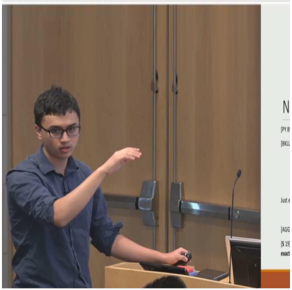
Sampling via Stochastic Localization

Andrea Montanari

Stanford University

July 18, 2022

Ahmed El Alaoui



Mark Sellke

Yuchen Wu

Outline

FOCS 2022 (with El Alaoui, Sellke)

In preparation (with Wu)

From Stochastic Localization to sampling

Sampling from μ

I want to generate

$$\boldsymbol{x}^* \sim \mu(\mathrm{d}\boldsymbol{x}) \quad \text{given } \mu \text{ probability distribution on } \mathbb{R}^{\boldsymbol{n}}\,.$$

- Non log-concave
- High-dimensional, $n \geq 100$

Sampling from μ

I want to generate

$$\boldsymbol{x}^* \sim \mu(\mathrm{d}\boldsymbol{x}) \quad \text{given} \;\; \mu \in \mathscr{P}(\mathbb{R}^{\boldsymbol{n}}),$$

- Non log-concave
- High-dimensional, $n \geq 100$

Example #1: Ising models (binary graphical models)

$$\mu(\mathrm{d}\boldsymbol{x}) = \frac{1}{\boldsymbol{Z}(\beta)}\, e^{\beta\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle/2}\, \nu_0^{\otimes n}(\mathrm{d}\boldsymbol{x}) \qquad \nu_0 := \mathsf{Unif}(\{+1,-1\})\,.$$

Example #2: Sparse inverse problems

$$\mu(\mathrm{d}\boldsymbol{x}) = \frac{1}{\boldsymbol{Z}(\beta)}\, \exp\Big\{ -\frac{\beta}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2\Big\}\, \overline{\nu}_s^{\otimes n}(\mathrm{d}\boldsymbol{x})\,,$$
$$\overline{\nu}_{\mathrm{s}} := (1-\mathrm{s})\delta_0 + s\phi(x)\mathrm{d}x\,.$$

$$\mu(\mathrm{d}\boldsymbol{x}) = \frac{1}{Z(\beta)}\, e^{\beta\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle/2}\, \nu_0^{\otimes n}(\mathrm{d}\boldsymbol{x}) \qquad \nu_0 := \mathsf{Unif}(\{+1,-1\})\,.$$

Monte Carlo approach
Random walk on the vertices of the hypercube $\{+1,+1\}^n$ with invariant measure μ

This talk
Stochastic process $\boldsymbol{m}_t \in \mathbb{R}^n$, $t \in [0,\infty]$:

$$\boldsymbol{m}_0 = \int \boldsymbol{x}\, \mu(\mathrm{d}\boldsymbol{x}) \;\;\longrightarrow\;\; \boldsymbol{x}^* = \boldsymbol{m}_\infty \sim \mu$$

$$\mu(\mathrm{d}\boldsymbol{x}) = \frac{1}{Z(\beta)}\, e^{\beta\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle/2}\, \nu_0^{\otimes n}(\mathrm{d}\boldsymbol{x}) \qquad \nu_0 := \mathsf{Unif}(\{+1,-1\})\,.$$

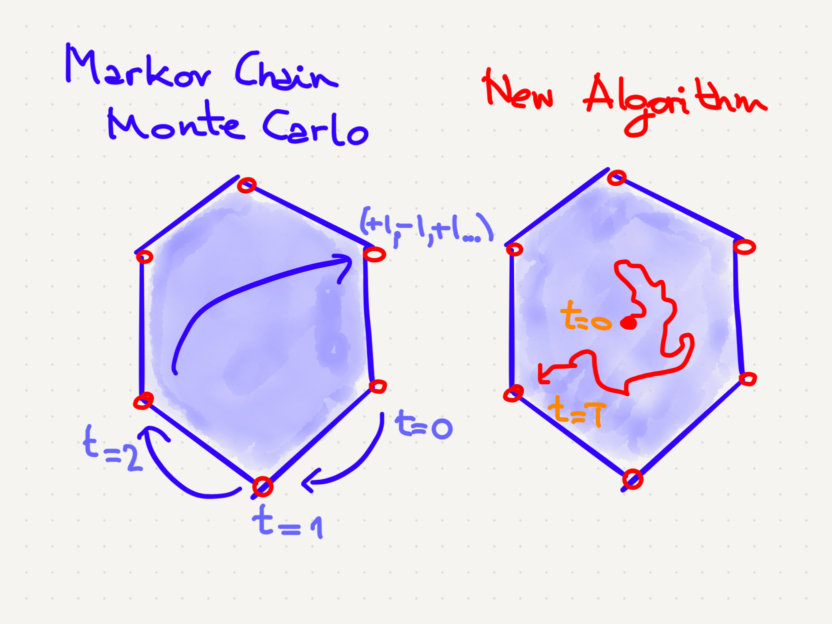
Monte Carlo approach

Random walk on the vertices of the hypercube $\{+1,+1\}^n$ with invariant measure μ

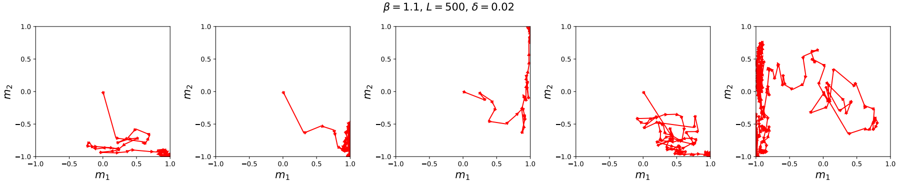
This talk

Stochastic process $\boldsymbol{m}_t \in \mathbb{R}^n$, $t \in [0,\infty]$:

$$\boldsymbol{m}_0 = \int \boldsymbol{x}\, \mu(\mathrm{d}\boldsymbol{x}) \;\;\longrightarrow\;\; \boldsymbol{x}^* = \boldsymbol{m}_\infty \sim \mu$$


Markov Chain
Monte Carlo
New Algorithm
(+1,-1,+1,...)
t=0
t=1
t=2
t=0
t=T

Appetizer



Constructing the process

Stochastic localization (Eldan 2016, 2020):

Stochastic process $(\mu_t)_{t\in[0,T]}$ taking values $\mathscr{P}(\mathbb{R}^n)$:
[If you prefer, think of $\mu_t \in \mathscr{P}(\{+1,-1\}^n)$]

Properties:

- $\mu_0 = \mu \quad \longrightarrow \quad \mu_T = \delta_{x^*},\ x^* \sim \mu.$
- μ_t is a martingale

El Alaoui, Montanari, 2021; Chen, Eldan, 2022

Constructing the process

Stochastic localization (Eldan 2016, 2020):

Stochastic process $(\mu_t)_{t\in[0,T]}$ taking values $\mathscr{P}(\mathbb{R}^n)$:
[If you prefer, think of $\mu_t \in \mathscr{P}(\{+1,-1\}^n)$]

Properties:

- $\mu_0 = \mu \quad \longrightarrow \quad \mu_T = \delta_{\boldsymbol{x}^*},\ \boldsymbol{x}^* \sim \mu.$
- μ_t is a martingale

El Alaoui, Montanari, 2021; Chen, Eldan, 2022

From stochastic localization to sampling

$$\mu_0 = \mu \;\; \longrightarrow \;\; \mu_T = \delta_{\boldsymbol{x}^*},\; \boldsymbol{x}^* \sim \mu\,.$$

- If I can generate the process $(\mu_t)_{t\in[0,T]}$, then I can sample.
- Can track the barycenter

$$\boldsymbol{m}_t := \int \boldsymbol{x}\,\mu_t(\mathrm{d}\boldsymbol{x}) \;\; \longrightarrow \;\; \boldsymbol{x}^* \sim \mu$$

Sounds crazy …right?

From stochastic localization to sampling

$$\mu_0 = \mu \quad \longrightarrow \quad \mu_T = \delta_{\boldsymbol{x}^*},\ \boldsymbol{x}^* \sim \mu\,.$$

- If I can generate the process $(\mu_t)_{t\in[0,T]}$, then I can sample.
- Can track the barycenter

$$\boldsymbol{m}_t := \int \boldsymbol{x}\,\mu_t(\mathrm{d}\boldsymbol{x}) \quad \longrightarrow \quad \boldsymbol{x}^* \sim \mu$$

Sounds crazy ...right?

