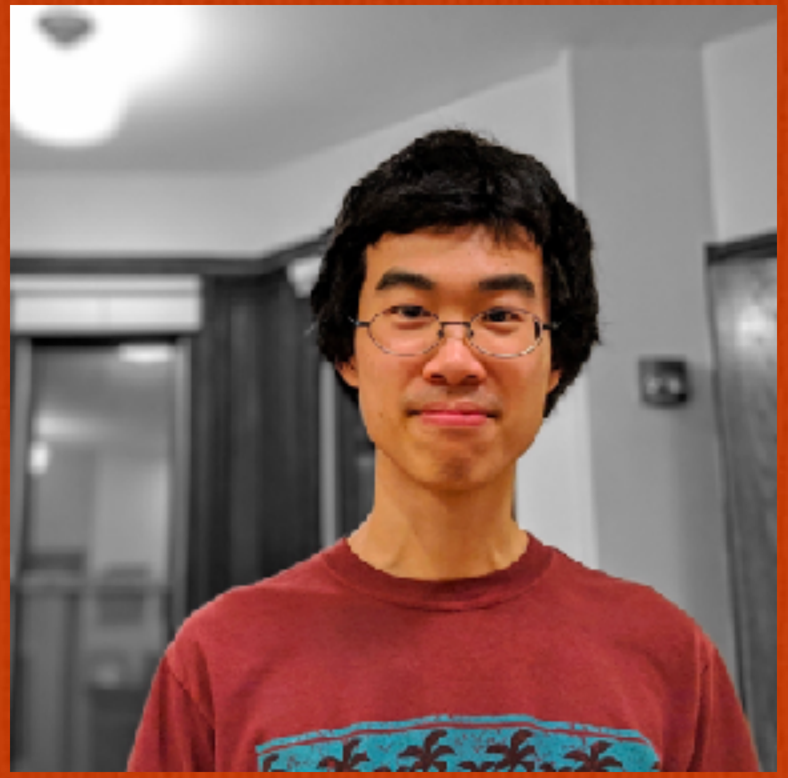
When MCMC Meets Variational Methods

Frederic Koehler (Stanford => UChicago)

Holden Lee (JHU)



Andrej Risteski (CMU)

Reminder:

Ising models,

sampling,

and all that

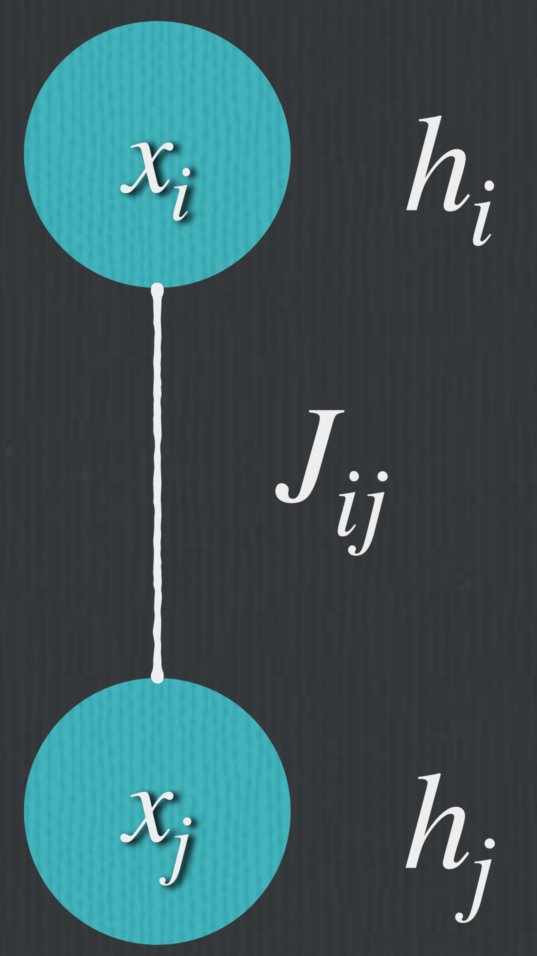
Ising Model

$$p_{J,h}(x) \propto \exp\left(\frac{1}{2}\langle x, Jx\rangle + \langle h, x\rangle\right), \qquad x \in \{\pm 1\}^n$$

