Weak Poincaré Inequalities and Simulated Annealing

Brice Huang
(MIT)

Sidhanth Mohanty
(MIT)

Amit Rajaraman
(MIT)

David X. Wu
(Berkeley)

Available at arXiv:2411.09075

Motivation

Motivating Problem

Given a high-dimensional probability distribution μ, efficiently sample a point from μ.

Motivation

Motivating Problem

Given a high-dimensional probability distribution μ, efficiently sample a point from μ.

Useful in various downstream tasks:

- optimization
- inference
- integration…

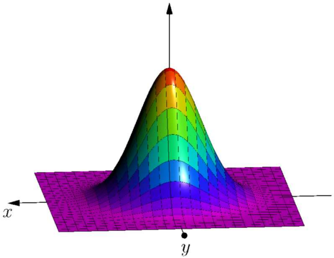
Example distribution: strongly log-concave distribution

Example distribution: strongly log-concave distribution

Distribution μ over $\mathbb{R}^n$, with density $\mu(x) \propto e^{-V(x)}$, with V strongly convex ($\nabla^2 V \succeq \alpha \mathbf{Id}$ uniformly).

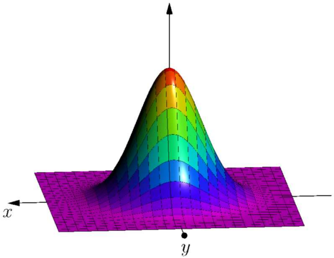
Example distribution: strongly log-concave distribution

Distribution μ over $\mathbb{R}^n$, with density $\mu(x) \propto e^{-V(x)}$, with V strongly convex ($\nabla^2 V \succeq \alpha \mathbf{Id}$ uniformly).



Example distribution: strongly log-concave distribution

Distribution μ over $\mathbb{R}^n$, with density $\mu(x) \propto e^{-V(x)}$, with V strongly convex ($\nabla^2 V \succeq \alpha \mathbf{Id}$ uniformly).



Classical result that such distributions can be efficiently sampled from. (see Chewi [Che23])

[Che23]: S Chewi. Log-concave sampling.

How do we sample?

How do we sample?

Markov Chain Monte Carlo!

How do we sample?

Markov Chain Monte Carlo!

Design a Markov chain P with stationary distribution μ.

How do we sample?

Markov Chain Monte Carlo!

Design a Markov chain P with stationary distribution μ.
Run it for $T = \mathbf{poly}(n)$ time starting at x_0. Output x_T.

How do we sample?

Markov Chain Monte Carlo!

Design a Markov chain P with stationary distribution μ.
Run it for $T = \mathrm{poly}(n)$ time starting at x_0. Output x_T.
Hopefully, the random output x_T is distributed according to μ.

How do we sample?

Markov Chain Monte Carlo!

Design a Markov chain P with stationary distribution μ.
Run it for $T = \mathbf{poly}(n)$ time starting at x_0. Output x_T.
Hopefully, the random output x_T is distributed according to μ.

P *mixes* (from x_0) in time T if $d_{\mathrm{TV}}\left(x_T, \mu\right)$ is tiny.

What Markov chain?

For distributions supported on $\mathbb{R}^n$, *Langevin diffusion* is a canonical Markov chain for $\mu(x) \propto e^{-V(x)}$

What Markov chain?

For distributions supported on $\mathbb{R}^n$, *Langevin diffusion* is a canonical Markov chain for $\mu(x) \propto e^{-V(x)}$:

$$X_{t+\delta} - X_t = -\delta \nabla V(X_t) + \sqrt{2\delta}\ \underset{\mathcal{N}(0,\mathrm{Id})}{\underset{\uparrow}{g_t}} \quad \text{for } \delta \text{ infinitesimal}$$

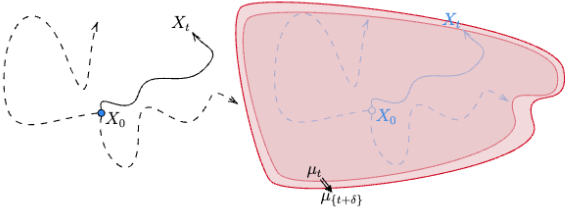


Figure from Yuansi Chen

What Markov chain?

For distributions supported on $\mathbb{R}^n$, *Langevin diffusion* is a canonical Markov chain for $\mu(x) \propto e^{-V(x)}$:

$$X_{t+\delta} - X_t = -\delta \nabla V(X_t) + \sqrt{2\delta} \underset{\substack{\uparrow \\ \mathcal{N}(0,\mathrm{Id})}}{g_t} \quad \text{for } \delta \text{ infinitesimal}$$

Definition can be modified for distributions supported on the sphere, say.

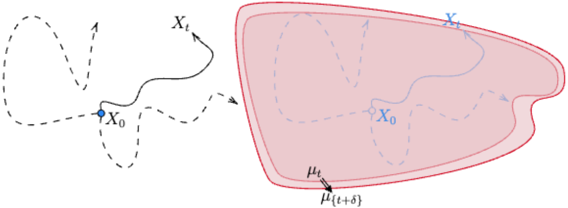


Figure from Yuansi Chen

Mixing times

Mixing is shown by proving that for *any* distribution ν,

$$\mathrm{distance}\,(P\nu \| \mu) \leqslant \left(1 - \frac{1}{\mathrm{poly}(n)}\right) \mathrm{distance}\,(\nu \| \mu) \ .$$

Mixing times

Mixing is shown by proving that for *any* distribution ν,

$$\operatorname{distance}(P\nu \| \mu) \leqslant \left(1 - \frac{1}{\operatorname{poly}(n)}\right) \operatorname{distance}(\nu \| \mu) \ .$$

There are now many many ways to show this:

- Coupling [BDJ96, BD97a]
- Path coupling [Jer95, BD97b, BD97c]
- Canonical paths [JS89, JSV04, HMMR05]
- Curvature considerations/Bakry–Émery theory [BÉ06, Vil09, EHMT17, CMS24]
- Zerofreeness [LY52, Bar16a, Bar16b, CLV24]
- Correlation decay [DSVW04, Wei04, Wei06, CLV21, CLMM23]
- Spectral independence [ALGV19, ALG21, Liu23, AJK^{+}24]
- Entropic independence [AJK^{+}21a, AJK^{+}21b, CCYZ24]
- Stochastic localization [EKZ22, CE22, AKV24, LMRW24]

Reevaluating the basics

Mixing is shown by proving that for *any* distribution ν,

$$\mathrm{distance}\left(P\nu \| \mu\right) \leqslant \left(1 - \frac{1}{\mathrm{poly}(n)}\right) \mathrm{distance}\left(\nu \| \mu\right)\ .$$

Reevaluating the basics

Mixing is shown by proving that for *any* distribution ν,

$$\mathrm{distance}\left(P\nu \| \mu\right) \leqslant \left(1 - \frac{1}{\mathrm{poly}(n)}\right) \mathrm{distance}\left(\nu \| \mu\right) \ .$$

...Cannot always guarantee this for all ν!

Reevaluating the basics

Mixing is shown by proving that for *any* distribution ν,

$$\mathrm{distance}\left(P\nu \| \mu\right) \leqslant \left(1 - \frac{1}{\mathrm{poly}(n)}\right) \mathrm{distance}\left(\nu \| \mu\right) \ .$$

…Cannot always guarantee this for all ν! Could I show mixing from some problem-specific initialization?

Reevaluating the basics

Mixing is shown by proving that for *any* distribution ν,

$$\mathrm{distance}\,(P\nu \| \mu) \leqslant \left(1 - \frac{1}{\mathrm{poly}(n)}\right) \mathrm{distance}\,(\nu \| \mu)\ .$$

…Cannot always guarantee this for all ν! Could I show mixing from some problem-specific initialization? Most previous work [LS93, Lov99, LV23, AEGP23] has been restricted to sampling from convex bodies.

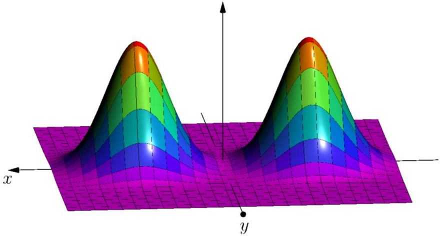
[LS93]: L Lovász and M Simonovits. Random walks in a convex body and an improved volume algorithm.
[Lov99]: L Lovász. Hit-and-run mixes fast.
[LV23]: A Laddha and SS Vempala. Convergence of Gibbs sampling: Coordinate hit-and-run mixes fast.
[AEGP23]: AE Alaoui, R Eldan, R Gheissari, and A Piana. Fast relaxation of the random field Ising dynamics.

Reevaluating the basics

Mixing is shown by proving that for *any* distribution ν,

$$\text{distance}\left(P\nu \| \mu\right) \leqslant \left(1 - \frac{1}{\text{poly}(n)}\right) \text{distance}\left(\nu \| \mu\right) \ .$$

…Cannot always guarantee this for all ν! Could I show mixing from some problem-specific initialization?
Most previous work [LS93, Lov99, LV23, AEGP23] has been restricted to sampling from convex bodies.



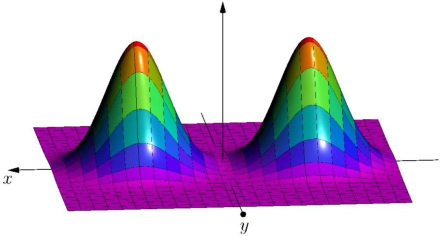
[LS93]: L Lovász and M Simonovits. Random walks in a convex body and an improved volume algorithm.
[Lov99]: L Lovász. Hit-and-run mixes fast.
[LV23]: A Laddha and SS Vempala. Convergence of Gibbs sampling: Coordinate hit-and-run mixes fast.
[AEGP23]: AE Alaoui, R Eldan, R Gheissari, and A Piana. Fast relaxation of the random field Ising dynamics.

Reevaluating the basics

Mixing is shown by proving that for *any* distribution ν,

$$\text{distance}\,(P\nu \| \mu) \leqslant \left(1 - \frac{1}{\text{poly}(n)}\right) \text{distance}\,(\nu \| \mu)\ .$$

…Cannot always guarantee this for all ν! Could I show mixing from some problem-specific initialization? Most previous work [LS93, Lov99, LV23, AEGP23] has been restricted to sampling from convex bodies.



Maybe if I initialize with equal mass in each cluster, I do mix.

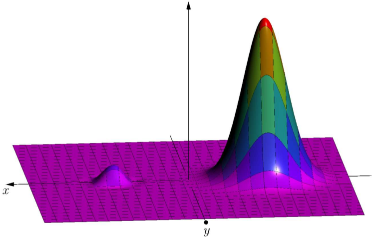
[LS93]: L Lovász and M Simonovits. Random walks in a convex body and an improved volume algorithm.
[Lov99]: L Lovász. Hit-and-run mixes fast.
[LV23]: A Laddha and SS Vempala. Convergence of Gibbs sampling: Coordinate hit-and-run mixes fast.
[AEGP23]: AE Alaoui, R Eldan, R Gheissari, and A Piana. Fast relaxation of the random field Ising dynamics.

Reevaluating the basics

Mixing is shown by proving that for *any* distribution ν,

$$\text{distance}\left(P\nu \| \mu\right) \leqslant \left(1 - \frac{1}{\text{poly}(n)}\right) \text{distance}\left(\nu \| \mu\right) \ .$$

…Cannot always guarantee this for all ν! Could I show mixing from some problem-specific initialization? Most previous work [LS93, Lov99, LV23, AEGP23] has been restricted to sampling from convex bodies.