One example of this program you all know

- Draw $\boldsymbol{x}^* \sim \mu$
- For $t \in \{0, 1, \dots, T = n\}$, define

$$\mu_t(\,\cdot\,) = \mathbb{P}_\mu\big(\boldsymbol{x} \in \,\cdot\, \big| x_1 = x_1^*, \dots, x_t = x_t^*\big)\,.$$

$$\mu_0 = \mu \;\longrightarrow\; \mu_n = \delta_{\boldsymbol{x}^*}$$

Does this give an algorithm?

$$\mu\big(x \in \cdot \,\big|x_1^*, \dots, x_t^*\big) := \mathbb{P}_\mu\big(x \in \cdot \,\big|x_1 = x_1^*, \dots, x_t = x_t^*\big)\,.$$

Algorithm: Sequential Sampling

Data: Probability measure μ
Result: Sample $x^* \sim \mu$
for $t \in \{1, \dots, n\}$ **do**
  Sample $x_t^* \sim \mu\big(x_t \in \cdot \,\big|x_1^*, \dots, x_{t-1}^*\big)$;
end
return x^*

Criticism

Algorithm: Sequential Sampling

Data: Probability measure μ
Result: Sample $\boldsymbol{x}^* \sim \mu$
for $t \in \{1, \ldots, n\}$ **do**
  Sample $x_t^* \sim \mu\big(x_t \in \cdot \,\big|x_1^*, \ldots, x_{t-1}^*\big)$;
end
return $\boldsymbol{x}^*$

- Requires an 'oracle' to compute $\mu\big(\,\cdot\, \big|x_1^*, \ldots, x_{t-1}^*\big)$
- Particularly difficult for continuous variables.
- Ordering is arbitrary.

A more interesting process

$$y_t = t x^* + B_t \quad \sim \mathsf{N}(t \cdot x_*, t\, \mathrm{I}_d)\,,$$

$$x^* \sim \mu\,, \quad B_\cdot \sim \text{Brownian Motion}, \quad x^* \perp\!\!\!\perp B_\cdot\,.$$

$$\mu_t(\,\cdot\,) = \mathbb{P}_\mu\big(x \in \cdot\,\big|y_t\big)\,, \quad t \in [0, \infty]$$

$$\mu_0 = \mu \;\longrightarrow\; \mu_\infty = \delta_{x^*}$$

A more interesting process

$$\boldsymbol{y}_t = t\boldsymbol{x}^* + \boldsymbol{B}_t \quad \sim \mathsf{N}(t \cdot \boldsymbol{x}_*, t\,\mathrm{I}_d)\,,$$

$$\boldsymbol{x}^* \sim \mu\,, \;\; \boldsymbol{B}_\cdot \sim \text{Brownian Motion}, \;\; \boldsymbol{x}^* \perp\!\!\!\perp \boldsymbol{B}_\cdot\,.$$

$$\mu_t(\,\cdot\,) = \mathbb{P}_\mu\big(\boldsymbol{x} \in \,\cdot\, \big|\boldsymbol{y}_t\big)\,, \quad t \in [0, \infty]$$

$$\mu_0 = \mu \;\longrightarrow\; \mu_\infty = \delta_{x^*}$$

A more interesting process

$$\boldsymbol{y}_t = t\boldsymbol{x}^* + \boldsymbol{B}_t \quad \sim \mathsf{N}(t \cdot \boldsymbol{x}_*, t\,\mathrm{I}_d)\,,$$

$$\boldsymbol{x}^* \sim \mu\,, \;\; \boldsymbol{B}_{\cdot} \sim \text{Brownian Motion}, \;\; \boldsymbol{x}^* \perp\!\!\!\perp \boldsymbol{B}_{\cdot}\,.$$

$$\mu_t(\,\cdot\,) = \mathbb{P}_{\mu}\big(\boldsymbol{x} \in \,\cdot\, \big| \boldsymbol{y}_t\big)\,, \quad t \in [0, \infty]$$

$$\mu_0 = \mu \;\longrightarrow\; \mu_\infty = \delta_{\boldsymbol{x}^*}$$

Does this give an algorithm?

$$y_t = tx^* + B_t\,, \quad (x^* \sim \mu) \perp\!\!\!\perp B.$$

Proposition

There exists a Brownian motion G. such that y_t solves the SDE:

$$dy_t = m_t(y_t)\,dt + dG_t\,,$$

$$\text{where} \quad m_t(y) := \mathbb{E}\{x^*|y_t = y\} = \int x\,\mu_t(dx)\,.$$

Proof: Exercise (Coming up).

Does this give an algorithm?

$$\boldsymbol{y}_t = t\boldsymbol{x}^* + \boldsymbol{B}_t\,, \quad (\boldsymbol{x}^* \sim \mu) \perp\!\!\!\perp \boldsymbol{B}.$$

Proposition

There exists a Brownian motion $\boldsymbol{G}$. *such that* $\boldsymbol{y}_t$ *solves the SDE:*

$$d\boldsymbol{y}_t = \boldsymbol{m}_t(\boldsymbol{y}_t)\, dt + d\boldsymbol{G}_t\,,$$

$$\textit{where} \quad \boldsymbol{m}_t(\boldsymbol{y}) := \mathbb{E}\{\boldsymbol{x}^* | \boldsymbol{y}_t = \boldsymbol{y}\} = \int \boldsymbol{x}\; \mu_t(d\boldsymbol{x})\,.$$

Proof: Exercise (Coming up).

Is there an algorithm?

$$\mathrm{d}\boldsymbol{y}_t = \boldsymbol{m}_t(\boldsymbol{y}_t)\,\mathrm{d}t + \mathrm{d}\boldsymbol{G}_t$$

Algorithm: Sampling via SL

Data: Probability measure μ
Result: Sample $\boldsymbol{x}^* \sim \mu$
for $t \in [0, \delta, 2\delta, \dots, T-\delta]$ **do**
    Sample $\boldsymbol{g}_t \sim \mathsf{N}(0, \mathrm{I}_d)$ independent of past;
    Set $\boldsymbol{y}_{t+\delta} = \boldsymbol{y}_t + \hat{\boldsymbol{m}}_t(\boldsymbol{y}_t)\,\delta + \sqrt{\delta}\,\boldsymbol{g}_t$;
end
Set $\boldsymbol{x}^* =$Round$(\hat{\boldsymbol{m}}_T(\boldsymbol{y}_T))$;
return $\boldsymbol{x}^*$

- Requires an approximation of $\hat{\boldsymbol{m}}_t(\boldsymbol{y}) \approx \mathbb{E}\{\boldsymbol{x}|\boldsymbol{y}_t = \boldsymbol{y}\}$

Every talk should contain a proof and a joke

$$\mathrm{d}\boldsymbol{y}_t = \boldsymbol{m}_t(\boldsymbol{y}_t)\,\mathrm{d}t + \mathrm{d}\boldsymbol{G}_t$$

Every talk should contain a proof and a joke

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

$$\boldsymbol{G}_t := \boldsymbol{y}_t - \int_0^t \boldsymbol{m}_s(\boldsymbol{y}_s)\,\mathrm{d}s$$

-

$$\begin{aligned}\boldsymbol{G}_{t+\delta} - \boldsymbol{G}_t &:= \boldsymbol{B}_{t+\delta} - \boldsymbol{B}_t + \delta\boldsymbol{x}^* - \int_t^{t+\delta} \boldsymbol{m}_s(\boldsymbol{y}_s)\,\mathrm{d}s \\ &= \boldsymbol{B}_{t+\delta} - \boldsymbol{B}_t + \delta\boldsymbol{x}^* - \delta\mathbb{E}\{\boldsymbol{x}^*|\boldsymbol{y}_t\} + o(\delta)\,.\end{aligned}$$

-

$$\begin{aligned}&\mathbb{E}\{\boldsymbol{G}_{t+\delta} - \boldsymbol{G}_t|\mathcal{F}_t^{\boldsymbol{B}}\} = o(\delta)\,,\\ &\mathbb{E}\{(\boldsymbol{G}_{t+\delta} - \boldsymbol{G}_t)(\boldsymbol{G}_{t+\delta} - \boldsymbol{G}_t)^{\mathsf{T}}|\mathcal{F}_t^{\boldsymbol{B}}\} = \delta\,\mathrm{I}_d + o(\delta)\,.\end{aligned}$$