- ☐ $J : n \times n$ is an arbitrary symmetric interaction matrix, and $h \in \mathbb{R}^n$ is a vector of bias/external field

- ☐ General class including models of magnetism, spin glasses, neurons, social networks, Bayesian statistics…

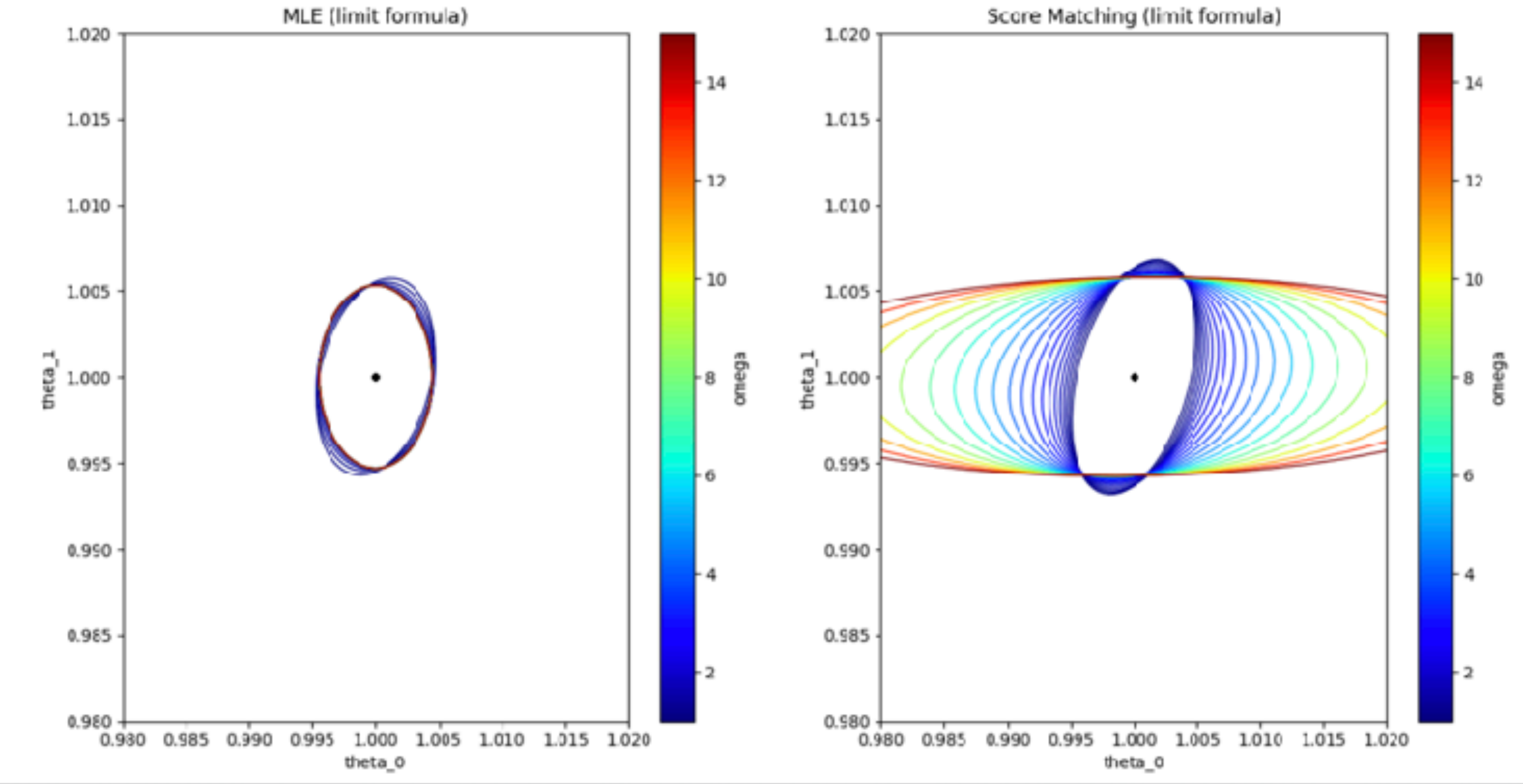
- ☐ Maximum entropy distribution. Analogue of Gaussian.