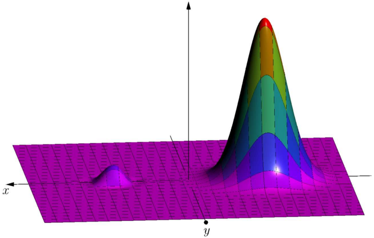
[LS93]: L Lovász and M Simonovits. Random walks in a convex body and an improved volume algorithm.
[Lov99]: L Lovász. Hit-and-run mixes fast.
[LV23]: A Laddha and SS Vempala. Convergence of Gibbs sampling: Coordinate hit-and-run mixes fast.
[AEGP23]: AE Alaoui, R Eldan, R Gheissari, and A Piana. Fast relaxation of the random field Ising dynamics.

Reevaluating the basics

Mixing is shown by proving that for *any* distribution ν,

$$\text{distance}\,(P\nu \| \mu) \leqslant \left(1 - \frac{1}{\text{poly}(n)}\right) \text{distance}\,(\nu \| \mu)\ .$$

…Cannot always guarantee this for all ν! Could I show mixing from some problem-specific initialization? Most previous work [LS93, Lov99, LV23, AEGP23] has been restricted to sampling from convex bodies.



Maybe if I initialize in the larger cluster, I do mix.

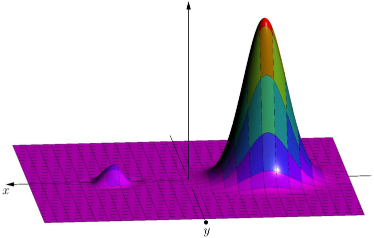
[LS93]: L Lovász and M Simonovits. Random walks in a convex body and an improved volume algorithm.
[Lov99]: L Lovász. Hit-and-run mixes fast.
[LV23]: A Laddha and SS Vempala. Convergence of Gibbs sampling: Coordinate hit-and-run mixes fast.
[AEGP23]: AE Alaoui, R Eldan, R Gheissari, and A Piana. Fast relaxation of the random field Ising dynamics.

Reevaluating the basics

Mixing is shown by proving that for *any* distribution ν,

$$\text{distance}\,(P\nu\|\mu) \leqslant \left(1 - \frac{1}{\text{poly}(n)}\right) \text{distance}\,(\nu\|\mu)\ .$$

…Cannot always guarantee this for all ν! Could I show mixing from some problem-specific initialization? Most previous work [LS93, Lov99, LV23, AEGP23] has been restricted to sampling from convex bodies.



Maybe if I initialize in the larger cluster, I do mix.

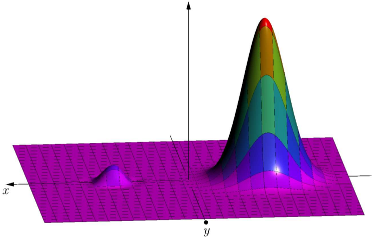
Is there a more principled approach to designing “good initializations”?

Reevaluating the basics

Mixing is shown by proving that for *any* distribution ν,

$$\mathrm{distance}\left(P\nu \| \mu\right) \leqslant \left(1 - \frac{1}{\mathrm{poly}(n)}\right) \mathrm{distance}\left(\nu \| \mu\right) .$$

…Cannot always guarantee this for all ν! Could I show mixing from some problem-specific initialization? Most previous work [LS93, Lov99, LV23, AEGP23] has been restricted to sampling from convex bodies.



Maybe if I initialize in the larger cluster, I do mix.

Is there a more principled approach to designing "good initializations"?
How do we prove rapid mixing from non-worst-case initializations?

Simulated annealing

Say I have distribution $\mu(x) \propto e^{-V(x)}$.

Simulated annealing

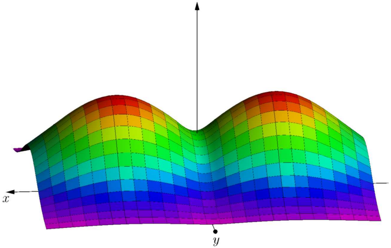
Say I have distribution $\mu(x) \propto e^{-V(x)}$. In simulated annealing, consider $\mu_\beta(x) \propto e^{-\beta V(x)}$.

Simulated annealing

Say I have distribution $\mu(x) \propto e^{-V(x)}$. In simulated annealing, consider $\mu_\beta(x) \propto e^{-\beta V(x)}$. Sampling is easy at small β, possibly difficult at $\beta = 1$.

Simulated annealing

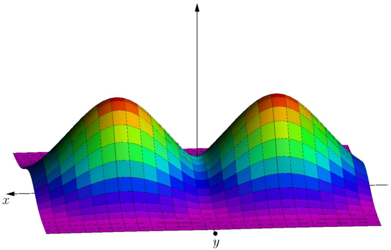
Say I have distribution $\mu(x) \propto e^{-V(x)}$. In simulated annealing, consider $\mu_\beta(x) \propto e^{-\beta V(x)}$. Sampling is easy at small β, possibly difficult at $\beta = 1$.



$\beta = 0.1$

Simulated annealing

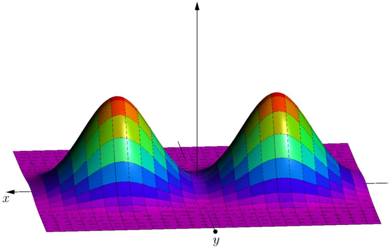
Say I have distribution $\mu(x) \propto e^{-V(x)}$. In simulated annealing, consider $\mu_\beta(x) \propto e^{-\beta V(x)}$. Sampling is easy at small β, possibly difficult at $\beta = 1$.



$\beta = 0.2$

Simulated annealing

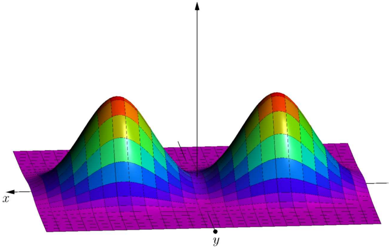
Say I have distribution $\mu(x) \propto e^{-V(x)}$. In simulated annealing, consider $\mu_\beta(x) \propto e^{-\beta V(x)}$. Sampling is easy at small β, possibly difficult at $\beta = 1$.



$\beta = 0.4$

Simulated annealing

Say I have distribution $\mu(x) \propto e^{-V(x)}$. In simulated annealing, consider $\mu_\beta(x) \propto e^{-\beta V(x)}$. Sampling is easy at small β, possibly difficult at $\beta = 1$.
Slowly increase β from 0 to 1, so μ_β provides a good initialization for the Markov chain at $\mu_{\beta+\delta}$.

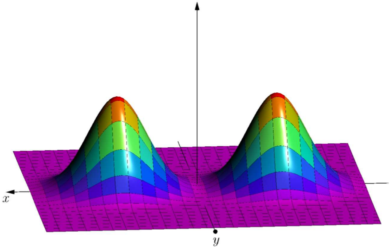


$\beta = 0.4$

Simulated annealing

Say I have distribution $\mu(x) \propto e^{-V(x)}$. In simulated annealing, consider $\mu_\beta(x) \propto e^{-\beta V(x)}$. Sampling is easy at small β, possibly difficult at $\beta = 1$.

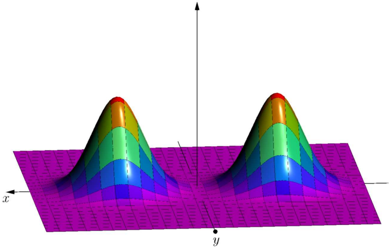
Slowly increase β from 0 to 1, so μ_β provides a good initialization for the Markov chain at $\mu_{\beta+\delta}$.



$\beta = 0.6$

Simulated annealing

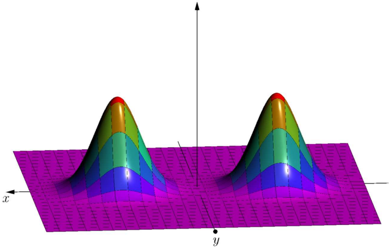
Say I have distribution $\mu(x) \propto e^{-V(x)}$. In simulated annealing, consider $\mu_\beta(x) \propto e^{-\beta V(x)}$. Sampling is easy at small β, possibly difficult at $\beta = 1$.
Slowly increase β from 0 to 1, so μ_β provides a good initialization for the Markov chain at $\mu_{\beta+\delta}$.



$\beta = 0.8$

Simulated annealing

Say I have distribution $\mu(x) \propto e^{-V(x)}$. In simulated annealing, consider $\mu_\beta(x) \propto e^{-\beta V(x)}$. Sampling is easy at small β, possibly difficult at $\beta = 1$.
Slowly increase β from 0 to 1, so μ_β provides a good initialization for the Markov chain at $\mu_{\beta+\delta}$.



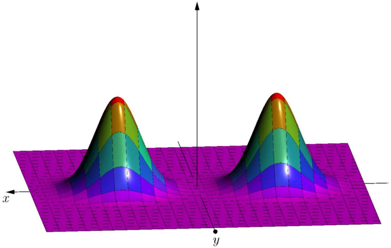
$\beta = 1.0$

Simulated annealing

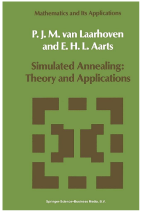
Say I have distribution $\mu(x) \propto e^{-V(x)}$. In simulated annealing, consider $\mu_\beta(x) \propto e^{-\beta V(x)}$. Sampling is easy at small β, possibly difficult at $\beta = 1$.

Slowly increase β from 0 to 1, so μ_β provides a good initialization for the Markov chain at $\mu_{\beta+\delta}$.

Used very often in practice!



$\beta = 1.0$

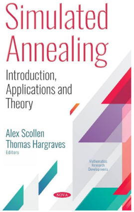


Mathematics and Its Applications

P. J. M. van Laarhoven
and E. H. L. Aarts

Simulated Annealing:
Theory and Applications

Springer-Science+Business Media, B.V.



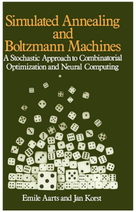
Simulated
Annealing

Introduction,
Applications and
Theory

Alex Scollen
Thomas Hargraves
Editors

Mathematics
Research
Developments

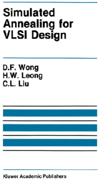
NOVA



Simulated Annealing
and
Boltzmann Machines

A Stochastic Approach to Combinatorial
Optimization and Neural Computing

Emile Aarts and Jan Korst



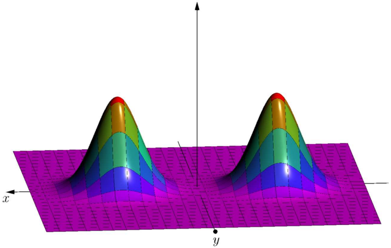
Simulated
Annealing for
VLSI Design

D.F. Wong
H.W. Leong
C.L. Liu

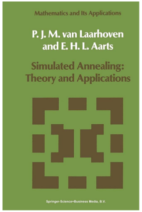
Kluwer Academic Publishers

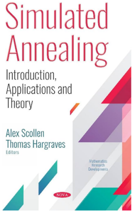
Simulated annealing

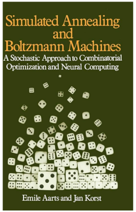
Say I have distribution $\mu(x) \propto e^{-V(x)}$. In simulated annealing, consider $\mu_\beta(x) \propto e^{-\beta V(x)}$. Sampling is easy at small β, possibly difficult at $\beta = 1$.
Slowly increase β from 0 to 1, so μ_β provides a good initialization for the Markov chain at $\mu_{\beta+\delta}$.
Used very often in practice!

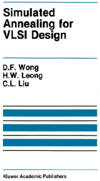


$\beta = 1.0$


Mathematics and Its Applications
P. J. M. van Laarhoven
and E. H. L. Aarts
Simulated Annealing:
Theory and Applications
Springer-Science+Business Media, B.V.