…by Levy's characterization of BM…

Every talk should contain a proof and a joke

- $y_t = tx^* + B_t$

$$G_t := y_t - \int_0^t m_s(y_s)\,\mathrm{d}s$$

-

$$\begin{aligned} G_{t+\delta} - G_t &:= B_{t+\delta} - B_t + \delta x^* - \int_t^{t+\delta} m_s(y_s)\,\mathrm{d}s \\ &= B_{t+\delta} - B_t + \delta x^* - \delta \mathbb{E}\{x^* | y_t\} + o(\delta)\,. \end{aligned}$$

-

$$\begin{aligned} &\mathbb{E}\{G_{t+\delta} - G_t | \mathcal{F}_t^{B}\} = o(\delta)\,, \\ &\mathbb{E}\{(G_{t+\delta} - G_t)(G_{t+\delta} - G_t)^{\mathsf{T}} | \mathcal{F}_t^{B}\} = \delta\,\mathrm{I}_d + o(\delta)\,. \end{aligned}$$

…by Levy's characterization of BM…

Every talk should contain a proof and a joke

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

$$\boldsymbol{G}_t := \boldsymbol{y}_t - \int_0^t \boldsymbol{m}_s(\boldsymbol{y}_s)\,\mathrm{d}s$$

-

$$\begin{aligned}\boldsymbol{G}_{t+\delta} - \boldsymbol{G}_t &:= \boldsymbol{B}_{t+\delta} - \boldsymbol{B}_t + \delta\boldsymbol{x}^* - \int_t^{t+\delta} \boldsymbol{m}_s(\boldsymbol{y}_s)\,\mathrm{d}s \\ &= \boldsymbol{B}_{t+\delta} - \boldsymbol{B}_t + \delta\boldsymbol{x}^* - \delta\mathbb{E}\{\boldsymbol{x}^*|\boldsymbol{y}_t\} + o(\delta)\,.\end{aligned}$$

-

$$\begin{aligned}&\mathbb{E}\{\boldsymbol{G}_{t+\delta} - \boldsymbol{G}_t|\mathcal{F}_t^{\boldsymbol{B}}\} = o(\delta)\,,\\ &\mathbb{E}\{(\boldsymbol{G}_{t+\delta} - \boldsymbol{G}_t)(\boldsymbol{G}_{t+\delta} - \boldsymbol{G}_t)^{\mathsf{T}}|\mathcal{F}_t^{\boldsymbol{B}}\} = \delta\,\mathrm{I}_d + o(\delta)\,.\end{aligned}$$

…by Levy's characterization of BM…

Every talk should contain a proof and a joke

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

$$\boldsymbol{G}_t := \boldsymbol{y}_t - \int_0^t \boldsymbol{m}_s(\boldsymbol{y}_s)\,\mathrm{d}s$$

- $$\begin{aligned}\boldsymbol{G}_{t+\delta} - \boldsymbol{G}_t &:= \boldsymbol{B}_{t+\delta} - \boldsymbol{B}_t + \delta\boldsymbol{x}^* - \int_t^{t+\delta} \boldsymbol{m}_s(\boldsymbol{y}_s)\,\mathrm{d}s \\ &= \boldsymbol{B}_{t+\delta} - \boldsymbol{B}_t + \delta\boldsymbol{x}^* - \delta\mathbb{E}\{\boldsymbol{x}^*|\boldsymbol{y}_t\} + o(\delta)\,.\end{aligned}$$

- $$\begin{aligned}&\mathbb{E}\{\boldsymbol{G}_{t+\delta} - \boldsymbol{G}_t|\mathcal{F}_t^{\boldsymbol{B}}\} = o(\delta)\,,\\ &\mathbb{E}\{(\boldsymbol{G}_{t+\delta} - \boldsymbol{G}_t)(\boldsymbol{G}_{t+\delta} - \boldsymbol{G}_t)^{\mathsf{T}}|\mathcal{F}_t^{\boldsymbol{B}}\} = \delta\,\mathrm{I}_d + o(\delta)\,.\end{aligned}$$

…by Levy's characterization of BM…

Where do we stand?

- General sampling procedure
- Requires an approximation $\hat{\boldsymbol{m}}_t(\boldsymbol{y}) \approx \mathbb{E}\{\boldsymbol{x}|\boldsymbol{y}_t = \boldsymbol{y}\}$

What comes next?

- Two challenge problems:
  - Sherrington-Kirkpatrick model
  - $\mathbb{Z}_2$-synchronization

Where do we stand?

- General sampling procedure
- Requires an approximation $\hat{\boldsymbol{m}}_t(\boldsymbol{y}) \approx \mathbb{E}\{\boldsymbol{x}|\boldsymbol{y}_t = \boldsymbol{y}\}$

What comes next?

- Two challenge problems:
  - Sherrington-Kirkpatrick model
  - $\mathbb{Z}_2$-synchronization

Sherrington-Kirkpatrick model

Sherrington-Kirkpatrick model

Probability distribution over $x \in \{+1, -1\}^n$

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

$\boldsymbol{W} \sim \mathrm{GOE}(n)$:

- $\boldsymbol{W} = \boldsymbol{W}^{\mathsf{T}}$, $(W_{ij})_{i\le j}$ independent.
- $i<j$: $W_{ij} \sim \mathsf{N}(0, 1/n)$, $W_{ii} \sim \mathsf{N}(0, 2/n)$

Want to sample from $\mu_{\boldsymbol{W},\beta}$ *conditional on* $\boldsymbol{W} \sim \mathrm{GOE}(n)$.

Sherrington-Kirkpatrick model

Probability distribution over $x \in \{+1, -1\}^n$

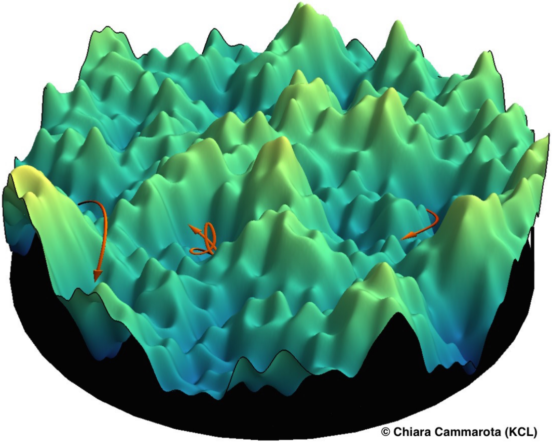
$$\mu_{\boldsymbol{W},\beta}(\boldsymbol{x}) = \frac{1}{Z_n(\beta)}\, e^{\beta\langle \boldsymbol{x}, \boldsymbol{W}\boldsymbol{x}\rangle/2}\,.$$

$\boldsymbol{W} \sim \mathrm{GOE}(n)$:

- $\boldsymbol{W} = \boldsymbol{W}^{\mathsf{T}}$, $(W_{ij})_{i\le j}$ independent.
- $i < j$: $W_{ij} \sim \mathsf{N}(0, 1/n)$, $W_{ii} \sim \mathsf{N}(0, 2/n)$

Want to sample from $\mu_{\boldsymbol{W},\beta}$ *conditional on* $\boldsymbol{W} \sim \mathrm{GOE}(n)$.

Can we sample with this energy function?



History: Structure of $\mu_{\boldsymbol{W},\beta}$

1975 Sherrington and Kirkpatrick: simplified model for magnetic materials.

1980 Parisi 'solution' (asymptotics of $Z_n(\beta)$)
Replica method: Non-rigorous.

⋮

A quarter century of mathematical efforts

⋮

2005 Talagrand proves Parisi's formula. (More general proof: Panchenko)

2021 *Parisi receives the Nobel Prize for Physics.*

History: Structure of $\mu_{\boldsymbol{W},\beta}$

1975 Sherrington and Kirkpatrick: simplified model for magnetic materials.

1980 Parisi 'solution' (asymptotics of $Z_n(\beta)$)
Replica method: Non-rigorous.

⋮

A quarter century of mathematical efforts

⋮

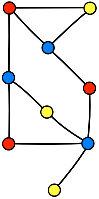
2005 Talagrand proves Parisi's formula. (More general proof: Panchenko)

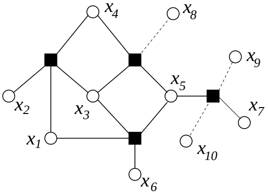
2021 *Parisi receives the Nobel Prize for Physics.*

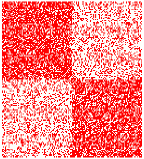
Why interesting?