Sampling and Statistical Inference

- ☐ Sampling: given parameters J, h generate $X \sim p_{J,h}$
- ☐ Many connections to statistical inference.
- ☐ Bayesian posteriors often Ising.
  - ☐ Ising is the posterior on $X \sim Uni\{\pm 1\}^n$ given $Y = X + N(0,\Sigma)$ for some Σ, Y
- ☐ Sampling used to optimally estimate J,h from samples (via MLE.)

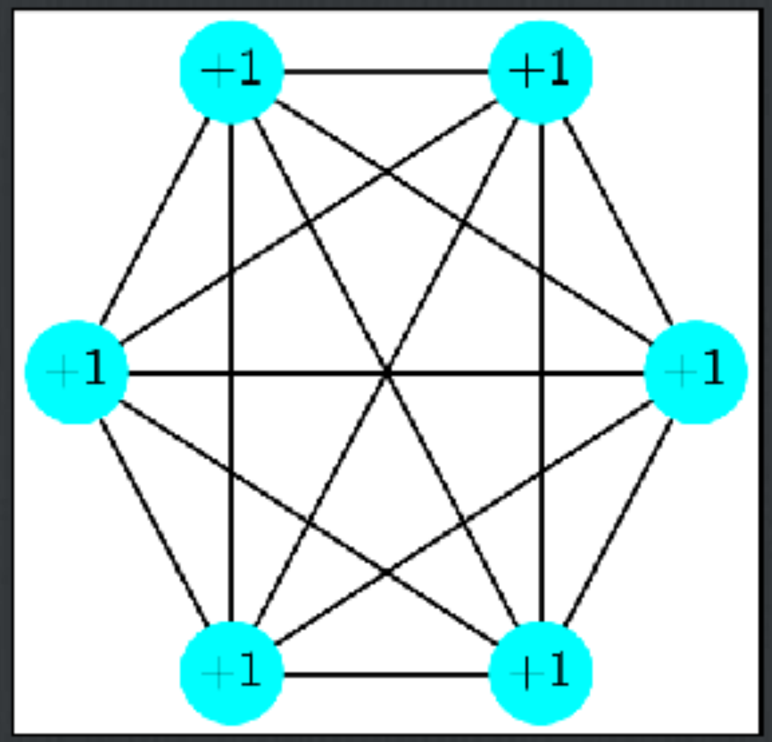
stochastic block model (community detection) posterior is Ising



Alternatives to MLE (maximum likelihood estimation) are usually not as accurate…

How to sample? Typically MCMC. Especially nice is the Gibbs sampler/Glauber dynamics:

- ☐ Let initial $X \in \{\pm 1\}^n$ be arbitrary.
- ☐ For t = 1 to T :
    - ☐ Pick $i \in [n]$ randomly and resample $X_i \mid X_{\sim i}$



But: it is hard to know if it is working! Sometimes it doesn't work!

- ☐ How many steps to mix to correct distribution? Can be exponential time…
- ☐ Some Ising models are NP-hard to sample from.

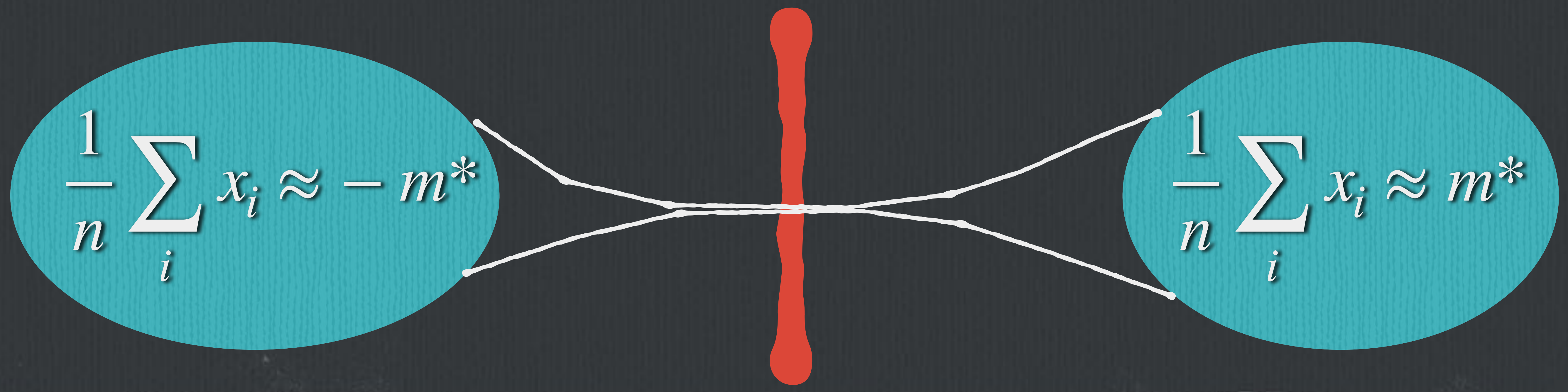
What can we say theoretically?