Simulated
Annealing
Introduction,
Applications and
Theory
Alex Scollen
Thomas Hargraves
Editors
Mathematics Research Developments
NOVA

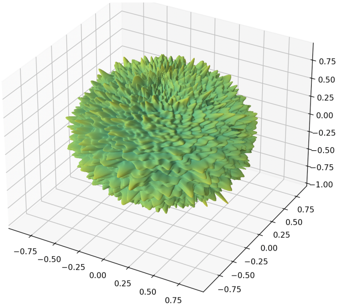

Simulated Annealing
and
Boltzmann Machines
A Stochastic Approach to Combinatorial
Optimization and Neural Computing
Emile Aarts and Jan Korst


Simulated
Annealing for
VLSI Design
D.F. Wong
H.W. Leong
C.L. Liu
Kluwer Academic Publishers

How do we prove rapid mixing from non-worst-case initializations?

1. Results
    - Sampling from spherical spin glasses
    - Sampling from data-based initializations
2. Techniques

Spherical 4-spin glass

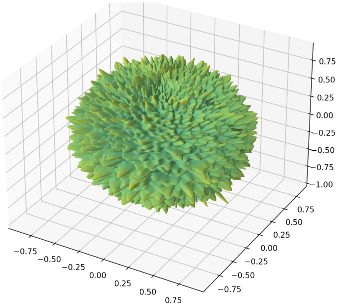


Spherical 4-spin glass

Set

$$H(\sigma) = \frac{\gamma}{N^{3/2}} \langle G, \sigma^{\otimes 4} \rangle$$

for $\sigma \in \sqrt{N} \cdot \mathbb{S}^{N-1}$ for G a random rank-4 tensor ($G_{i_1,\ldots,i_4} \sim \mathcal{N}(0,1)$), and set $\mu(\sigma) \propto e^{H(\sigma)}$.



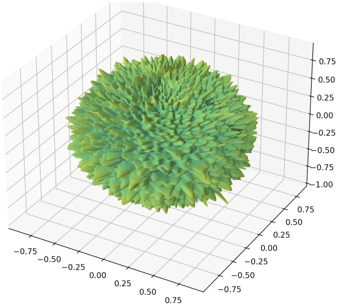
Spherical 4-spin glass

Set

$$H(\sigma) = \frac{\gamma}{N^{3/2}} \langle G, \sigma^{\otimes 4} \rangle$$

for $\sigma \in \sqrt{N} \cdot \mathbb{S}^{N-1}$ for G a random rank-4 tensor ($G_{i_1,\dots,i_4} \sim \mathcal{N}(0,1)$), and set $\mu(\sigma) \propto e^{H(\sigma)}$.

Extremely well-studied in statistical physics. Subject of 2021 Nobel Prize in Physics!



Results: Sampling from spherical 4-spin models

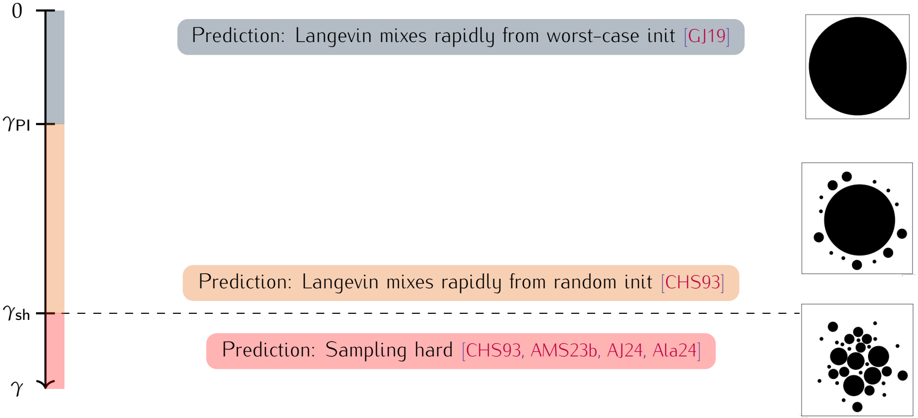
$H(\sigma) = \frac{1}{N^{3/2}} \langle G, \sigma^{\otimes 4} \rangle$

$\mu(\sigma) \propto e^{\gamma H(\sigma)}$ for $\sigma \in \sqrt{N} \cdot \mathbb{S}^{N-1}$

Results: Sampling from spherical 4-spin models

$$H(\sigma) = \frac{1}{N^{3/2}} \langle G, \sigma^{\otimes 4} \rangle$$

$$\mu(\sigma) \propto e^{\gamma H(\sigma)} \text{ for } \sigma \in \sqrt{N} \cdot \mathbb{S}^{N-1}$$



Prediction: Langevin mixes rapidly from worst-case init [GJ19]

Prediction: Langevin mixes rapidly from random init [CHS93]

Prediction: Sampling hard [CHS93, AMS23b, AJ24, Ala24]

[AJ24]: GB Arous and A Jagannath. Shattering versus metastability in spin glasses.

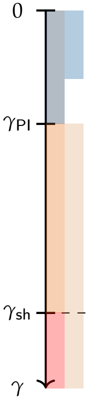
[Ala24]: AE Alaoui. Near-optimal shattering in the ising pure p-spin and rarity of solutions returned by stable algorithms.

[AMS23b]: AE Alaoui, A Montanari, and M Sellke. Shattering in pure spherical spin glasses.
[CHS93]: A Crisanti, H Horner, and HJ Sommers. The spherical p-spin interaction spin-glass model: the dynamics.
[GJ19]: R Gheissari and A Jagannath. On the spectral gap of spherical spin glass dynamics.

Results: Sampling from spherical 4–spin models

$$H(\sigma) = \frac{1}{N^{3/2}} \langle G, \sigma^{\otimes 4} \rangle$$

$$\mu(\sigma) \propto e^{\gamma H(\sigma)} \text{ for } \sigma \in \sqrt{N} \cdot \mathbb{S}^{N-1}$$



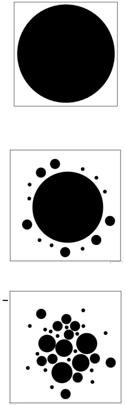
Prediction: Langevin mixes rapidly from worst-case init [GJ19]

Known: Langevin mixes rapidly from worst-case init [GJ19, AJK+24]

Known: Langevin mixes slowly from worst-case init [GJ19, AJ24]

Prediction: Langevin mixes rapidly from random init [CHS93]

Prediction: Sampling hard [CHS93, AMS23b, AJ24, Ala24]



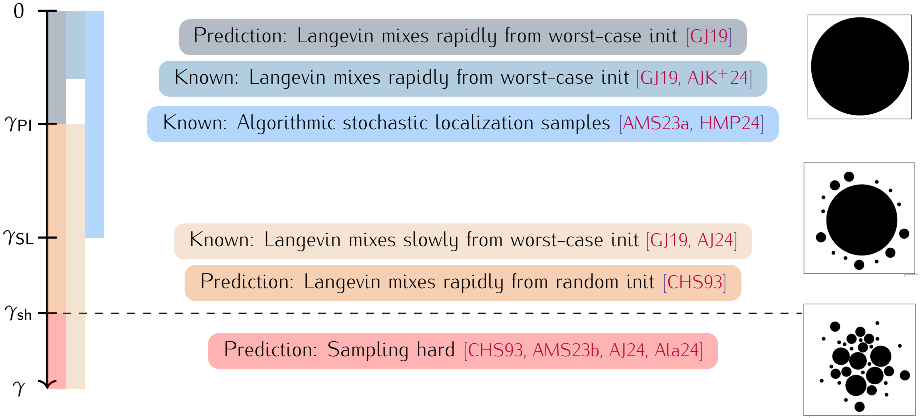
[AJ24]: GB Arous and A Jagannath. Shattering versus metastability in spin glasses.
[AJK+24]: N Anari, V Jain, F Koehler, HT Pham, and TD Vuong. Universality of spectral independence with applications to fast mixing in spin glasses.
[Ala24]: AE Alaoui. Near-optimal shattering in the ising pure p-spin and rarity of solutions returned by stable algorithms.

[AMS23b]: AE Alaoui, A Montanari, and M Sellke. Shattering in pure spherical spin glasses.
[CHS93]: A Crisanti, H Horner, and HJ Sommers. The spherical p-spin interaction spin-glass model: the dynamics.
[GJ19]: R Gheissari and A Jagannath. On the spectral gap of spherical spin glass dynamics.

Results: Sampling from spherical 4–spin models

$$H(\sigma) = \frac{1}{N^{3/2}}\langle G, \sigma^{\otimes 4}\rangle$$

$$\mu(\sigma) \propto e^{\gamma H(\sigma)} \text{ for } \sigma \in \sqrt{N} \cdot \mathbb{S}^{N-1}$$



Prediction: Langevin mixes rapidly from worst-case init [GJ19]

Known: Langevin mixes rapidly from worst-case init [GJ19, AJK+24]

Known: Algorithmic stochastic localization samples [AMS23a, HMP24]

Known: Langevin mixes slowly from worst-case init [GJ19, AJ24]

Prediction: Langevin mixes rapidly from random init [CHS93]

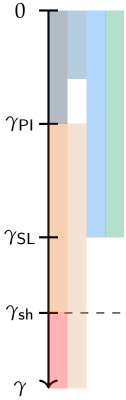
Prediction: Sampling hard [CHS93, AMS23b, AJ24, Ala24]

[AJ24]: GB Arous and A Jagannath. Shattering versus metastability in spin glasses.
[AJK+24]: N Anari, V Jain, F Koehler, HT Pham, and TD Vuong. Universality of spectral independence with applications to fast mixing in spin glasses.
[Ala24]: AE Alaoui. Near-optimal shattering in the ising pure p-spin and rarity of solutions returned by stable algorithms.
[AMS23a]: AE Alaoui, A Montanari, and M Sellke. Sampling from mean-field Gibbs measures via diffusion processes.
[AMS23b]: AE Alaoui, A Montanari, and M Sellke. Shattering in pure spherical spin glasses.
[CHS93]: A Crisanti, H Horner, and HJ Sommers. The spherical p-spin interaction spin-glass model: the dynamics.
[GJ19]: R Gheissari and A Jagannath. On the spectral gap of spherical spin glass dynamics.
[HMP24]: B Huang, A Montanari, and HT Pham. Sampling from Spherical Spin Glasses in Total Variation via Algorithmic Stochastic Localization.

Results: Sampling from spherical 4–spin models

$$H(\sigma) = \frac{1}{N^{3/2}}\langle G, \sigma^{\otimes 4}\rangle$$

$$\mu(\sigma) \propto e^{\gamma H(\sigma)} \text{ for } \sigma \in \sqrt{N}\cdot\mathbb{S}^{N-1}$$



Prediction: Langevin mixes rapidly from worst-case init [GJ19]

Known: Langevin mixes rapidly from worst-case init [GJ19, AJK+24]

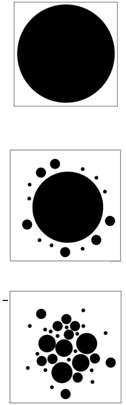
Known: Algorithmic stochastic localization samples [AMS23a, HMP24]

This work: simulated annealing samples

Known: Langevin mixes slowly from worst-case init [GJ19, AJ24]

Prediction: Langevin mixes rapidly from random init [CHS93]

Prediction: Sampling hard [CHS93, AMS23b, AJ24, Ala24]



[AJ24]: GB Arous and A Jagannath. Shattering versus metastability in spin glasses.
[AJK+24]: N Anari, V Jain, F Koehler, HT Pham, and TD Vuong. Universality of spectral independence with applications to fast mixing in spin glasses.
[Ala24]: AE Alaoui. Near-optimal shattering in the ising pure p-spin and rarity of solutions returned by stable algorithms.
[AMS23a]: AE Alaoui, A Montanari, and M Sellke. Sampling from mean-field Gibbs measures via diffusion processes.
[AMS23b]: AE Alaoui, A Montanari, and M Sellke. Shattering in pure spherical spin glasses.
[CHS93]: A Crisanti, H Horner, and HJ Sommers. The spherical p-spin interaction spin-glass model: the dynamics.
[GJ19]: R Gheissari and A Jagannath. On the spectral gap of spherical spin glass dynamics.
[HMP24]: B Huang, A Montanari, and HT Pham. Sampling from Spherical Spin Glasses in Total Variation via Algorithmic Stochastic Localization.

