quad Z \sim \mathcal{N}(0,1).$$

- $(a_t^{(k)}, b_t^{(k)})$ determined recursively, converge to $(a_t^{\infty}, b_t^{\infty})$.

State Evolution for AMP

From (a_t^∞, b_t^∞), one can read off the asymptotic MSE

$$E_* = \lim_{k\to\infty} \operatorname*{p-lim}_{n\to\infty} \mathbb{E}\|\hat{\boldsymbol{m}}_t^{(k)} - \boldsymbol{x}\|_2^2.$$

State Evolution for AMP

From (a_t^∞, b_t^∞), one can read off the asymptotic MSE

$$E_* = \lim_{k\to\infty} \operatorname*{p-lim}_{n\to\infty} \mathbb{E}\|\hat{\boldsymbol{m}}_t^{(k)} - \boldsymbol{x}\|_2^2.$$

If we can show

$$E_* \approx MMSE(t) \equiv \mathbb{E}\|\boldsymbol{m}_t - \boldsymbol{x}\|_2^2,$$

then we conclude $\hat{\boldsymbol{m}}_t^{(k)} \approx \boldsymbol{m}_t$.

State Evolution for AMP

From (a_t^∞, b_t^∞), one can read off the asymptotic MSE

$$E_* = \lim_{k\to\infty} \operatorname*{p-lim}_{n\to\infty} \mathbb{E}\|\hat{\boldsymbol{m}}_t^{(k)} - \boldsymbol{x}\|_2^2.$$

If we can show

$$E_* \approx MMSE(t) \equiv \mathbb{E}\|\boldsymbol{m}_t - \boldsymbol{x}\|_2^2,$$

then we conclude $\hat{\boldsymbol{m}}_t^{(k)} \approx \boldsymbol{m}_t$.

I-MMSE Area Law [Guo-Shamai-Verdu 04, Deshpande-Abbe-Montanari 15]:

$$\int_0^\infty \mathrm{MMSE}(t)\,\mathrm{d}t = 2 \cdot \mathrm{Ent}(\boldsymbol{x}^*).$$

- Verify explicitly that $\int_0^\infty E_*(t)$ asympotically matches $\mathrm{Ent}(\boldsymbol{x}^*)$.

Conclusion of Step 1: Rough Estimate for $\boldsymbol{m}_t$

$$\hat{\boldsymbol{m}}_t^{(k+1)} = \tanh\left(\beta \boldsymbol{G}\hat{\boldsymbol{m}}_t^{(k)} + \boldsymbol{y}_t - b_k \hat{\boldsymbol{m}}_t^{(k-1)}\right),$$

Proposition (El Alaoui-Montanari-S 22)

For $\beta < 1$ and any $\varepsilon, t \geq 0$ there exists $k_0(t, \varepsilon)$ such that for all $k \geq k_0$,

$$\lim_{n \to \infty} \mathbb{P}\left[\|\hat{\boldsymbol{m}}_t^{(k)} - \boldsymbol{m}_t\| \leq \varepsilon\sqrt{n}\right] = 1.$$

Step 2: Refined Estimate of $\boldsymbol{m}_t$

Surprisingly, this is not quite enough.

- Two types of error: SDE δ-discretization and $\hat{\boldsymbol{m}}_t^{(k)} \approx \boldsymbol{m}_t$.
- Simply sending $(\delta, k) \to (0, \infty)$ doesn't work. Not Lipschitz enough.

Step 2: Refined Estimate of $\boldsymbol{m}_t$

Surprisingly, this is not quite enough.

- Two types of error: SDE δ-discretization and $\hat{\boldsymbol{m}}_t^{(k)} \approx \boldsymbol{m}_t$.
- Simply sending $(\delta, k) \to (0, \infty)$ doesn't work. Not Lipschitz enough.

Second step: by construction, $\hat{\boldsymbol{m}}_t^{(k)}$ is an approximate stationary point for the TAP free energy:

$$F_{TAP}(\boldsymbol{m}, \boldsymbol{y}_t) = -\frac{\beta}{2}\langle \boldsymbol{m}, \boldsymbol{G}\boldsymbol{m}\rangle - \langle \boldsymbol{y}_t, \boldsymbol{m}\rangle - \sum_{i=1}^{n} h(m_i).$$

- Refine $\hat{\boldsymbol{m}}_t^{(k)}$ to $\hat{\boldsymbol{m}}_t = \arg\min_{\boldsymbol{m}} F_{TAP}(\boldsymbol{m}, \boldsymbol{y}_t)$ via gradient descent.