Useful example: Curie-Weiss model

$$p(x) = \frac{1}{Z} \exp\left(\frac{\beta}{2n} \left(\sum_i x_i \right)^2 \right)$$

- Rapid mixing for $\beta < 1$ (Dobrushin).

- When $\beta > 1$: **bottleneck emerges** between two clusters of spins. Gibbs sampler is **exponentially slow to mix** (runtime c^n)…



Gibbs Sampler at High Temperature

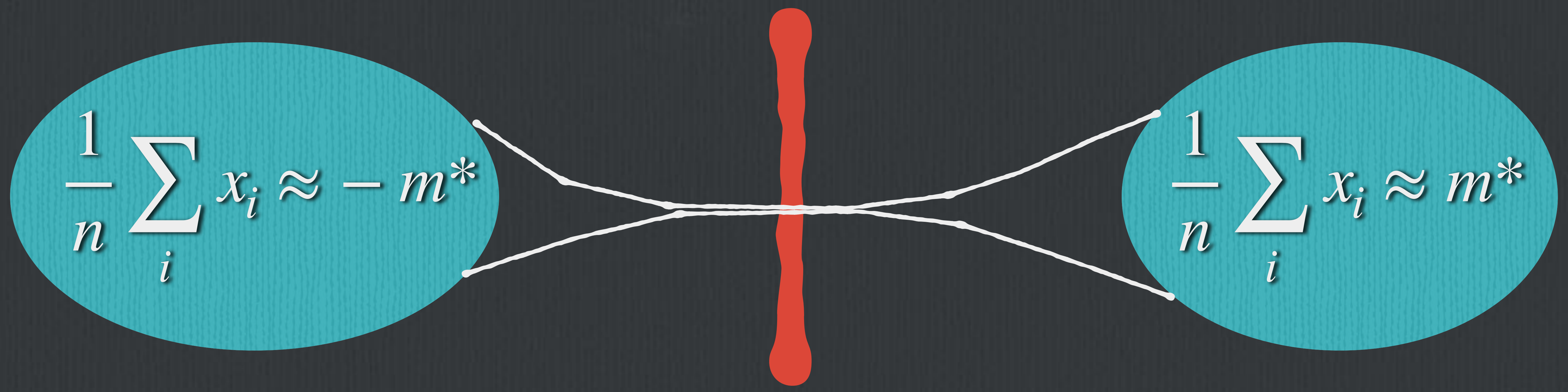
Rapid mixing of Gibbs/Glauber for "small" J is a generic phenomena.
Lots of work for many years…

Theorem […, BB '19, EKZ '21, AJKPV '22]: if $\lambda_{max}(J) - \lambda_{min}(J) < 1$, then Gibbs sampler mixes rapidly.

Not true for "larger" J due to bottleneck (previous slide). What are the alternatives to Gibbs?

Variational Inference Picture

- Variational inference: **approximate by a simpler distribution**. Popular alternative to MCMC, comes from statistical physics.