Prototype of complex/random/rough probability measures

- Proper colorings of random graph; Random K-satisfiability assignments; High-dimensional statistics posteriors; Stochastic block model; Coding theory; Glassy materials; …







History: Sampling

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

Physicists: Can be sampled efficiently for any $\beta < 1$ (and not for $\beta > 1$)

⋮

No progress for 40 years

⋮

Bodineau, Bauerschmidt, 2020, Eldan, Koehler, Zeitouni 2021:
Sampling algorithms with guarantees $\beta < 1/4$

Our contribution: Efficient sampling (in weaker sense) for $\beta < 1/2$
(New algorithm)

History: Sampling

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

Physicists: Can be sampled efficiently for any $\beta < 1$ (and not for $\beta > 1$)

⋮

No progress for 40 years

⋮

Bodineau, Bauerschmidt, 2020, Eldan, Koehler, Zeitouni 2021:
Sampling algorithms with guarantees $\beta < 1/4$

Our contribution: Efficient sampling (in weaker sense) for $\beta < 1/2$
(New algorithm)

Metric for our approximation guarantee

Given $\mu_1, \mu_2 \in \mathscr{P}(\mathbb{R}^n)$,

$$W_{2,n}(\mu_1, \mu_2) := \left\{ \inf_{\gamma \in \mathsf{Coupl}(\mu_1,\mu_2)} \int \frac{1}{\boldsymbol{n}} \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2 \, \gamma(\mathrm{d}\boldsymbol{x}_1, \mathrm{d}\boldsymbol{x}_2) \right\}^{1/2}$$

Intuition: If $f \in [-1,+1]^n \to \mathbb{R}$ is L/n-Lipschitz

$$\left| \int f(x_1)\mu_1(\mathrm{d}x_1) - \int f(x_2)\mu_2(\mathrm{d}x_2) \right| \le L\, W_{2,n}(\mu_1, \mu_2)\,.$$

Metric for our approximation guarantee

Given $\mu_1, \mu_2 \in \mathscr{P}(\mathbb{R}^n)$,

$$W_{2,n}(\mu_1, \mu_2) := \Big\{ \inf_{\gamma \in \mathsf{Coupl}(\mu_1,\mu_2)} \int \frac{1}{n} \|x_1 - x_2\|_2^2 \, \gamma(\mathrm{d}x_1, \mathrm{d}x_2) \Big\}^{1/2}$$

Intuition: If $f \in [-1,+1]^n \to \mathbb{R}$ is L/n-Lipschitz

$$\Big| \int f(x_1) \mu_1(\mathrm{d}x_1) - \int f(x_2) \mu_2(\mathrm{d}x_2) \Big| \le L \, W_{2,n}(\mu_1, \mu_2) \, .$$

Main result

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

Theorem (El Alaoui, M, Sellke, 2022)

For any $\beta < 1/2$, there exists an algorithm with complexity $O(n^2)$, that takes as input $\boldsymbol{W} \sim \mathrm{GOE}(n)$ and outputs $\boldsymbol{x} \sim \mu^{\rm alg}$ such that (with high probability with respect to $\boldsymbol{W}$), we have

$$W_{2,n}(\mu^{\rm alg}_{\boldsymbol{W},\beta}, \mu_{\boldsymbol{W},\beta}) \le o_n(1)\,.$$

- The algorithm is Stochastic-Localization Sampling
- $\beta < 1/2$ is a proof artifact: should work for all $\beta < 1$.
- Need to define $\hat{\boldsymbol{m}}_t(\boldsymbol{y})$ so that $\hat{\boldsymbol{m}}_t(\boldsymbol{y}) \approx \mathbb{E}[\boldsymbol{x}|\boldsymbol{y}_t = \boldsymbol{y}]$.

Main result

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

Theorem (El Alaoui, M, Sellke, 2022)

For any $\beta < 1/2$, there exists an algorithm with complexity $O(n^2)$, that takes as input $\boldsymbol{W} \sim \mathrm{GOE}(n)$ and outputs $\boldsymbol{x} \sim \mu^{\rm alg}$ such that (with high probability with respect to $\boldsymbol{W}$), we have

$$W_{2,n}(\mu^{\rm alg}_{\boldsymbol{W},\beta}, \mu_{\boldsymbol{W},\beta}) \le o_n(1)\,.$$

- The algorithm is Stochastic-Localization Sampling
- $\beta < 1/2$ is a proof artifact: should work for all $\beta < 1$.
- Need to define $\hat{\boldsymbol{m}}_t(\boldsymbol{y})$ so that $\hat{\boldsymbol{m}}_t(\boldsymbol{y}) \approx \mathbb{E}[\boldsymbol{x}|\boldsymbol{y}_t = \boldsymbol{y}]$.

Main result

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

Theorem (El Alaoui, M, Sellke, 2022)

For any $\beta < 1/2$, there exists an algorithm with complexity $O(n^2)$, that takes as input $\boldsymbol{W} \sim \mathrm{GOE}(n)$ and outputs $\boldsymbol{x} \sim \mu^{\mathrm{alg}}$ such that (with high probability with respect to $\boldsymbol{W}$), we have

$$W_{2,n}(\mu^{\mathrm{alg}}_{\boldsymbol{W},\beta}, \mu_{\boldsymbol{W},\beta}) \le o_n(1)\,.$$

- The algorithm is Stochastic-Localization Sampling
- $\beta < 1/2$ is a proof artifact: should work for all $\beta < 1$.
- Need to define $\hat{\boldsymbol{m}}_t(\boldsymbol{y})$ so that $\hat{\boldsymbol{m}}_t(\boldsymbol{y}) \approx \mathbb{E}[\boldsymbol{x}|\boldsymbol{y}_t = \boldsymbol{y}]$.

Hardness

Theorem (El Alaoui, M, Sellke, 2022)

Let $\mu^{\rm alg}_{\boldsymbol{W},\beta}$ *be the law of the output* $\mathsf{ALG}_n(\boldsymbol{W},\beta,\omega)$ *conditional on* $\boldsymbol{W}$*. If* ALG *is a stable algorithm, then, for any* $\beta > 1$*:*

$$\liminf_{n\to\infty} \mathbb{E}\big[W_{2,n}(\mu^{\rm alg}_{\boldsymbol{W},\beta},\ \mu_{\boldsymbol{W},\beta})\big] > 0\,.$$

Posterior expectation via AMP

$\hat{\boldsymbol{m}}_t(\boldsymbol{y}) \approx \mathbb{E}[\boldsymbol{x}|\boldsymbol{y}_t = \boldsymbol{y}]$

- Approximate Message Passing
- General class of iterative algorithms
- Rigorous (sharp) high-dimensional asymptotics

Next: Partial heuristic justification

Donoho, Maleki, M, 2009; Bolthausen, 2013; Bayati, M, 2011; Javanmard, M, 2013; ...

Form of the posterior

$$\begin{aligned}\mu_t(\boldsymbol{x}) &= \mathbb{P}_{\beta,\boldsymbol{W}}(\boldsymbol{x}|\boldsymbol{y}_t) \\ &= \frac{1}{Z_t'} \exp\left\{\frac{\beta}{2}\langle \boldsymbol{x}, \boldsymbol{W}\boldsymbol{x}\rangle - \frac{1}{2t}\|\boldsymbol{x} - t\boldsymbol{y}_t\|_2^2\right\} \\ &= \frac{1}{Z_t} \exp\left\{\frac{\beta}{2}\langle \boldsymbol{x}, \boldsymbol{W}\boldsymbol{x}\rangle + \langle \boldsymbol{y}_t, \boldsymbol{x}\rangle\right\}\end{aligned}$$

Form of the posterior

$$\begin{aligned}\mu_t(\boldsymbol{x}) &= \mathbb{P}_{\beta,\boldsymbol{W}}(\boldsymbol{x}|\boldsymbol{y}_t)\\ &= \frac{1}{Z_t'}\,\exp\Big\{\frac{\beta}{2}\langle\boldsymbol{x},\boldsymbol{W}\boldsymbol{x}\rangle - \frac{1}{2t}\|\boldsymbol{x}-t\boldsymbol{y}_t\|_2^2\Big\}\\ &= \frac{1}{Z_t}\,\exp\Big\{\frac{\beta}{2}\langle\boldsymbol{x},\boldsymbol{W}\boldsymbol{x}\rangle + \langle\boldsymbol{y}_t,\boldsymbol{x}\rangle\Big\}\end{aligned}$$