Results: Comparison to previous work

Theorem (HMRW)

Simulated annealing succeeds at sampling from μ to TV distance $e^{-cN^{0.2}}$ for $\gamma < \gamma_{\mathsf{SL}}$.

Results: Comparison to previous work

Theorem (HMRW)

Simulated annealing succeeds at sampling from μ to TV distance $e^{-cN^{0.2}}$ for $\gamma < \gamma_{\mathsf{SL}}$.

- Huang–Montanari–Pham [HMP24] show that algorithmic stochastic localization gets to $O(N^{-c})$ TV distance. This is intrinsic to the algorithm!

[HMP24]: B Huang, A Montanari, and HT Pham. Sampling from Spherical Spin Glasses in Total Variation via Algorithmic Stochastic Localization.

Results: Comparison to previous work

Theorem (HMRW)

Simulated annealing succeeds at sampling from μ to TV distance $e^{-cN^{0.2}}$ for $\gamma < \gamma_{\mathsf{SL}}$.

- Huang–Montanari–Pham [HMP24] show that algorithmic stochastic localization gets to $O(N^{-c})$ TV distance. This is intrinsic to the algorithm!
- Simulated annealing is what is run in practice.

[HMP24]: B Huang, A Montanari, and HT Pham. Sampling from Spherical Spin Glasses in Total Variation via Algorithmic Stochastic Localization.

Results: Comparison to previous work

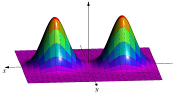
Theorem (HMRW)

Simulated annealing succeeds at sampling from μ to TV distance $e^{-cN^{0.2}}$ for $\gamma < \gamma_{\mathsf{SL}}$.

- Huang–Montanari–Pham [HMP24] show that algorithmic stochastic localization gets to $O(N^{-c})$ TV distance. This is intrinsic to the algorithm!
- Simulated annealing is what is run in practice.
- Algorithmic stochastic localization gets stuck at γ_{SL}, while we believe that simulated annealing should work all the way down to γ_{sh}.

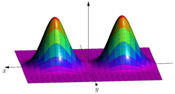
[HMP24]: B Huang, A Montanari, and HT Pham. Sampling from Spherical Spin Glasses in Total Variation via Algorithmic Stochastic Localization.

Results: Sampling from data-based initializations



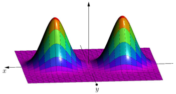
Let $\pi = \sum_{i=1}^{K} p_i \pi_i$ be a mixture of strongly log-concave distributions.

Results: Sampling from data-based initializations



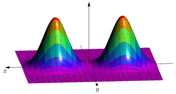
Let $\pi = \sum_{i=1}^{K} p_i \pi_i$ be a mixture of strongly log-concave distributions.
Given access to only the “score” $\nabla \log \pi$ (so you can run Langevin), cannot hope to sample without additional information (it is difficult to “find” the clusters). Worst-case mixing fails!

Results: Sampling from data-based initializations



Let $\pi = \sum_{i=1}^{K} p_i \pi_i$ be a mixture of strongly log-concave distributions.
Given access to only the “score” $\nabla \log \pi$ (so you can run Langevin), cannot hope to sample without additional information (it is difficult to “find” the clusters). Worst-case mixing fails!
Additional information: given a bunch of samples from π.

Results: Sampling from data-based initializations

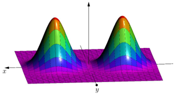


Let $\pi = \sum_{i=1}^{K} p_i \pi_i$ be a mixture of strongly log-concave distributions.
Given access to only the "score" $\nabla \log \pi$ (so you can run Langevin), cannot hope to sample without additional information (it is difficult to "find" the clusters). Worst-case mixing fails!
Additional information: given a bunch of samples from π.

Theorem (HMRW)

Suppose $\min p_i \geqslant p_*$. Let $x_1, x_2, \ldots, x_m$ be sampled according to π.

Results: Sampling from data-based initializations

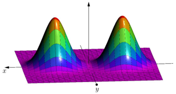


Let $\pi = \sum_{i=1}^{K} p_i \pi_i$ be a mixture of strongly log-concave distributions.
Given access to only the "score" $\nabla \log \pi$ (so you can run Langevin), cannot hope to sample without additional information (it is difficult to "find" the clusters). Worst-case mixing fails!
Additional information: given a bunch of samples from π.

Theorem (HMRW)

Suppose $\min p_i \geqslant p_*$. Let $x_1, x_2, \ldots, x_m$ be sampled according to π. For $m = \Omega\left(\frac{1}{p_* \varepsilon^2}\right)$, with high probability over the samples, Langevin diffusion initialized at $\frac{1}{m}\sum \delta_{x_i}$ run for $\mathrm{poly}(n)$ time samples from π to TV distance ε.

Results: Sampling from data-based initializations



Let $\pi = \sum_{i=1}^{K} p_i \pi_i$ be a mixture of strongly log-concave distributions.
Given access to only the "score" $\nabla \log \pi$ (so you can run Langevin), cannot hope to sample without additional information (it is difficult to "find" the clusters). Worst-case mixing fails!
Additional information: given a bunch of samples from π.

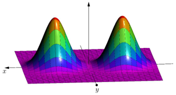
Theorem (HMRW)

Suppose $\min p_i \geqslant p_*$. Let $x_1, x_2, \ldots, x_m$ be sampled according to π. For $m = \Omega\left(\frac{1}{p_* \varepsilon^2}\right)$, with high probability over the samples, Langevin diffusion initialized at $\frac{1}{m}\sum \delta_{x_i}$ run for $\mathrm{poly}(n)$ time samples from π to TV distance ε.

Improves on (doubly exponential) dependence on $(1/p_*)$ from previous work by Koehler–Vuong [KV23].

[KV23]: F Koehler and TD Vuong. Sampling multimodal distributions with the vanilla score: Benefits of data-based initialization.

Results: Sampling from data-based initializations



Let $\pi = \sum_{i=1}^{K} p_i \pi_i$ be a mixture of strongly log-concave distributions.
Given access to only the "score" $\nabla \log \pi$ (so you can run Langevin), cannot hope to sample without additional information (it is difficult to "find" the clusters). Worst-case mixing fails!
Additional information: given a bunch of samples from π.

Theorem (HMRW)

Suppose $\min p_i \geqslant p_*$. Let $x_1, x_2, \ldots, x_m$ be sampled according to π. For $m = \Omega\left(\frac{1}{p_* \varepsilon^2}\right)$, with high probability over the samples, Langevin diffusion initialized at $\frac{1}{m} \sum \delta_{x_i}$ run for $\mathbf{poly}(n)$ time samples from π to TV distance ε.

Improves on (doubly exponential) dependence on $(1/p_*)$ from previous work by Koehler–Vuong [KV23].
This problem is studied in more detail in parallel independent work by Koehler–Lee–Vuong [KLV23].

[KV23]: F Koehler and TD Vuong. Sampling multimodal distributions with the vanilla score: Benefits of data-based initialization.
[KLV23]: F Koehler, H Lee, and TD Vuong. Efficiently learning and sampling multimodal distributions with data-based initialization.

Techniques

How do we show worst-case mixing?

Suppose I run Langevin diffusion initialized at ν_0, with distribution ν_t at time t.

How do we show worst-case mixing?

Suppose I run Langevin diffusion initialized at ν_0, with distribution ν_t at time t. Let $f_t = \frac{\mathrm{d}\nu_t}{\mathrm{d}\mu}$ be the likelihood ratio at time t.

How do we show worst-case mixing?

Suppose I run Langevin diffusion initialized at ν_0, with distribution ν_t at time t. Let $f_t = \frac{\mathrm{d}\nu_t}{\mathrm{d}\mu}$ be the likelihood ratio at time t. Then,

$$\chi^2\left(\nu_t \| \mu\right) = \mathbb{E}_\mu\left[\left(\frac{\mathrm{d}\nu_t}{\mathrm{d}\mu} - 1\right)^2\right]$$

How do we show worst-case mixing?

Suppose I run Langevin diffusion initialized at ν_0, with distribution ν_t at time t. Let $f_t = \frac{\mathrm{d}\nu_t}{\mathrm{d}\mu}$ be the likelihood ratio at time t. Then,

$$\chi^2\left(\nu_t \| \mu\right) = \mathbb{E}_\mu\left[\left(\frac{\mathrm{d}\nu_t}{\mathrm{d}\mu} - 1\right)^2\right] = \operatorname*{Var}_\mu[f_t]\,.$$

How do we show worst-case mixing?

Suppose I run Langevin diffusion initialized at ν_0, with distribution ν_t at time t. Let $f_t = \frac{\mathrm{d}\nu_t}{\mathrm{d}\mu}$ be the likelihood ratio at time t. Then,

$$\chi^2\left(\nu_t \| \mu\right) = \mathbb{E}_\mu\left[\left(\frac{\mathrm{d}\nu_t}{\mathrm{d}\mu} - 1\right)^2\right] = \operatorname*{Var}_\mu[f_t]\,.$$

Say I want to show that $\chi^2\left(\nu_t \| \mu\right)$ decays exponentially fast.

How do we show worst-case mixing?

Suppose I run Langevin diffusion initialized at ν_0, with distribution ν_t at time t. Let $f_t = \frac{d\nu_t}{d\mu}$ be the likelihood ratio at time t. Then,

$$\chi^2\left(\nu_t \| \mu\right) = \mathbb{E}_\mu\left[\left(\frac{d\nu_t}{d\mu} - 1\right)^2\right] = \operatorname*{Var}_\mu[f_t]\,.$$

Say I want to show that $\chi^2\left(\nu_t \| \mu\right)$ decays exponentially fast. Then, would like that

$$-\frac{d}{dt}\operatorname*{Var}_\mu[f_t] \geqslant \frac{1}{\mathrm{poly}(n)} \cdot \operatorname*{Var}_\mu[f_t]\,.$$

How do we show worst-case mixing?

Suppose I run Langevin diffusion initialized at ν_0, with distribution ν_t at time t. Let $f_t = \frac{\mathrm{d}\nu_t}{\mathrm{d}\mu}$ be the likelihood ratio at time t. Then,

$$\chi^2\left(\nu_t \| \mu\right) = \mathbb{E}_\mu\left[\left(\frac{\mathrm{d}\nu_t}{\mathrm{d}\mu} - 1\right)^2\right] = \operatorname*{Var}_\mu[f_t]\,.$$

Say I want to show that $\chi^2\left(\nu_t \| \mu\right)$ decays exponentially fast. Then, would like that

$$-\frac{\mathrm{d}}{\mathrm{d}t}\operatorname*{Var}_\mu[f_t] \geqslant \frac{1}{\mathrm{poly}(n)} \cdot \operatorname*{Var}_\mu[f_t]\,.$$

Lemma

Turns out that for Langevin, $-\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{Var}_\mu[f_t] = \mathbb{E}_\mu\|\nabla f_t\|_2^2$!

Poincaré inequalities

Poincaré inequality

For all functions f,

$$\mathbb{E}_\mu \|\nabla f\|_2^2 \geqslant \frac{1}{\mathrm{poly}(n)} \cdot \operatorname*{Var}_\mu[f] \ .$$

Poincaré inequalities imply rapid mixing from worst-case initializations!