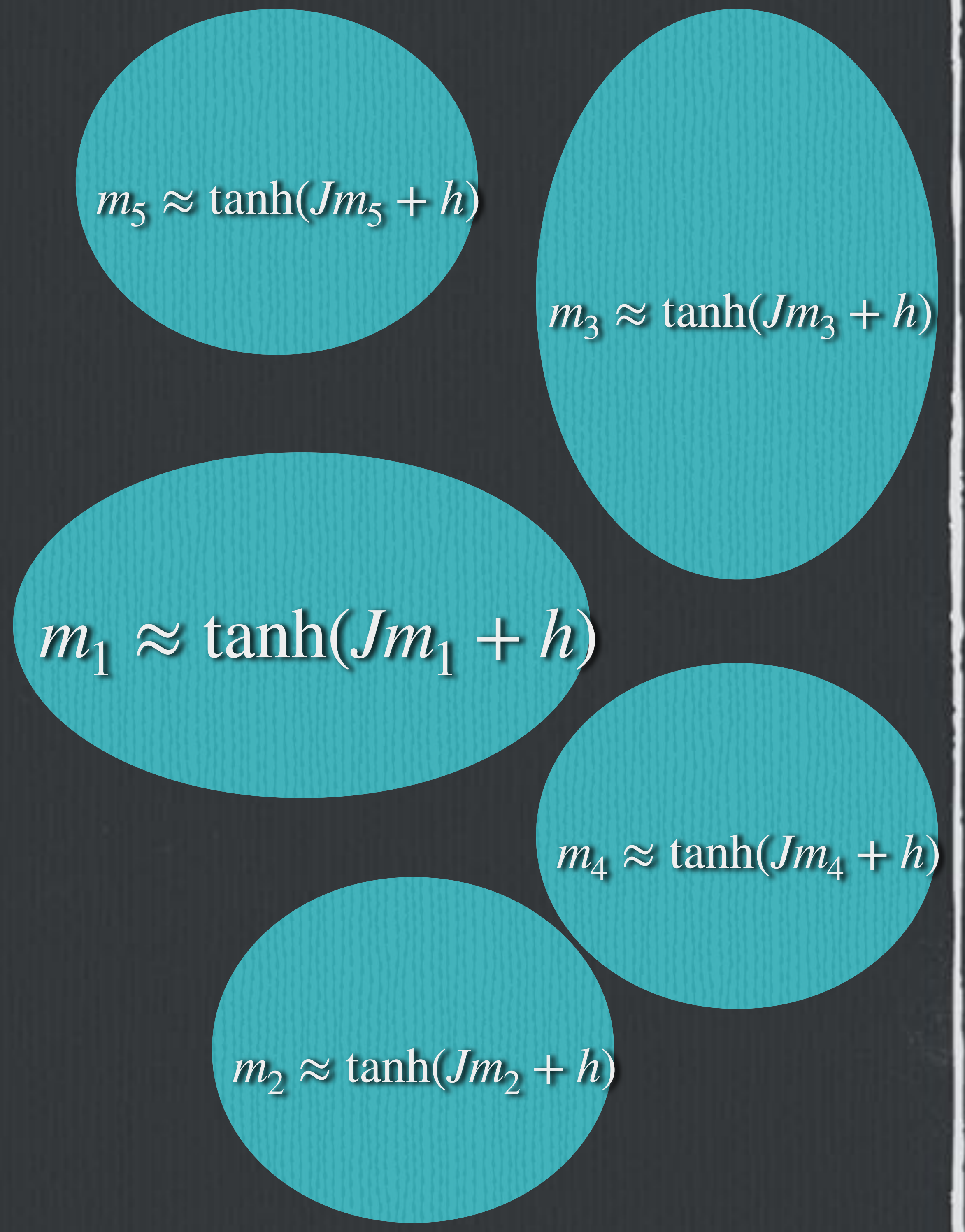


- When $\beta > 1$, Curie-Weiss model is close to a **mixture of two product measures** centered at fixed points of "mean-field equation" $m^* = \tanh(\beta m^* + h)$

Structure of low-rank Ising models

- Rigorous Naive Mean-Field Approximation [..., Eldan-Gross '18, Austin '19,...] shows approximate mixture of product decomposition for all low-rank J.
  - If rank J = $o(n)$ then Ising $\approx$ mixture of $2^{o(n)}$ product measures (with means m satisfying $m \approx \tanh(Jm + h)$).
  - Only a rough approximation ($o(n)$ in Wasserstein).
  - Not constructive (unless you use our results!)