Step 2: Refined Estimate of $\boldsymbol{m}_t$

Surprisingly, this is not quite enough.

- Two types of error: SDE δ-discretization and $\hat{\boldsymbol{m}}_t^{(k)} \approx \boldsymbol{m}_t$.
- Simply sending $(\delta, k) \to (0, \infty)$ doesn't work. Not Lipschitz enough.

Second step: by construction, $\hat{\boldsymbol{m}}_t^{(k)}$ is an approximate stationary point for the TAP free energy:

$$F_{TAP}(\boldsymbol{m}, \boldsymbol{y}_t) = -\frac{\beta}{2}\langle \boldsymbol{m}, \boldsymbol{G}\boldsymbol{m}\rangle - \langle \boldsymbol{y}_t, \boldsymbol{m}\rangle - \sum_{i=1}^{n} h(m_i).$$

- Refine $\hat{\boldsymbol{m}}_t^{(k)}$ to $\hat{\boldsymbol{m}}_t = \arg\min_{\boldsymbol{m}} F_{TAP}(\boldsymbol{m}, \boldsymbol{y}_t)$ via gradient descent.
- [Celentano 22]: F_{TAP} is strongly convex near $\boldsymbol{m}_t$ for $\beta < 1$.
  Implies $\boldsymbol{y}_t \mapsto \hat{\boldsymbol{m}}_t$ is C_β-Lipschitz. ($\boldsymbol{y}_t \mapsto \hat{\boldsymbol{m}}_t^{(k)}$ is C_β^k-Lipschitz.)

Some Last Remarks

This type of algorithm must be completely impractical, right?

Not quite...

Connection to Image Generation

Recall:

$$\boldsymbol{m}_t(\boldsymbol{y}_t) = \mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y}_t], \quad \boldsymbol{y}_t = t\boldsymbol{x}_t + \sqrt{t}\mathbf{g}, \quad \mathbf{g} \sim \mathcal{N}(0, I_n),$$
$$\boldsymbol{m}_t = \arg \min_{\phi:\mathbb{R}^n \to \mathbb{R}^n} \mathbb{E}[\|\phi(\boldsymbol{y}_t) - \boldsymbol{x}\|_2^2].$$

I.e.:

Bayes-optimal inversion of Gaussian noise suffices to sample.

Image Generation

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be i.i.d. natural images. Generate noisy versions $\boldsymbol{y}_i$.

Choose $\hat{\boldsymbol{m}}_t = \phi(\boldsymbol{y}_i)$ minimizing empirical loss

$$\frac{1}{n}\sum_{i=1}^{n} \|\phi(\boldsymbol{y}_i) - \boldsymbol{x}_i\|_2^2$$

...for $\phi \in \mathcal{F}$ constrained inside some **function class** such as **convolutional neural networks**.

Image Generation

These are diffusion models! [Song-Ermon 19], DALL-E 2, Imagen.

Image Generation

These are diffusion models! [Song-Ermon 19], DALL-E 2, Imagen.

- Equivalent setup: turn $\boldsymbol{x} \sim \mu$ into Gaussian noise with OU flow. Then simulate the time-reversal (corresponds to $\boldsymbol{y}_t/t$).
- Mean-estimation is done using "forward" sample paths.

Image Generation

These are diffusion models! [Song-Ermon 19], DALL-E 2, Imagen.

- Equivalent setup: turn $\boldsymbol{x} \sim \mu$ into Gaussian noise with OU flow. Then simulate the time-reversal (corresponds to $\boldsymbol{y}_t/t$).
- Mean-estimation is done using "forward" sample paths.

[Chen-Chewi-Li-Li-Salim-Zhang 22, Lee-Lu-Tan 22a,22b,22c]: estimating $\boldsymbol{m}_t$ in L^2 suffices for sampling if $\boldsymbol{y}_t \mapsto \boldsymbol{m}_t$ is globally Lipschitz.

- For us: proxy $\hat{\boldsymbol{m}}_t$ is **typically locally** Lipschitz near the sample path.

Summary

Stochastic localization for the SK model: interaction with high-dimensional probability enables a rigorous, end-to-end analysis.

Our algorithm produces Waserstein-approximate samples for $\beta < 1$. For $\beta > 1$, disorder chaos is a natural barrier for stable algorithms.

- What other distributions are stochastic localization sampleable?
- Sharp thresholds in related models.
  - **Shattering** may obstruct efficient sampling even when replica symmetric. Absent in SK model, expected for pure spherical p-spin.