Posterior expectation calculation

$$\mu_t(\boldsymbol{x}) = \frac{1}{Z_t} \exp\left\{ \frac{\beta}{2} \langle \boldsymbol{x}, \boldsymbol{W}\boldsymbol{x} \rangle + \langle \boldsymbol{y}_t, \boldsymbol{x} \rangle \right\}$$

Naive mean field

$$\mu_t(x_i) \propto \sum_{x_{-i}} e^{x_i[\beta\langle \boldsymbol{W}_{i,\cdot}, x_{-i}\rangle + y_{t,i}]} \exp\left\{ \frac{\beta}{2} \langle x_{-i}, \boldsymbol{W}_{-i,-i} x_{-i} \rangle + \langle y_{t,-i}, x_{-i} \rangle \right\}$$

$$m_{t,i} := \sum_{x_i} \mu_t(x_i) = \sum_{x} \tanh\left(\beta \langle \boldsymbol{W}_{i,-i}, x_{-i} \rangle + y_{t,i}\right) \mu_t(x)$$

Naive mean field

$$\hat{m}_{t,i} = \tanh\left(\beta \langle \boldsymbol{W}_{i,-i}, \hat{\boldsymbol{m}}_{t,-i} \rangle + y_{t,i}\right)$$

Assumes that $\langle \boldsymbol{W}_{i,-i}, x_{-i} \rangle$ concentrates under μ_t.

Posterior expectation calculation

$$\mu_t(\boldsymbol{x}) = \frac{1}{Z_t} \exp\left\{\frac{\beta}{2}\langle \boldsymbol{x}, \boldsymbol{W}\boldsymbol{x}\rangle + \langle \boldsymbol{y}_t, \boldsymbol{x}\rangle\right\}$$

Naive mean field

$$\mu_t(x_i) \propto \sum_{\boldsymbol{x}_{-i}} e^{x_i[\beta\langle \boldsymbol{W}_{i,\cdot}, \boldsymbol{x}_{-i}\rangle + y_{t,i}]} \exp\left\{\frac{\beta}{2}\langle \boldsymbol{x}_{-i}, \boldsymbol{W}_{-i,-i}\boldsymbol{x}_{-i}\rangle + \langle \boldsymbol{y}_{t,-i}, \boldsymbol{x}_{-i}\rangle\right\}$$

$$m_{t,i} := \sum_{x_i} \mu_t(x_i) = \sum_{x} \tanh\left(\beta\langle \boldsymbol{W}_{i,-i}, x_{-i}\rangle + y_{t,i}\right)\mu_t(x)$$

Naive mean field

$$\hat{m}_{t,i} = \tanh\left(\beta\langle \boldsymbol{W}_{i,-i}, \hat{\boldsymbol{m}}_{t,-i}\rangle + y_{t,i}\right)$$

Assumes that $\langle \boldsymbol{W}_{i,-i}, x_{-i}\rangle$ concentrates under μ_t.

Posterior expectation calculation

$$\mu_t(\boldsymbol{x}) = \frac{1}{Z_t}\,\exp\left\{\frac{\beta}{2}\langle \boldsymbol{x}, \boldsymbol{W}\boldsymbol{x}\rangle + \langle \boldsymbol{y}_t, \boldsymbol{x}\rangle\right\}$$

Naive mean field

$$\mu_t(x_i) \propto \sum_{\boldsymbol{x}_{-i}} e^{x_i[\beta\langle \boldsymbol{W}_{i,\cdot}, \boldsymbol{x}_{-i}\rangle + y_{t,i}]}\exp\left\{\frac{\beta}{2}\langle \boldsymbol{x}_{-i}, \boldsymbol{W}_{-i,-i}\boldsymbol{x}_{-i}\rangle + \langle \boldsymbol{y}_{t,-i}, \boldsymbol{x}_{-i}\rangle\right\}$$

$$m_{t,i} := \sum_{x_i} \mu_t(x_i) = \sum_{\boldsymbol{x}} \tanh\left(\beta\langle \boldsymbol{W}_{i,-i}, \boldsymbol{x}_{-i}\rangle + y_{t,i}\right)\mu_t(\boldsymbol{x})$$

Naive mean field

$$\hat{m}_{t,i} = \tanh\left(\beta\langle \boldsymbol{W}_{i,-i}, \hat{\boldsymbol{m}}_{t,-i}\rangle + y_{t,i}\right)$$

Assumes that $\langle \boldsymbol{W}_{i,-i}, \boldsymbol{x}_{-i}\rangle$ concentrates under μ_t.

Posterior expectation calculation

$$\mu_t(\boldsymbol{x}) = \frac{1}{Z_t} \exp\left\{\frac{\beta}{2}\langle \boldsymbol{x}, \boldsymbol{W}\boldsymbol{x}\rangle + \langle \boldsymbol{y}_t, \boldsymbol{x}\rangle\right\}$$

Naive mean field

$$\mu_t(x_i) \propto \sum_{\boldsymbol{x}_{-i}} e^{x_i[\beta\langle \boldsymbol{W}_{i,\cdot}, \boldsymbol{x}_{-i}\rangle + y_{t,i}]} \exp\left\{\frac{\beta}{2}\langle \boldsymbol{x}_{-i}, \boldsymbol{W}_{-i,-i}\boldsymbol{x}_{-i}\rangle + \langle \boldsymbol{y}_{t,-i}, \boldsymbol{x}_{-i}\rangle\right\}$$

$$m_{t,i} := \sum_{x_i} \mu_t(x_i) = \sum_{\boldsymbol{x}} \tanh\left(\beta\langle \boldsymbol{W}_{i,-i}, \boldsymbol{x}_{-i}\rangle + y_{t,i}\right)\mu_t(\boldsymbol{x})$$

Naive mean field

$$\hat{m}_{t,i} = \tanh\left(\beta\langle \boldsymbol{W}_{i,-i}, \hat{\boldsymbol{m}}_{t,-i}\rangle + y_{t,i}\right)$$

Assumes that $\langle \boldsymbol{W}_{i,-i}, \boldsymbol{x}_{-i}\rangle$ concentrates under μ_t.

From naive mean field to AMP

Naive mean field

$$\hat{\boldsymbol{m}}_{t,i} = \tanh\left(\beta\langle \boldsymbol{W}_{i,-i}, \hat{\boldsymbol{m}}_{t,-i}\rangle + y_{t,i}\right),$$
$$\hat{\boldsymbol{m}}_t \approx \tanh\left(\beta \boldsymbol{W}\hat{\boldsymbol{m}}_t + \boldsymbol{y}_t\right).$$

Naive mean field iteration

$$\hat{\boldsymbol{m}}_t^{(k+1)} = \tanh\left(\beta \boldsymbol{W}\hat{\boldsymbol{m}}_t^{(k)} + \boldsymbol{y}_t\right), \quad \hat{\boldsymbol{m}}_t^{(0)} = \boldsymbol{0}$$

From naive mean field to AMP

Naive mean field iteration

$$\hat{\boldsymbol{m}}_t^{(k+1)} = \tanh\left(\beta \boldsymbol{W} \hat{\boldsymbol{m}}_t^{(k)} + \boldsymbol{y}_t\right), \quad \hat{\boldsymbol{m}}_0^{(k)} = \boldsymbol{0}$$

AMP (Approximate Message Passing)

$$\hat{\boldsymbol{m}}_t^{(k+1)} = \tanh\left(\beta \boldsymbol{W} \hat{\boldsymbol{m}}_t^{(k)} + \boldsymbol{y}_t - \mathsf{b}_k \hat{\boldsymbol{m}}^{(k-1)}\right),$$
$$\mathsf{b}_k := \beta^2\left(1 - \frac{1}{n}\|\boldsymbol{m}^{(k)}\|_2^2\right)$$

- 'Onsager' term $-\mathsf{b}_k \hat{\boldsymbol{m}}^{(k-1)}$ accounts for non-Gaussianity in $\boldsymbol{W}\hat{\boldsymbol{m}}_t^{(k)}$.
- Thouless, Anderson, Palmer , 1977