$$p_{J,h} \approx$$



Mixture of products

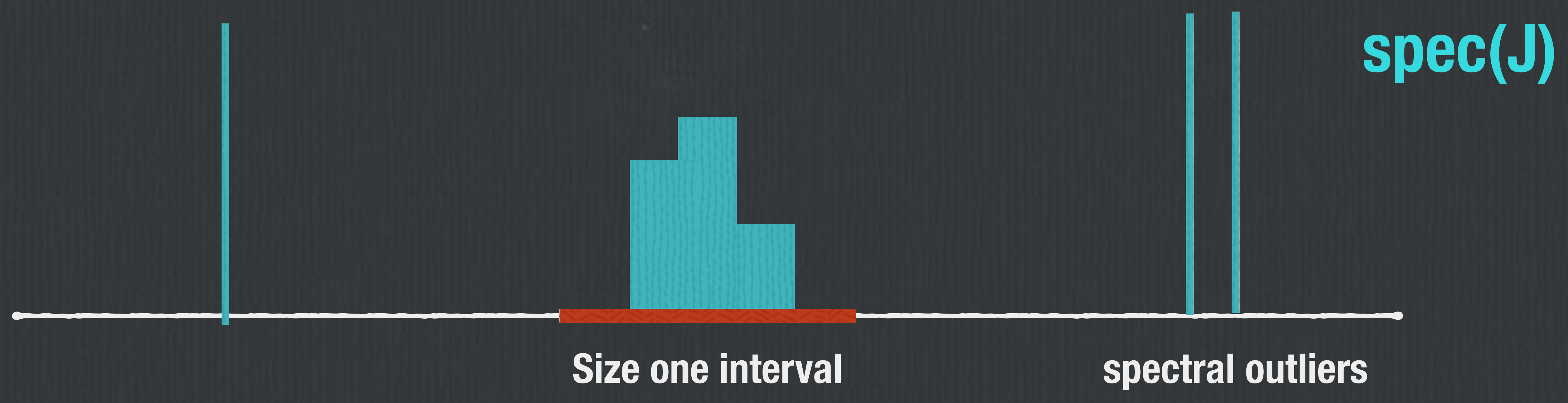
MCMC versus Variational Inference

- Variational inference world
  - GOOD: makes sense in multimodal/low-temperature settings.
  - BAD: only approximates the true distribution, structural results are not algorithmic.
- MCMC world
  - GOOD: when it works, it really samples from the true distribution.
  - BAD: Gibbs sampler fails in multimodal case. Unclear how to fix…

MCMC meets Variational Inference

Theorem (this work): new sampler for approximately low-rank Ising models!

Runtime parameterized by (# of spectral outliers/“threshold rank” of J).



$$J = J_{\mathsf{LR}} + J_{\mathsf{small}} \qquad \lambda_{max}(J_{\mathsf{small}}) - \lambda_{min}(J_{\mathsf{small}}) < 1$$