Poincaré inequalities

Poincaré inequality

For all functions f,

$$\mathbb{E}_{\mu}\|\nabla f\|_2^2 \geqslant \frac{1}{\mathrm{poly}(n)} \cdot \underset{\mu}{\mathrm{Var}}[f] \, .$$

Poincaré inequalities imply rapid mixing from worst-case initializations!
Can be appropriately generalized to other Markov chains.

Poincaré inequalities

Poincaré inequality

For all functions f,

$$\mathbb{E}_\mu \|\nabla f\|_2^2 \geqslant \frac{1}{\mathrm{poly}(n)} \cdot \operatorname*{Var}_\mu[f] \, .$$

Poincaré inequalities imply rapid mixing from worst-case initializations!
Can be appropriately generalized to other Markov chains.
Poincaré inequalities are equivalent to the more familiar spectral gaps.

Poincaré inequalities

Poincaré inequality

For all functions f,

$$\mathbb{E}_{\mu}\|\nabla f\|_2^2 \geqslant \frac{1}{\mathrm{poly}(n)} \cdot \operatorname*{Var}_{\mu}[f] \, .$$

Poincaré inequalities imply rapid mixing from worst-case initializations!
Can be appropriately generalized to other Markov chains.
Poincaré inequalities are equivalent to the more familiar spectral gaps. All the techniques from earlier show rapid mixing by proving Poincaré inequalities/showing large spectral gaps.

An observation

This only cares about functions f_t encountered along the trajectory of the chain!

Weak Poincaré inequalities

This only cares about functions f_t encountered along the trajectory of the chain!

Weak Poincaré inequality (WPI)

μ satisfies a WPI with error functional $\mathbf{Defect}$ if for all f,

$$\mathbb{E}_\mu \|\nabla f\|^2 \geqslant \rho \left(\operatorname*{Var}_\mu[f] - \mathrm{Defect}(f) \right) .$$

Weak Poincaré inequalities

This only cares about functions f_t encountered along the trajectory of the chain!

Weak Poincaré inequality (WPI)

μ satisfies a WPI with error functional **Defect** if for all f,

$$\mathbb{E}_\mu \|\nabla f\|^2 \geqslant \rho \left(\operatorname*{Var}_{\mu}[f] - \mathrm{Defect}(f) \right) .$$

If μ satisfies a WPI,

$$\chi^2 \left(\nu_T \| \mu\right) \leqslant e^{-\rho T} \chi^2 \left(\nu_0 \| \mu\right) + e^{-\rho T} \int_0^T \rho e^{\rho t}\, \mathrm{Defect}(f_t)\, \mathrm{d}t .$$

Weak Poincaré inequalities

This only cares about functions f_t encountered along the trajectory of the chain!

Weak Poincaré inequality (WPI)

μ satisfies a WPI with error functional Defect if for all f,

$$\mathbb{E}_\mu \|\nabla f\|^2 \geqslant \rho \left(\operatorname*{Var}_\mu[f] - \mathrm{Defect}(f) \right) .$$

If μ satisfies a WPI,

$$\chi^2 \left(\nu_T \| \mu\right) \leqslant e^{-\rho T} \chi^2 \left(\nu_0 \| \mu\right) + e^{-\rho T} \int_0^T \rho e^{\rho t} \, \mathrm{Defect}(f_t) \, \mathrm{d}t \, .$$

If **Defect** is small (on average) along the trajectory of your Markov chain, it succeeds at sampling!

Weak Poincaré inequalities

This only cares about functions f_t encountered along the trajectory of the chain!

Weak Poincaré inequality (WPI)

μ satisfies a WPI with error functional **Defect** if for all f,

$$\mathbb{E}_\mu \|\nabla f\|^2 \geqslant \rho \left(\operatorname*{Var}_\mu[f] - \mathbf{Defect}(f) \right) .$$

If μ satisfies a WPI,

$$\chi^2\left(\nu_T \| \mu\right) \leqslant e^{-\rho T} \chi^2\left(\nu_0 \| \mu\right) + e^{-\rho T} \int_0^T \rho e^{\rho t}\, \mathbf{Defect}(f_t)\, \mathrm{d}t .$$

If **Defect** is small (on average) along the trajectory of your Markov chain, it succeeds at sampling!
So far, not new [Aid98], even for sampling guarantees [RW01].

[Aid98]: S Aida. Uniform positivity improving property, Sobolev inequalities, and spectral gaps.
[RW01]: M Röckner and FY Wang. Weak Poincaré inequalities and L^2-convergence rates of Markov semigroups

Weak Poincaré inequalities for simulated annealing

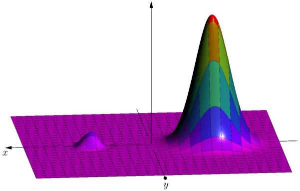
$$\mathbb{E}\|\nabla f\|_2^2 \geqslant \rho\left(\mathrm{Var}_\mu[f] - \mathrm{Defect}(f)\right)$$
$$\chi^2\left(\nu_T\|\mu\right) \leqslant e^{-\rho T}\chi^2\left(\nu_0\|\mu\right) + \mathbb{E}_t[\mathrm{Defect}(f_t)]$$

A weak Poincaré inequality with $\mathrm{Defect}(f) = \varepsilon \cdot \overset{\max f - \min f}{\overset{\downarrow}{\mathrm{osc}(f)}}{}^2$ implies simulated annealing samples!

Weak Poincaré inequalities for simulated annealing

$$\mathbb{E}\|\nabla f\|_2^2 \geqslant \rho\left(\mathrm{Var}_\mu[f] - \mathrm{Defect}(f)\right)$$
$$\chi^2(\nu_T\|\mu) \leqslant e^{-\rho T}\chi^2(\nu_0\|\mu) + \mathbb{E}_t[\mathrm{Defect}(f_t)]$$

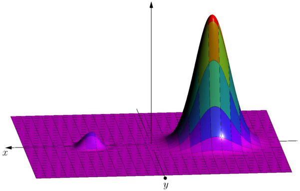
A weak Poincaré inequality with $\mathrm{Defect}(f) = \varepsilon \cdot \overset{\max f - \min f \atop \downarrow}{\mathrm{osc}(f)^2}$ implies simulated annealing samples!



Weak Poincaré inequalities for simulated annealing

$$\mathbb{E}\|\nabla f\|_2^2 \geqslant \rho\left(\mathrm{Var}_\mu[f] - \mathrm{Defect}(f)\right)$$
$$\chi^2(\nu_T\|\mu) \leqslant e^{-\rho T}\chi^2(\nu_0\|\mu) + \mathbb{E}_t[\mathrm{Defect}(f_t)]$$

A weak Poincaré inequality with $\mathbf{Defect}(f) = \varepsilon \cdot \overset{\displaystyle \max f - \min f \atop \downarrow}{\mathrm{osc}(f)^2}$ implies simulated annealing samples!

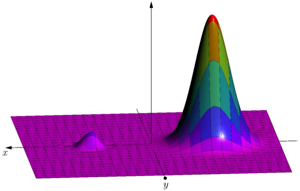


TV-close to true Poincaré $\implies$ $\mathbf{Defect} = \varepsilon \cdot \mathrm{osc}(f)^2$.

Weak Poincaré inequalities for simulated annealing

$$\mathbb{E}\|\nabla f\|_2^2 \geqslant \rho\left(\mathrm{Var}_\mu[f] - \mathrm{Defect}(f)\right)$$
$$\chi^2\left(\nu_T\|\mu\right) \leqslant e^{-\rho T}\chi^2\left(\nu_0\|\mu\right) + \mathbb{E}_t[\mathrm{Defect}(f_t)]$$

A weak Poincaré inequality with $\mathbf{Defect}(f) = \varepsilon \cdot \overset{\substack{\max f - \min f \\ \downarrow}}{\mathrm{osc}(f)^2}$ implies simulated annealing samples!

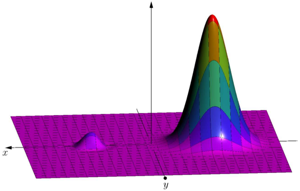


TV-close to true Poincaré $\Longrightarrow$ $\mathbf{Defect} = \varepsilon \cdot \mathrm{osc}(f)^2$.
Weak Poincaré inequalities can be proved using localization schemes! (generalizations of spectral independence, stochastic localization...)

Weak Poincaré inequalities for simulated annealing

$$\mathbb{E}\|\nabla f\|_2^2 \geqslant \rho \left(\mathrm{Var}_\mu[f] - \mathrm{Defect}(f)\right)$$
$$\chi^2\left(\nu_T \| \mu\right) \leqslant e^{-\rho T} \chi^2\left(\nu_0 \| \mu\right) + \mathbb{E}_t[\mathrm{Defect}(f_t)]$$

A weak Poincaré inequality with $\mathbf{Defect}(f) = \varepsilon \cdot \overset{\substack{\max f - \min f \\ \downarrow}}{\mathrm{osc}(f)^2}$ implies simulated annealing samples!



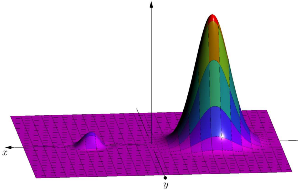
TV-close to true Poincaré $\implies$ $\mathbf{Defect} = \varepsilon \cdot \mathrm{osc}(f)^2$.
Weak Poincaré inequalities can be proved using localization schemes! (generalizations of spectral independence, stochastic localization...)

A quick jargon-infested elaboration for the expert.

Weak Poincaré inequalities for simulated annealing

$$\mathbb{E}\|\nabla f\|_2^2 \geqslant \rho\left(\mathrm{Var}_\mu[f] - \mathrm{Defect}(f)\right)$$
$$\chi^2(\nu_T\|\mu) \leqslant e^{-\rho T}\chi^2(\nu_0\|\mu) + \mathbb{E}_t[\mathrm{Defect}(f_t)]$$

A weak Poincaré inequality with $\mathbf{Defect}(f) = \varepsilon \cdot \overset{\max f - \min f}{\overset{\downarrow}{\mathrm{osc}(f)}}{}^2$ implies simulated annealing samples!



TV-close to true Poincaré $\Longrightarrow$ $\mathbf{Defect} = \varepsilon \cdot \mathrm{osc}(f)^2$.
Weak Poincaré inequalities can be proved using localization schemes! (generalizations of spectral independence, stochastic localization…)

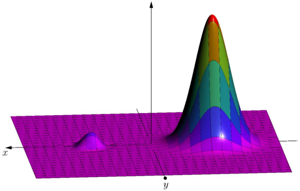
A quick jargon-infested elaboration for the expert.

When using a localization scheme to prove fast mixing,

Weak Poincaré inequalities for simulated annealing

$$\mathbb{E}\|\nabla f\|_2^2 \geqslant \rho\left(\mathrm{Var}_\mu[f] - \mathrm{Defect}(f)\right)$$
$$\chi^2(\nu_T\|\mu) \leqslant e^{-\rho T}\chi^2(\nu_0\|\mu) + \mathbb{E}_t[\mathrm{Defect}(f_t)]$$

A weak Poincaré inequality with $\mathbf{Defect}(f) = \varepsilon \cdot \overset{\max f - \min f}{\overset{\downarrow}{\mathrm{osc}(f)^2}}$ implies simulated annealing samples!



TV-close to true Poincaré $\implies$ $\mathbf{Defect} = \varepsilon \cdot \mathrm{osc}(f)^2$.
Weak Poincaré inequalities can be proved using localization schemes! (generalizations of spectral independence, stochastic localization...)

A quick jargon-infested elaboration for the expert.