From naive mean field to AMP

Naive mean field iteration

$$\hat{\boldsymbol{m}}_t^{(k+1)} = \tanh\left(\beta \boldsymbol{W} \hat{\boldsymbol{m}}_t^{(k)} + \boldsymbol{y}_t\right), \quad \hat{\boldsymbol{m}}_0^{(k)} = \boldsymbol{0}$$

AMP (Approximate Message Passing)

$$\hat{\boldsymbol{m}}_t^{(k+1)} = \tanh\left(\beta \boldsymbol{W} \hat{\boldsymbol{m}}_t^{(k)} + \boldsymbol{y}_t - \mathsf{b}_k \hat{\boldsymbol{m}}^{(k-1)}\right),$$

$$\mathsf{b}_k := \beta^2\left(1 - \frac{1}{n}\|\boldsymbol{m}^{(k)}\|_2^2\right)$$

- 'Onsager' term $-\mathsf{b}_k \hat{\boldsymbol{m}}^{(k-1)}$ accounts for non-Gaussianity in $\boldsymbol{W}\hat{\boldsymbol{m}}_t^{(k)}$.
- Thouless, Anderson, Palmer , 1977

Accuracy of AMP

Proposition (Abbe, Deshpande, M, 2017; El Alaoui, M, Sellke, 2022)

Let $\boldsymbol{m}_t := \sum_{\boldsymbol{x}} \boldsymbol{x}\, \mu_t(\boldsymbol{x})$*, with* $\beta < 1$*. Then, for any* $t \in [0, \infty)$*,* $\varepsilon \in (0, 1)$*, there exists* $k_0 = k_0(t, \varepsilon)$ *such that, for any* $k \geq k_0$*:*

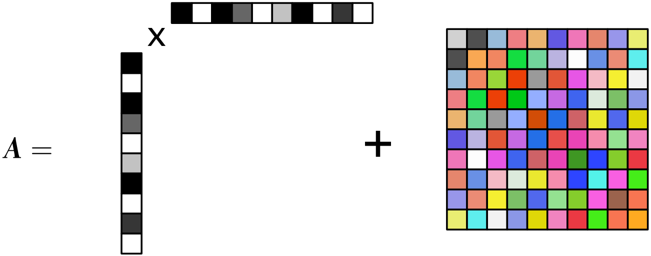
$$\operatorname*{p-lim}_{n\to\infty} \frac{1}{n} \big\| \hat{\boldsymbol{m}}_t^{(k)} - \boldsymbol{m}_t \big\|_2^2 \leq \varepsilon \,.$$

- We set $\hat{\boldsymbol{m}}(\boldsymbol{y}_t) := \hat{\boldsymbol{m}}_t^{(K)}$ for some large K.

In the proof of main result, tweaking of $\hat{\boldsymbol{m}}_t^{(k)}$ for technical reasons.

$\mathbb{Z}_2$-synchronization

Rank-one matrix estimation



$$A = \frac{\beta}{n} x_0 x_0^{\mathsf{T}} + W$$

Low-rank matrix estimation

$$\boldsymbol{A} = \frac{\beta}{n}\boldsymbol{x}_0\boldsymbol{x}_0^{\mathsf{T}} + \boldsymbol{W}\,.$$

- Signal $\boldsymbol{x}_0 \sim \mu_0^{\otimes n}$ $\qquad ((x_{0,i})_{i\le n} \sim_{iid} \mu_0)$
- Noise $\boldsymbol{W} \sim \mathrm{GOE}(n) \quad \perp\!\!\!\perp \quad \boldsymbol{x}_0$
- β = signal-to-noise ratio
- **Given $\boldsymbol{A}$, want to estimate $\boldsymbol{x}_0$**

$\mathbb{Z}_2$-synchronization

$$\boldsymbol{A} = \frac{\beta}{n}\boldsymbol{x}_0\boldsymbol{x}_0^{\mathsf{T}} + \boldsymbol{W}\,.$$

- Signal $\boldsymbol{x}_0 \sim \mathsf{Unif}(\{+1,-1\}^n)$
- Noise $\boldsymbol{W} \sim \text{GOE}(n) \quad \perp\!\!\!\perp \quad \boldsymbol{x}_0$
- $\beta =$ signal-to-noise ratio
- Given $\boldsymbol{A}$, want to estimate $\boldsymbol{x}_0 \in \{+1,-1\}^n$

General group synchronization

$$A_{ij} = \frac{\beta}{n} x_{0,i}^{-1} x_{0,j} + \text{noise}\,.$$

- Signal $x_{0,i} \in \mathfrak{G}$
- Given A, want to estimate x_0
- $\mathbb{Z}_2$-synchronization: special case $\mathfrak{G} = (\{+1, -1\}, \cdot\,)$.

$\mathbb{Z}_2$-synchronization: Bayesian posterior

$$A = \frac{\beta}{n} x_0 x_0^{\mathsf{T}} + W\,, \qquad x_0 \sim \mathsf{Unif}(\{+1,-1\}^n)\,.$$

$$\mu_{A,\beta}(x) := \mathbb{P}(x|A)\,, \qquad x \in \{+1,-1\}^n$$

$$= \frac{1}{Z'(A)} \exp\Big\{ -\frac{n}{4}\Big\|A - \frac{\beta}{n} x x^{\mathsf{T}}\Big\|_2^2 \Big\}$$

$$= \frac{1}{Z(A)} \exp\Big\{ \frac{\beta}{2}\langle x, Ax\rangle \Big\}$$

Same structure as before; different matrix distribution

$\mathbb{Z}_2$-synchronization: Bayesian posterior

$$\boldsymbol{A} = \frac{\beta}{\boldsymbol{n}} \boldsymbol{x}_0 \boldsymbol{x}_0^{\mathsf{T}} + \boldsymbol{W}, \qquad \boldsymbol{x}_0 \sim \mathsf{Unif}(\{+1,-1\}^n).$$

$$\mu_{\boldsymbol{A},\beta}(\boldsymbol{x}) := \mathbb{P}(\boldsymbol{x}|\boldsymbol{A}), \qquad \boldsymbol{x} \in \{+1,-1\}^n$$

$$= \frac{1}{Z'(\boldsymbol{A})} \exp\Big\{ -\frac{\boldsymbol{n}}{4} \Big\| \boldsymbol{A} - \frac{\beta}{\boldsymbol{n}} \boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} \Big\|_2^2 \Big\}$$

$$= \frac{1}{Z(\boldsymbol{A})} \exp\Big\{ \frac{\beta}{2} \langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x} \rangle \Big\}$$

Same structure as before; different matrix distribution

$\mathbb{Z}_2$-synchronization: Bayesian posterior

$$A = \frac{\beta}{n} x_0 x_0^{\mathsf{T}} + W\,, \qquad x_0 \sim \mathsf{Unif}(\{+1,-1\}^n)\,.$$

$$\mu_{A,\beta}(x) := \mathbb{P}(x|A)\,, \qquad x \in \{+1,-1\}^n$$

$$= \frac{1}{Z'(A)} \exp\Big\{ -\frac{n}{4}\Big\|A - \frac{\beta}{n} x x^{\mathsf{T}}\Big\|_2^2 \Big\}$$

$$= \frac{1}{Z(A)} \exp\Big\{ \frac{\beta}{2}\langle x, Ax\rangle \Big\}$$

Same structure as before; different matrix distribution

Main result

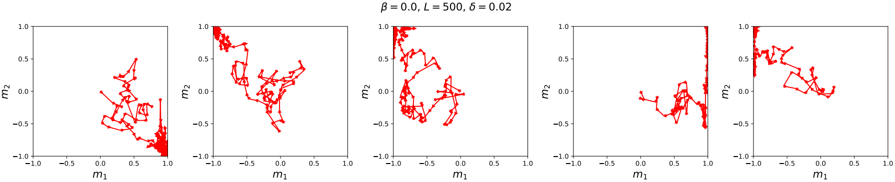
Theorem (M, Wu, 2022)

There exists $\beta_0 \in (0, \infty)$ s.t., for any $\beta \geq \beta_0$, there exists an SL sampling algorithm with complexity $O(n^2)$, that takes as input $\boldsymbol{A}$ (as above) and outputs $\boldsymbol{x} \sim \mu^{\rm alg}$ such that (whp)

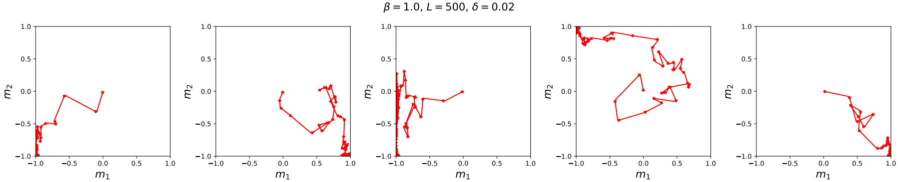
$$W_{2,n}(\mu^{\rm alg}, \mu_{\boldsymbol{A}}) = o_n(1)\,.$$

- Different regime: high SNR.
- It should work with optimal $\beta_0 = 1$, but…
- For $\beta < 1$ non-trivial estimation is impossible.