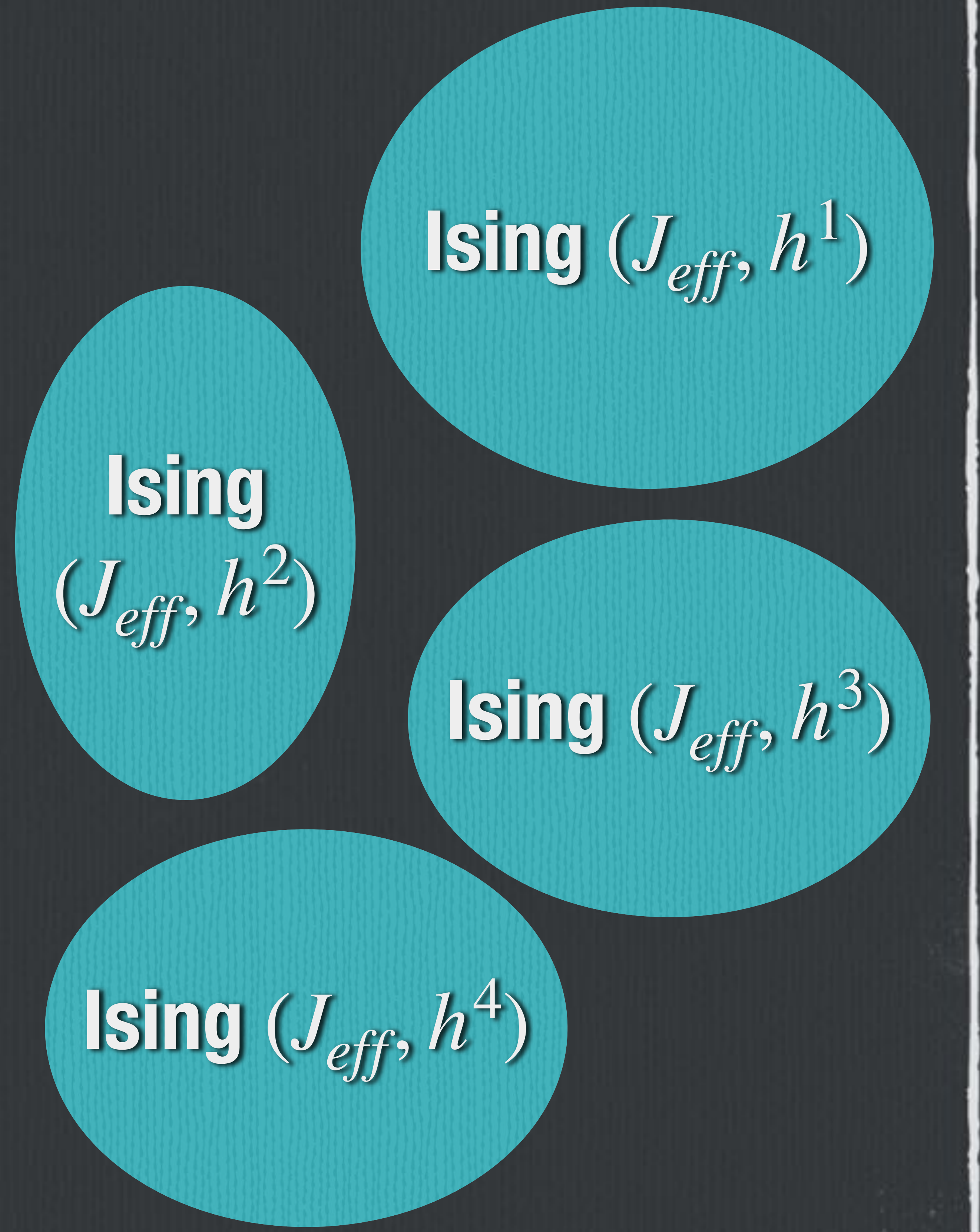
- ☐ Runtime $n^{O(\#outliers)}$. Close to lower bound from ETH [JKR ’19].
- ☐ Negative eigenvalues must be $O(1)$ in size. Large negative eigenvalue problem is NP-hard.

Proof Ideas

Key: a new structural result

- We decompose the Ising model as a mixture of high-temperature Ising models: $(J_{eff}, h^j)_{j=1}^{M}$ with $\|J_{eff}\|_{OP}$ small.
- Unlike mixture of product measures: decomposition is (1) constructive, and (2) very accurate.
- Enables polynomial-time sampling of the real distribution!