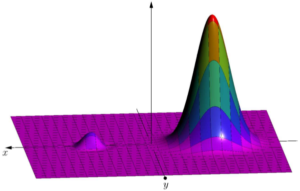
When using a localization scheme to prove fast mixing,

- Bounded influence for **all** pinnings/control on **all** localization paths $\implies$ Poincaré

Weak Poincaré inequalities for simulated annealing

$$\mathbb{E}\|\nabla f\|_2^2 \geqslant \rho \left(\mathrm{Var}_\mu[f] - \mathrm{Defect}(f)\right)$$
$$\chi^2\left(\nu_T \| \mu\right) \leqslant e^{-\rho T} \chi^2\left(\nu_0 \| \mu\right) + \mathbb{E}_t[\mathrm{Defect}(f_t)]$$

A weak Poincaré inequality with $\mathbf{Defect}(f) = \varepsilon \cdot \overset{\max f - \min f}{\overset{\downarrow}{\mathrm{osc}(f)^2}}$ implies simulated annealing samples!



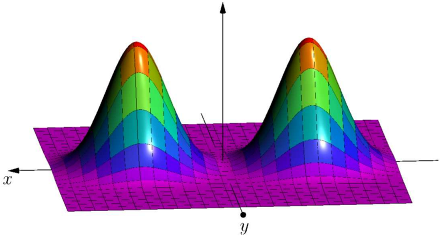
TV-close to true Poincaré $\implies$ $\mathbf{Defect} = \varepsilon \cdot \mathrm{osc}(f)^2$.
Weak Poincaré inequalities can be proved using localization schemes! (generalizations of spectral independence, stochastic localization...)

A quick jargon-infested elaboration for the expert.

When using a localization scheme to prove fast mixing,

- Bounded influence for **all** pinnings/control on **all** localization paths $\implies$ Poincaré
- Bounded influence for **most** pinnings/control on **most** localization paths $\implies$ weak Poincaré

Weak Poincaré inequalities from symmetry

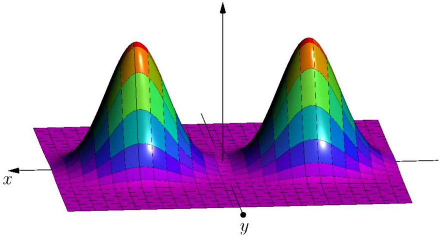


$$\mathbb{E}\|\nabla f\|_2^2 \geqslant \rho \left(\mathrm{Var}_\mu[f] - \mathrm{Defect}(f)\right)$$

$$\chi^2\left(\nu_T \| \mu\right) \leqslant e^{-\rho T} \chi^2\left(\nu_0 \| \mu\right) + \mathbb{E}_t[\mathrm{Defect}(f_t)]$$

Weak Poincaré inequalities from symmetry

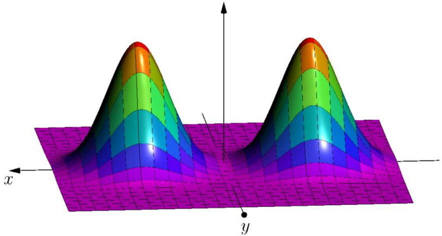
$$\mathbb{E}\|\nabla f\|_2^2 \geqslant \rho \left(\mathrm{Var}_\mu[f] - \mathrm{Defect}(f) \right)$$
$$\chi^2 (\nu_T \| \mu) \leqslant e^{-\rho T} \chi^2 (\nu_0 \| \mu) + \mathbb{E}_t [\mathrm{Defect}(f_t)]$$



Symmetric function $\implies$ **Defect** $= 0$.

Weak Poincaré inequalities from symmetry

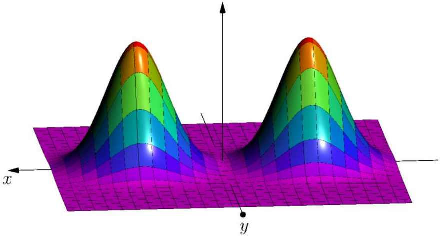
$$\mathbb{E}\|\nabla f\|_2^2 \geqslant \rho\left(\mathrm{Var}_\mu[f] - \mathrm{Defect}(f)\right)$$
$$\chi^2(\nu_T\|\mu) \leqslant e^{-\rho T}\chi^2(\nu_0\|\mu) + \mathbb{E}_t[\mathrm{Defect}(f_t)]$$



Symmetric function $\implies$ **Defect** $= 0$.
Approximately symmetric $\implies$ **Defect** small.

Proof: Sampling from data-based initializations

$$\mathbb{E}\|\nabla f\|_2^2 \geqslant \rho\left(\mathrm{Var}_\mu[f] - \mathrm{Defect}(f)\right)$$
$$\chi^2\left(\nu_T\|\mu\right) \leqslant e^{-\rho T}\chi^2\left(\nu_0\|\mu\right) + \mathbb{E}_t[\mathrm{Defect}(f_t)]$$



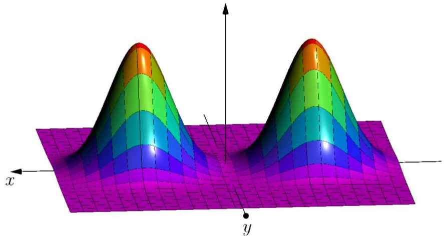
Symmetric function $\implies$ **Defect** $= 0$.
Approximately symmetric $\implies$ **Defect** small.

Let $\pi = \sum_{i=1}^K p_i\pi_i$ be a mixture of strongly log-concave distributions.

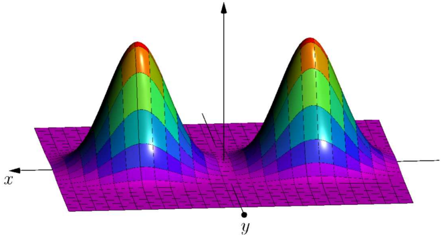
Theorem (HMRW)

Suppose $\min p_i \geqslant p_*$. Let $x_1, x_2, \ldots, x_m$ be sampled according to π. For $m = \Omega\left(\frac{1}{p_*\varepsilon^2}\right)$, with high probability over the samples, Langevin diffusion initialized at $\frac{1}{m}\sum \delta_{x_i}$ run for $\mathbf{poly}(n)$ time samples from π to TV distance ε.

Intuition

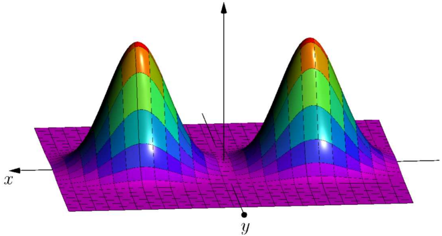


Intuition



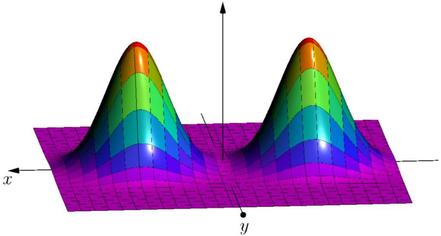
If the clusters were far apart, you expect to get about the right fraction of points per cluster at the start.

Intuition



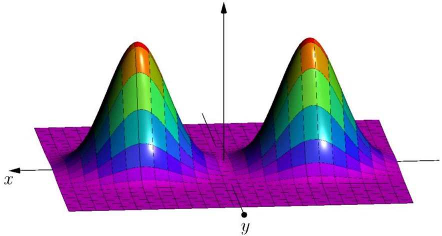
If the clusters were far apart, you expect to get about the right fraction of points per cluster at the start.
Approximately symmetric initialization, **Defect** starts small.

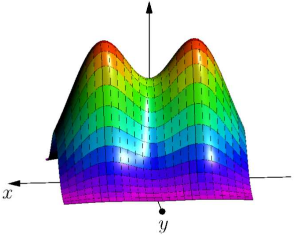
Intuition



If the clusters were far apart, you expect to get about the right fraction of points per cluster at the start.
Approximately symmetric initialization, **Defect** starts small.
Mass does not travel between clusters, **Defect** stays small.

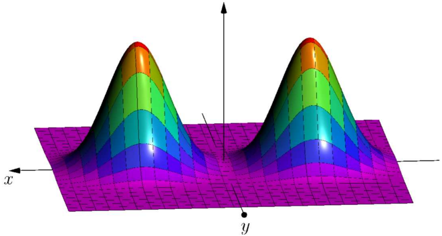
Intuition



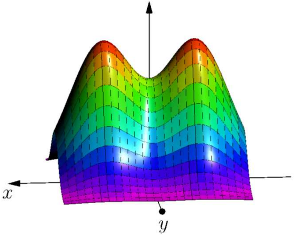


If the clusters were far apart, you expect to get about the right fraction of points per cluster at the start.
Approximately symmetric initialization, **Defect** starts small.
Mass does not travel between clusters, **Defect** stays small.

Intuition

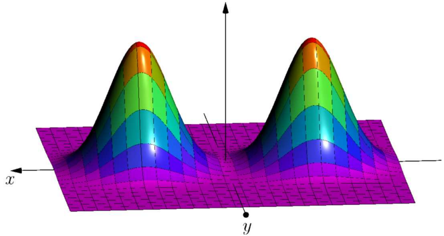


If the clusters were far apart, you expect to get about the right fraction of points per cluster at the start.
Approximately symmetric initialization, **Defect** starts small.
Mass does not travel between clusters, **Defect** stays small.

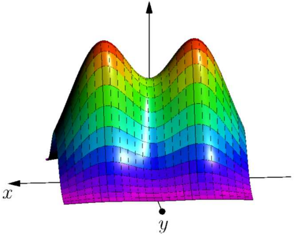


Defect starts off small for the same reason.

Intuition

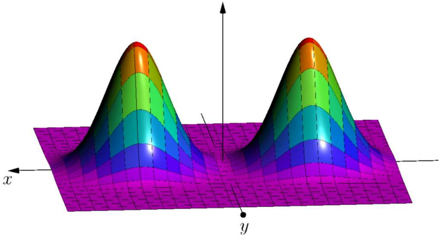


If the clusters were far apart, you expect to get about the right fraction of points per cluster at the start.
Approximately symmetric initialization, **Defect** starts small.
Mass does not travel between clusters, **Defect** stays small.

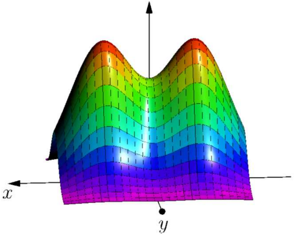


Defect starts off small for the same reason.
Mass can travel between clusters, but it should do so in a symmetric fashion.

Intuition

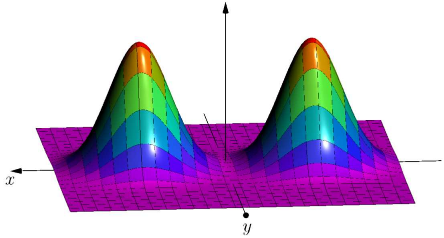


If the clusters were far apart, you expect to get about the right fraction of points per cluster at the start.
Approximately symmetric initialization, **Defect** starts small.
Mass does not travel between clusters, **Defect** stays small.

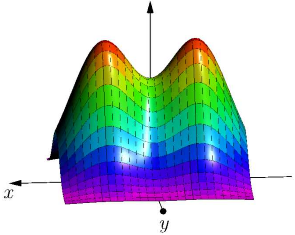


Defect starts off small for the same reason.
Mass can travel between clusters, but it should do so in a symmetric fashion.
Defect should stay small?

Intuition



If the clusters were far apart, you expect to get about the right fraction of points per cluster at the start.
Approximately symmetric initialization, **Defect** starts small.
Mass does not travel between clusters, **Defect** stays small.