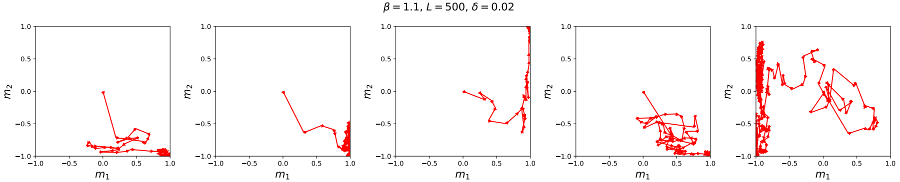
Trajectories: 2-dim projections; $\beta = 0$



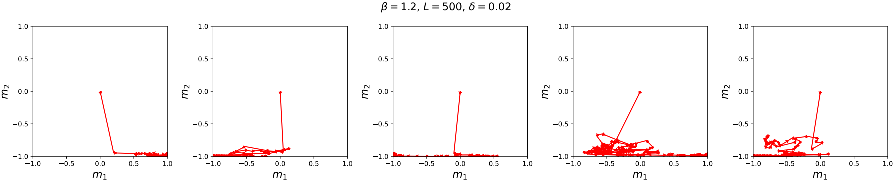
Trajectories: 2-dim projections; $\beta = 1$



Trajectories: 2-dim projections; $\beta = 1.1$



Trajectories: 2-dim projections; $\beta = 1.2$



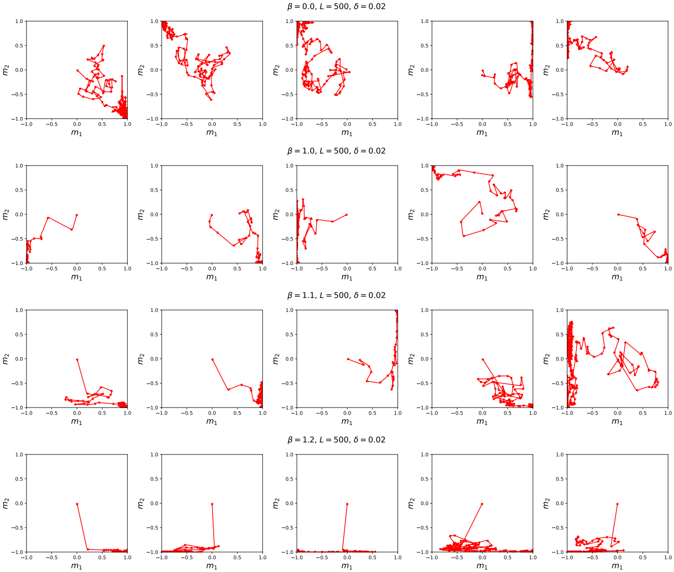
Trajectories: 2-dim projections

$\beta = 0$

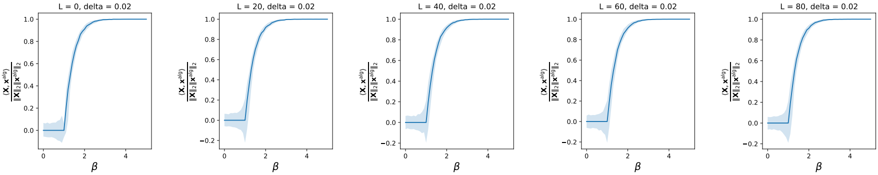
$\beta = 1$

$\beta = 1.1$

$\beta = 1.2$



Accuracy vs SNR: Increasing T

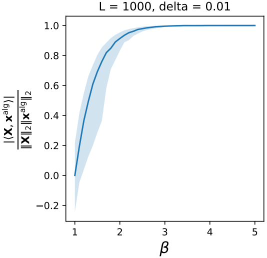


$T = 0$ $T = 0.4$ $T = 0.8$ $T = 1.2$ $T = 1.6$

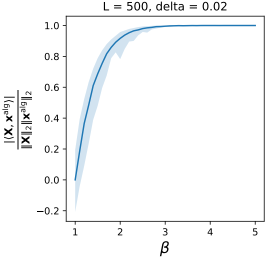
- $L := T/\delta$ number of SL steps.
- Continuous line: Bayes optimal accuracy[1], $n \to \infty$.
- 95% confidence bands over 300 runs.
- Martingale property + Concentration $\Rightarrow$ Independent of T

[1] Deshpande, Abbe, M, 2015

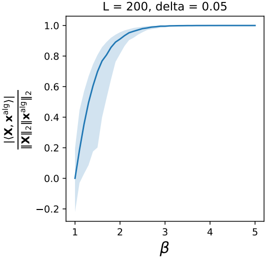
Accuracy vs SNR: Varying stepsize



$\delta = 0.01$



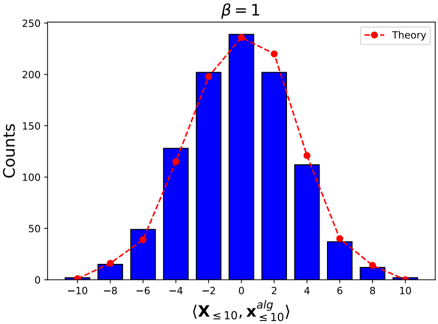
$\delta = 0.02$



$\delta = 0.05$

- $T = L\delta$ constant
- Continuous line: Bayes optimal accuracy, $n \to \infty$.

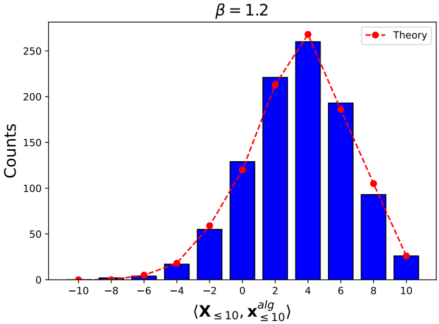
Low-dimensional projections, $\beta = 1$



For a fixed $\boldsymbol{W} \sim \mathrm{GOE}(n)$:

- Empirical distribution of $\sum_{i \le 10} x_{0,i} x_i^*$ (300 runs).
- Dashed line: Theoretical prediction (AMP).

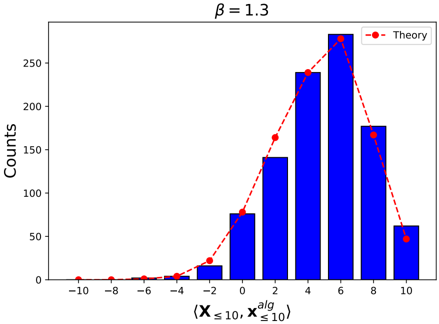
Low-dimensional projections, $\beta = 1.2$



For a fixed $\boldsymbol{W} \sim \text{GOE}(n)$:

- Empirical distribution of $\sum_{i \le 10} x_{0,i} x_i^*$ (300 runs).
- Dashed line: Theoretical prediction (AMP).

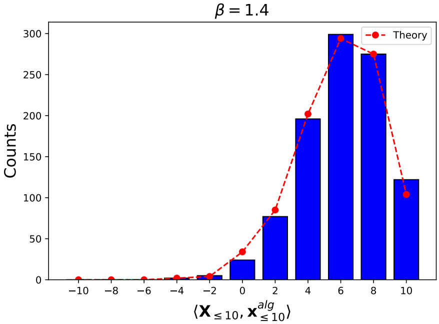
Low-dimensional projections, $\beta = 1.3$



For a fixed $\boldsymbol{W} \sim \text{GOE}(n)$:

- Empirical distribution of $\sum_{i\le 10} x_{0,i} x_i^*$ (300 runs).
- Dashed line: Theoretical prediction (AMP).