Defect starts off small for the same reason.
Mass can travel between clusters, but it should do so in a symmetric fashion.
Defect should stay small? (controlling this is essentially the source of the doubly exponential dependence in previous work)

Proving a weak Poincaré inequality

$$\mathbb{E}\|\nabla f\|_2^2 \geqslant \rho\left(\mathrm{Var}_\mu[f] - \mathrm{Defect}(f)\right)$$
$$\chi^2\left(\nu_T \| \mu\right) \leqslant e^{-\rho T}\chi^2\left(\nu_0 \| \mu\right) + \mathbb{E}_t[\mathrm{Defect}(f_t)]$$

Proving a weak Poincaré inequality

$$\mathbb{E}\|\nabla f\|_2^2 \geqslant \rho\left(\mathrm{Var}_\mu[f] - \mathrm{Defect}(f)\right)$$
$$\chi^2\left(\nu_T\|\mu\right) \leqslant e^{-\rho T}\chi^2\left(\nu_0\|\mu\right) + \mathbb{E}_t[\mathrm{Defect}(f_t)]$$

(proof on board) Will show

$$\mathbb{E}\|\nabla f\|_2^2 \gtrsim \mathrm{Var}[f] - \underbrace{\sum_{i=1}^{K} p_i\left(\mathbb{E}_{\pi_i}[f]^2 - \mathbb{E}_\pi[f]^2\right)}_{\mathrm{Defect}(f)} .$$

Proving a weak Poincaré inequality

$\mathbb{E}\|\nabla f\|_2^2 \geqslant \rho \left(\mathrm{Var}_\mu[f] - \mathrm{Defect}(f)\right)$

$\chi^2\left(\nu_T \| \mu\right) \leqslant e^{-\rho T} \chi^2\left(\nu_0 \| \mu\right) + \mathbb{E}_t[\mathrm{Defect}(f_t)]$

(proof on board) Will show

$$\mathbb{E}\|\nabla f\|_2^2 \gtrsim \mathrm{Var}[f] - \underbrace{\sum_{i=1}^{K} p_i \left(\mathbb{E}_{\pi_i}[f]^2 - \mathbb{E}_\pi[f]^2\right)}_{\mathrm{Defect}(f)} .$$

This is a random variable depending on the samples $x_1, \ldots, x_m$. Would like to show that it is small with high probability (over the samples) along the path of the Markov chain.

Controlling the error

$$\mathrm{Defect}(f_t) = \sum_{i=1}^{K} p_i \left(\mathbb{E}_{\pi_i}[f_t]^2 - 1 \right) \; .$$

Controlling the error

$$\mathrm{Defect}(f_t) = \sum_{i=1}^{K} p_i \left(\mathbb{E}_{\pi_i}[f_t]^2 - 1 \right) \ .$$

(proof on board)

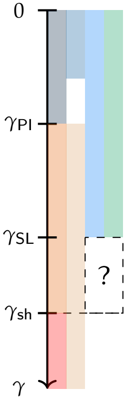
Controlling the error

$$\mathbf{Defect}(f_t) = \sum_{i=1}^{K} p_i \left(\mathbb{E}_{\pi_i}[f_t]^2 - 1 \right) \, .$$

(proof on board)

So $\mathbf{Defect}(f_t)$ is small with high probability! We are done!

Open Questions I: Sampling down to the shattering threshold



Prediction: Langevin mixes rapidly from worst-case init [GJ19]

Langevin mixes rapidly from worst-case init [GJ19, AJK$^+$24]

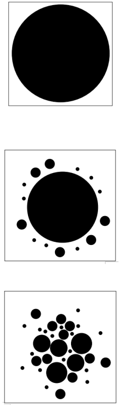
Algorithmic stochastic localization samples [AMS23a, HMP24]

This work: simulated annealing samples

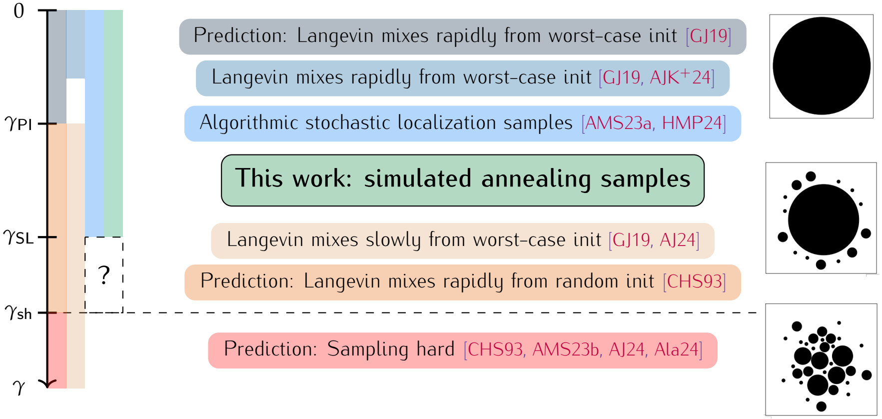
Langevin mixes slowly from worst-case init [GJ19, AJ24]

Prediction: Langevin mixes rapidly from random init [CHS93]

Prediction: Sampling hard [CHS93, AMS23b, AJ24, Ala24]



Open Questions I: Sampling down to the shattering threshold



Prediction: Langevin mixes rapidly from worst-case init [GJ19]

Langevin mixes rapidly from worst-case init [GJ19, AJK$^+$24]

Algorithmic stochastic localization samples [AMS23a, HMP24]

This work: simulated annealing samples

Langevin mixes slowly from worst-case init [GJ19, AJ24]

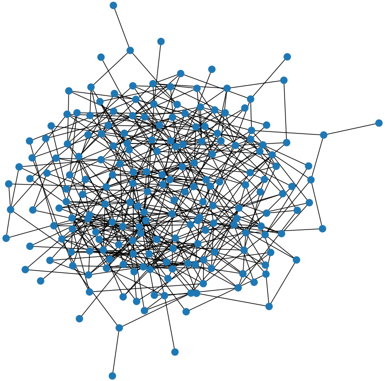
Prediction: Langevin mixes rapidly from random init [CHS93]

Prediction: Sampling hard [CHS93, AMS23b, AJ24, Ala24]

How do we close this gap? It seems like our proof strategy gets stuck...

Open Questions II: Annealing for inference

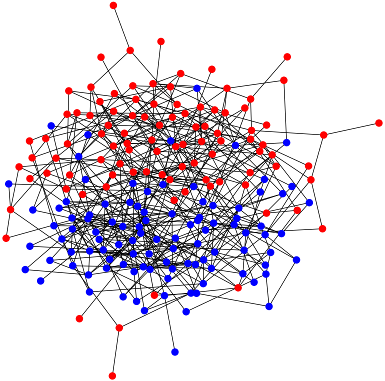
Inference problem: Infer $\mathbf{x} \sim \{\bullet, \bullet\}^n$ after observing a sparse random graph with "community structure" $\mathbf{x}$.



Open Questions II: Annealing for inference

Inference problem: Infer $\mathbf{x} \sim \{\bullet, \bullet\}^n$ after observing a sparse random graph with “community structure” $\mathbf{x}$.

Connect vertices with probability $\frac{d+\lambda\sqrt{d}}{n}$ if ●···● or ●···●

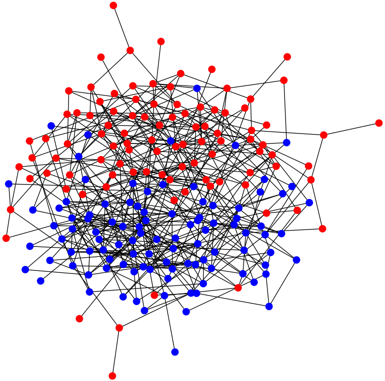


Open Questions II: Annealing for inference

Inference problem: Infer $\mathbf{x} \sim \{$🔴, 🔵$\}^n$ after observing a sparse random graph with “community structure” $\mathbf{x}$.

Connect vertices with probability $\frac{d+\lambda\sqrt{d}}{n}$ if 🔴···🔴 or 🔵···🔵

Connect vertices with probability $\frac{d-\lambda\sqrt{d}}{n}$ if 🔴···🔵 or 🔵···🔴

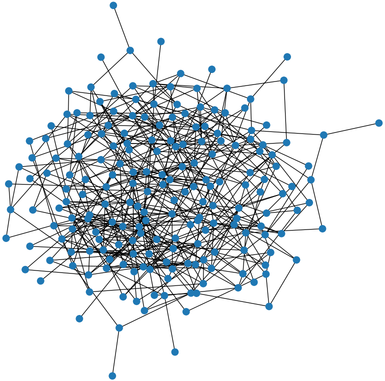


Open Questions II: Annealing for inference

Inference problem: Infer $\mathbf{x} \sim$ {🔴, 🔵}^{n} after observing a sparse random graph with “community structure” $\mathbf{x}$.

Connect vertices with probability $\frac{d+\lambda\sqrt{d}}{n}$ if 🔴···🔴 or 🔵···🔵

Connect vertices with probability $\frac{d-\lambda\sqrt{d}}{n}$ if 🔴···🔵 or 🔵···🔴



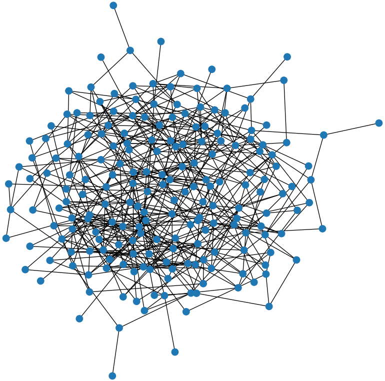
Open Questions II: Annealing for inference

Inference problem: Infer $\mathbf{x} \sim \{\bullet, \bullet\}^n$ after observing a sparse random graph with “community structure” $\mathbf{x}$.

Connect vertices with probability $\frac{d+\lambda\sqrt{d}}{n}$ if ●···● or ●···●

Connect vertices with probability $\frac{d-\lambda\sqrt{d}}{n}$ if ●···● or ●···●

Annealing run on the posterior of the stochastic block model appears to perform optimally... why?



Open Questions III: Worst-case approximation algorithms from non-worst-case initializations

For max-cut in a graph, SDPs provide a 0.878-approximation to the max-cut.

Open Questions III: Worst-case approximation algorithms from non-worst-case initializations

For max-cut in a graph, SDPs provide a 0.878-approximation to the max-cut.

Conjecture (Folklore, inspired by Trevisan [Tre01])

On bounded degree graphs, should be able to approximate to $0.878 + \Omega\left(\frac{1}{\sqrt{d}}\right)$!

[Tre01]: L Trevisan. Non-approximability results for optimization problems on bounded degree instances.

Open Questions III: Worst-case approximation algorithms from non-worst-case initializations

For max-cut in a graph, SDPs provide a 0.878-approximation to the max-cut.

Conjecture (Folklore, inspired by Trevisan [Tre01])

On bounded degree graphs, should be able to approximate to $0.878 + \Omega\left(\frac{1}{\sqrt{d}}\right)$!

Hsieh–Kothari [HK22] do local updates to the SDP solution to get $0.878 + \Omega\left(\frac{1}{d^2}\right)$.

[HK22]: JT Hsieh and P Kothari. Approximating Max-Cut on Bounded Degree Graphs: Tighter Analysis of the FKL Algorithm.

[Tre01]: L Trevisan. Non-approximability results for optimization problems on bounded degree instances.

Open Questions III: Worst-case approximation algorithms from non-worst-case initializations

For max-cut in a graph, SDPs provide a 0.878-approximation to the max-cut.

Conjecture (Folklore, inspired by Trevisan [Tre01])

On bounded degree graphs, should be able to approximate to $0.878 + \Omega\left(\frac{1}{\sqrt{d}}\right)$!

Hsieh–Kothari [HK22] do local updates to the SDP solution to get $0.878 + \Omega\left(\frac{1}{d^2}\right)$.