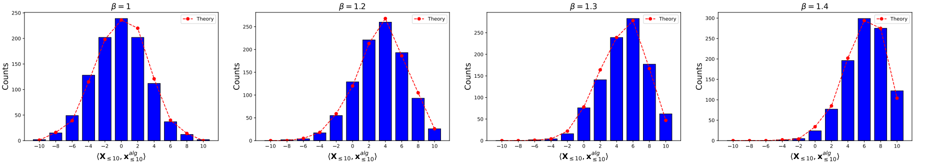
Low-dimensional projections, $\beta = 1.4$



For a fixed $\boldsymbol{W} \sim \mathrm{GOE}(n)$:

- Empirical distribution of $\sum_{i \leq 10} x_{0,i} x_i^*$ (300 runs).
- Dashed line: Theoretical prediction (AMP).

Low-dimensional projections



$\beta = 1$ $\beta = 1.2$ $\beta = 1.3$ $\beta = 1.4$

- $\langle \boldsymbol{x}_{0,\leq 10}, \boldsymbol{x}^*_{\leq 10} \rangle$ at high SNR

A different viewpoint and connections

Interpretation: Reduction of sampling to denoising

$$\boldsymbol{m}_t(\boldsymbol{y}) = \mathbb{E}[\boldsymbol{x}|\boldsymbol{y}]\,, \quad \frac{1}{t}\boldsymbol{y} = \boldsymbol{x} + \frac{1}{\sqrt{t}}\,\mathrm{g}\,, \;\; \mathrm{g} \sim \mathsf{N}(0, \mathrm{I}_d)\,,$$

$$\boldsymbol{m}_t(\,\cdot\,) = \arg \min_{\boldsymbol{\phi}:\mathbb{R}^d \to \mathbb{R}^s} \mathbb{E}\big\{\|\boldsymbol{\phi}(\boldsymbol{y}) - \boldsymbol{x}\|_2^2\big\}$$

If I have an optimal denoiser for Gaussian noise, I have a sampler

Interpretation: Reduction of sampling to denoising

$$\boldsymbol{m}_t(\boldsymbol{y}) = \mathbb{E}[\boldsymbol{x}|\boldsymbol{y}]\,, \quad \frac{1}{t}\boldsymbol{y} = \boldsymbol{x} + \frac{1}{\sqrt{t}}\,\boldsymbol{g}\,, \;\; \boldsymbol{g} \sim \mathsf{N}(0, \mathrm{I}_d)\,,$$

$$\boldsymbol{m}_t(\,\cdot\,) = \arg \min_{\boldsymbol{\phi}:\mathbb{R}^d \to \mathbb{R}^s} \mathbb{E}\big\{\|\boldsymbol{\phi}(\boldsymbol{y}) - \boldsymbol{x}\|_2^2\big\}$$

If I have an optimal denoiser for Gaussian noise, I have a sampler

Estimate $\boldsymbol{m}_t(\,\cdot\,)$ from data

$\boldsymbol{m}_t(\,\cdot\,)$

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}\{\|\boldsymbol{\phi}(\boldsymbol{y}) - \boldsymbol{x}\|_2^2\} \\ \text{subj. to} \quad & \boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R} \;\; \text{measurable.} \end{aligned}$$

Assume we have data $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \sim_{iid} \mu$

$\hat{\boldsymbol{m}}_t(\,\cdot\,)$: generate $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$

$$\begin{aligned} \text{minimize} \quad & \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{\phi}(\boldsymbol{y}_i) - \boldsymbol{x}_i\|_2^2 \\ \text{subj. to} \quad & \boldsymbol{\phi} \in \mathcal{F} \;\; \text{(function class)} \end{aligned}$$

Estimate $\boldsymbol{m}_t(\,\cdot\,)$ from data

$\boldsymbol{m}_t(\,\cdot\,)$

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}\{\|\boldsymbol{\phi}(\boldsymbol{y}) - \boldsymbol{x}\|_2^2\} \\ \text{subj. to} \quad & \boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R} \;\; \text{measurable.} \end{aligned}$$

Assume we have data $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \sim_{iid} \mu$

$\hat{\boldsymbol{m}}_t(\,\cdot\,)$: generate $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$

$$\begin{aligned} \text{minimize} \quad & \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{\phi}(\boldsymbol{y}_i) - \boldsymbol{x}_i\|_2^2 \\ \text{subj. to} \quad & \boldsymbol{\phi} \in \mathcal{F} \;\; \text{(function class)} \end{aligned}$$

For instance … μ = distr of natural images

$\boldsymbol{m}_t(\,\cdot\,)$

$$\begin{aligned}\text{minimize} \quad & \mathbb{E}\{\|\boldsymbol{\phi}(\boldsymbol{y}) - \boldsymbol{x}\|_2^2\} \\ \text{subj. to} \quad & \boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R} \;\; \text{measurable.}\end{aligned}$$

$\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ natural images

$\hat{\boldsymbol{m}}_t(\,\cdot\,)$

$$\begin{aligned}\text{minimize} \quad & \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{\phi}(\boldsymbol{y}_i) - \boldsymbol{x}_i\|_2^2 \\ \text{subj. to} \quad & \boldsymbol{\phi} \in \text{ConvNeuralNets}\end{aligned}$$

For instance … μ = distr of natural images

$\boldsymbol{m}_t(\,\cdot\,)$

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}\big\{\|\boldsymbol{\phi}(\boldsymbol{y}) - \boldsymbol{x}\|_2^2\big\} \\ \text{subj. to} \quad & \boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R} \;\; \text{measurable.} \end{aligned}$$

$\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ natural images

$\hat{\boldsymbol{m}}_t(\,\cdot\,)$

$$\begin{aligned} \text{minimize} \quad & \frac{1}{n}\sum_{i=1}^{n} \|\phi(\boldsymbol{y}_i) - x_i\|_2^2 \\ \text{subj. to} \quad & \phi \in \text{ConvNeuralNets} \end{aligned}$$

For instance … μ = distr of natural images

$\boldsymbol{m}_t(\,\cdot\,)$

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}\{\|\boldsymbol{\phi}(\boldsymbol{y}) - \boldsymbol{x}\|_2^2\} \\ \text{subj. to} \quad & \boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R} \ \text{ measurable.} \end{aligned}$$

$\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ natural images

$\hat{\boldsymbol{m}}_t(\,\cdot\,)$

$$\begin{aligned} \text{minimize} \quad & \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{\phi}(\boldsymbol{y}_i) - \boldsymbol{x}_i\|_2^2 \\ \text{subj. to} \quad & \boldsymbol{\phi} \in \text{ConvNeuralNets} \end{aligned}$$

Diffusions!

- State-of-the-art generative method in deep learning
- Song Ermon, 2019
- Song, Sohl-Dickstein, Kingma, Kumar, Ermon, Poole, 2020

Conclusion

Conclusion

- Stochastic Localization/Diffusions: A new approach to sampling
- Proved efficient sampling:
  - Sherrington-Kirkpatrick
  - $\mathbb{Z}_2$-synchronization
- Connection to diffusions in DL

Thanks!

Conclusion

- Stochastic Localization/Diffusions: A new approach to sampling
- Proved efficient sampling:
  - Sherrington-Kirkpatrick
  - $\mathbb{Z}_2$-synchronization
- Connection to diffusions in DL

Thanks!

Hardness

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

Algorithm:

$$\mathsf{ALG}_n : (\boldsymbol{W}, \beta, \omega) \mapsto \mathsf{ALG}_n(\boldsymbol{W}, \beta, \omega) \in [-1,1]^n$$

Stable algorithm:

$$\lim_{s\to 0} \operatorname*{p-lim}_{n\to\infty} W_{2,n}(\mu^{\rm alg}_{\boldsymbol{W},\beta}, \mu^{\rm alg}_{\boldsymbol{W}_s,\beta}) = 0\,, \qquad \boldsymbol{W}_s = \sqrt{1-s^2}\boldsymbol{W} + s\boldsymbol{W}'\,. \tag{1}$$

Theorem (El Alaoui, M, Sellke, 2022)

Let $\mu^{\rm alg}_{\boldsymbol{W},\beta}$ be the law of the output $\mathsf{ALG}_n(\boldsymbol{W}, \beta, \omega)$ conditional on $\boldsymbol{W}$. If ALG is a stable algorithm, then, for any $\beta > 1$:

$$\liminf_{n\to\infty} \mathbb{E}\big[W_{2,n}(\mu^{\rm alg}_{\boldsymbol{W},\beta},\ \mu_{\boldsymbol{W},\beta})\big] > 0\,.$$