Does running MCMC initialized at the SDP solution do anything?

[HK22]: JT Hsieh and P Kothari. Approximating Max-Cut on Bounded Degree Graphs: Tighter Analysis of the FKL Algorithm.

[Tre01]: L Trevisan. Non-approximability results for optimization problems on bounded degree instances.

Open Questions III: Worst-case approximation algorithms from non-worst-case initializations

For max-cut in a graph, SDPs provide a 0.878-approximation to the max-cut.

Conjecture (Folklore, inspired by Trevisan [Tre01])

On bounded degree graphs, should be able to approximate to $0.878 + \Omega\left(\frac{1}{\sqrt{d}}\right)$!

Hsieh–Kothari [HK22] do local updates to the SDP solution to get $0.878 + \Omega\left(\frac{1}{d^2}\right)$.

Does running MCMC initialized at the SDP solution do anything?
Our framework describes how to get sampling guarantees from non-worst-case initializations... can we say anything about inference/optimization guarantees?

[HK22]: JT Hsieh and P Kothari. Approximating Max-Cut on Bounded Degree Graphs: Tighter Analysis of the FKL Algorithm.

[Tre01]: L Trevisan. Non-approximability results for optimization problems on bounded degree instances.

Open Questions III: Worst-case approximation algorithms from non-worst-case initializations

For max-cut in a graph, SDPs provide a 0.878-approximation to the max-cut.

Conjecture (Folklore, inspired by Trevisan [Tre01])

On bounded degree graphs, should be able to approximate to $0.878 + \Omega\left(\frac{1}{\sqrt{d}}\right)$!

Hsieh–Kothari [HK22] do local updates to the SDP solution to get $0.878 + \Omega\left(\frac{1}{d^2}\right)$.

Does running MCMC initialized at the SDP solution do anything?
Our framework describes how to get sampling guarantees from non-worst-case initializations... can we say anything about inference/optimization guarantees?
(Liu–Mohanty–Raghavendra–R–Wu [LMR$^+$24] describes how to do this from worst-case initializations)

[HK22]: JT Hsieh and P Kothari. Approximating Max-Cut on Bounded Degree Graphs: Tighter Analysis of the FKL Algorithm.
[LMR$^+$24]: K Liu, S Mohanty, P Raghavendra, AR, and DX Wu. Locally Stationary Distributions: A Framework for Analyzing Slow-Mixing Markov Chains
[Tre01]: L Trevisan. Non-approximability results for optimization problems on bounded degree instances.

Thank you! Questions?

Feel free to email at **`amit_r@mit.edu`**.

Bibliography I

[AEGP23] Ahmed El Alaoui, Ronen Eldan, Reza Gheissari, and Arianna Piana. Fast relaxation of the random field Ising dynamics. *arXiv preprint arXiv:2311.06171*, 2023.

[Aid98] Shigeki Aida. Uniform positivity improving property, Sobolev inequalities, and spectral gaps. *Journal of functional analysis*, 158(1):152–185, 1998.

[AJ24] Gérard Ben Arous and Aukosh Jagannath. Shattering versus metastability in spin glasses. *Communications on Pure and Applied Mathematics*, 77(1):139–176, 2024.

[AJK+21a] Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Entropic independence i: Modified log-sobolev inequalities for fractionally log-concave distributions and high-temperature ising models. *arXiv preprint arXiv:2106.04105*, 2021.

[AJK+21b] Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Entropic independence ii: optimal sampling and concentration via restricted modified log-sobolev inequalities. *arXiv preprint arXiv:2111.03247*, 2021.

Bibliography II

[AJK$^+$24] Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Universality of spectral independence with applications to fast mixing in spin glasses. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 5029–5056. SIAM, 2024.

[AKV24] Nima Anari, Frederic Koehler, and Thuy-Duong Vuong. Trickle-down in localization schemes and applications. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 1094–1105, 2024.

[Ala24] Ahmed El Alaoui. Near-optimal shattering in the ising pure p-spin and rarity of solutions returned by stable algorithms. *arXiv preprint arXiv:2412.03511*, 2024.

[ALG21] Nima Anari, Kuikui Liu, and Shayan Oveis Gharan. Spectral independence in high-dimensional expanders and applications to the hardcore model. *SIAM Journal on Computing*, (0):FOCS20–1, 2021.

[ALGV19] Nima Anari, Kuikui Liu, Shayan Oveis Gharan, and Cynthia Vinzant. Log-concave polynomials ii: high-dimensional walks and an fpras for counting bases of a matroid. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 1–12, 2019.

Bibliography III

[AMS23a] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from mean-field gibbs measures via diffusion processes. *arXiv preprint arXiv:2310.08912*, 2023.

[AMS23b] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Shattering in pure spherical spin glasses. *arXiv preprint arXiv:2307.04659*, 2023.

[Bar16a] Alexander Barvinok. Approximating permanents and hafnians. *arXiv preprint arXiv:1601.07518*, 2016.

[Bar16b] Alexander Barvinok. *Combinatorics and complexity of partition functions*, volume 30. Springer, 2016.

[BD97a] Russ Bubley and Martin Dyer. Graph orientations with no sink and an approximation for a hard case of# sat. In *Proceedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 248–257, 1997.

[BD97b] Russ Bubley and Martin Dyer. Path coupling: A technique for proving rapid mixing in markov chains. In *Proceedings 38th Annual Symposium on Foundations of Computer Science*, pages 223–231. IEEE, 1997.

Bibliography IV

[BD97c] Russ Bubley and ME Dyer. Path coupling, dobrushin uniqueness, and approximate counting. *RESEARCH REPORT SERIES-UNIVERSITY OF LEEDS SCHOOL OF COMPUTER STUDIES LU SCS RR*, 1997.

[BDJ96] Russ Bubley, Martin Dyer, and Mark Jerrum. *A new approach to polynomial-time generation of random points in convex bodies*. University of Edinburgh. Laboratory for Foundations of Computer Science, 1996.

[BÉ06] Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84: Proceedings*, pages 177–206. Springer, 2006.

[CCYZ24] Xiaoyu Chen, Zongchen Chen, Yitong Yin, and Xinyuan Zhang. Rapid mixing at the uniqueness threshold. *arXiv preprint arXiv:2411.03413*, 2024.

[CE22] Yuansi Chen and Ronen Eldan. Hit-and-run mixing via localization schemes. *arXiv preprint arXiv:2212.00297*, 2022.

[Che23] Sinho Chewi. Log-concave sampling. *Book draft available at https://chewisinho. github. io*, 2023.

Bibliography V

[CHS93] Alessandro Crisanti, Herbert Horner, and H-J Sommers. The spherical p-spin interaction spin-glass model: the dynamics. *Zeitschrift für Physik B Condensed Matter*, 92:257–271, 1993.

[CLMM23] Zongchen Chen, Kuikui Liu, Nitya Mani, and Ankur Moitra. Strong spatial mixing for colorings on trees and its algorithmic applications. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 810–845. IEEE, 2023.

[CLV21] Zongchen Chen, Kuikui Liu, and Eric Vigoda. Optimal mixing of glauber dynamics: Entropy factorization via high-dimensional expansion. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1537–1550, 2021.

[CLV24] Zongchen Chen, Kuikui Liu, and Eric Vigoda. Spectral independence via stability and applications to holant-type problems. *TheoretiCS*, 3, 2024.

[CMS24] Pietro Caputo, Florentin Münch, and Justin Salez. Entropy and curvature: beyond the peres-tetali conjecture. *arXiv preprint arXiv:2401.17148*, 2024.

[DSVW04] Martin Dyer, Alistair Sinclair, Eric Vigoda, and Dror Weitz. Mixing in time and space for lattice spin systems: A combinatorial view. *Random Structures & Algorithms*, 24(4):461–479, 2004.

Bibliography VI

[EHMT17] Matthias Erbar, Christopher Henderson, Georg Menz, and Prasad Tetali. Ricci curvature bounds for weakly interacting markov chains. 2017.

[EKZ22] Ronen Eldan, Frederic Koehler, and Ofer Zeitouni. A spectral condition for spectral gap: fast mixing in high-temperature ising models. *Probability theory and related fields*, 182(3):1035–1051, 2022.

[GJ19] Reza Gheissari and Aukosh Jagannath. On the spectral gap of spherical spin glass dynamics. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 55(2):756 – 776, 2019.

[HK22] Jun-Ting Hsieh and Pravesh K Kothari. Approximating max-cut on bounded degree graphs: Tighter analysis of the fkl algorithm. *arXiv preprint arXiv:2206.09204*, 2022.

[HMMR05] Shlomo Hoory, Avner Magen, Steven Myers, and Charles Rackoff. Simple permutations mix well. *Theoretical computer science*, 348(2-3):251–261, 2005.

[HMP24] Brice Huang, Andrea Montanari, and Huy Tuan Pham. Sampling from Spherical Spin Glasses in Total Variation via Algorithmic Stochastic Localization. *arXiv preprint arXiv:2404.15651*, 2024.

Bibliography VII

[Jer95] Mark Jerrum. A very simple algorithm for estimating the number of k-colorings of a low-degree graph. *Random Structures & Algorithms*, 7(2):157–165, 1995.

[JS89] Mark Jerrum and Alistair Sinclair. Approximating the permanent. *SIAM journal on computing*, 18(6):1149–1178, 1989.

[JSV04] Mark Jerrum, Alistair Sinclair, and Eric Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *Journal of the ACM (JACM)*, 51(4):671–697, 2004.

[KLV23] Frederic Koehler, Holden Lee, and Thuy-Duong Vuong. Efficiently learning and sampling multimodal distributions with data-based initialization. *arXiv preprint arXiv:2411.09117*, 2023.

[KV23] Frederic Koehler and Thuy-Duong Vuong. Sampling multimodal distributions with the vanilla score: Benefits of data-based initialization. *arXiv preprint arXiv:2310.01762*, 2023.

[Liu23] Kuikui Liu. *Spectral Independence a New Tool to Analyze Markov Chains*. PhD thesis, University of Washington, 2023.

Bibliography VIII

[LMR$^+$24] Kuikui Liu, Sidhanth Mohanty, Prasad Raghavendra, Amit Rajaraman, and David X Wu. Locally stationary distributions: A framework for analyzing slow-mixing markov chains. *arXiv preprint arXiv:2405.20849*, 2024.

[LMRW24] Kuikui Liu, Sidhanth Mohanty, Amit Rajaraman, and David X Wu. Fast mixing in sparse random ising models. *arXiv preprint arXiv:2405.06616*, 2024.

[Lov99] László Lovász. Hit-and-run mixes fast. *Mathematical programming*, 86:443–461, 1999.

[LS93] László Lovász and Miklós Simonovits. Random walks in a convex body and an improved volume algorithm. *Random structures & algorithms*, 4(4):359–412, 1993.

[LV23] Aditi Laddha and Santosh S Vempala. Convergence of gibbs sampling: Coordinate hit-and-run mixes fast. *Discrete & Computational Geometry*, 70(2):406–425, 2023.

[LY52] Tsung-Dao Lee and Chen-Ning Yang. Statistical theory of equations of state and phase transitions. ii. lattice gas and ising model. *Physical Review*, 87(3):410, 1952.

[RW01] Michael Röckner and Feng-Yu Wang. Weak Poincaré inequalities and L2-convergence rates of Markov semigroups. *Journal of Functional Analysis*, 185(2):564–603, 2001.

Bibliography IX

[Tre01] Luca Trevisan. Non-approximability results for optimization problems on bounded degree instances. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 453–461, 2001.

[Vil09] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[Wei04] Dror Weitz. *Mixing in time and space for discrete spin systems*. University of California, Berkeley, 2004.

[Wei06] Dror Weitz. Counting independent sets up to the tree threshold. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 140–149, 2006.