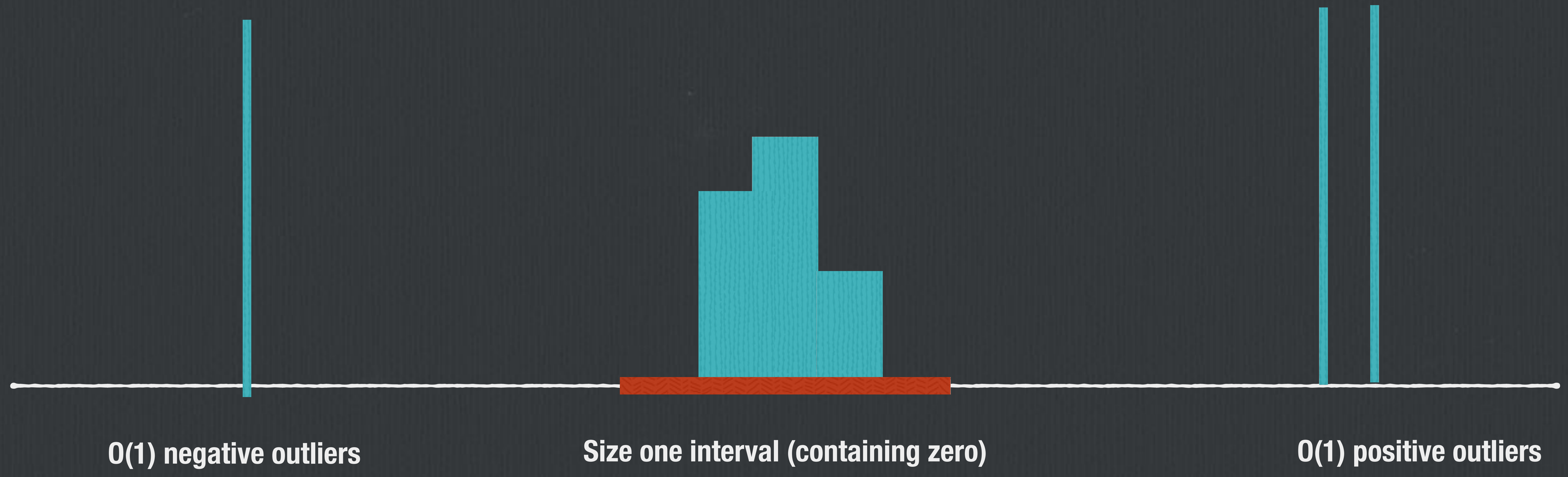
$$p_{J,h} \approx$$



Mixture of high-temperature models!

Two key principles

$$\mathrm{spec}(J)$$



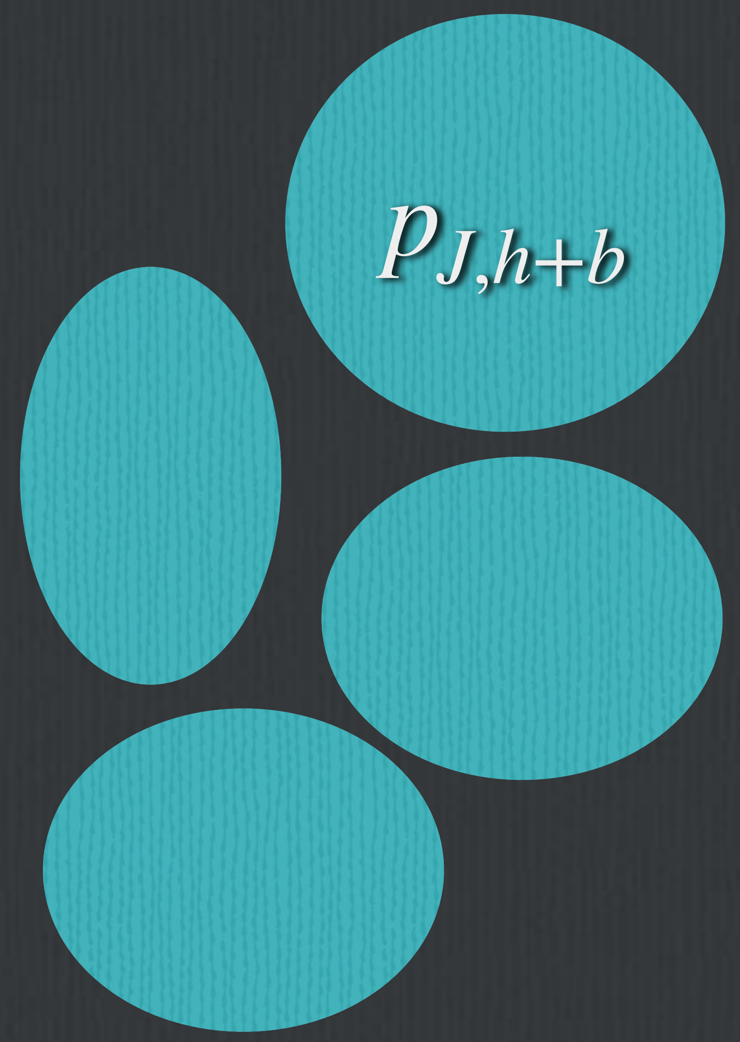
II. Eliminate negative eigenvalues: replace by an “effective field”.

I. Eliminate positive eigenvalues: low-dimensional decomposition

Step I: positive eigenvalues

- ☐ Suppose $J = J_{LR} + J'$ (LR = low rank) with J_{LR} PSD and J' small.
- ☐ Goal: find a mixing distribution q over a low-dimensional space so that approximately,

$$p_{J,h} \approx \int p_{J',h+b}\, q(b)\, db$$



Mixture of
high-temperature Ising
with weights $q(b)$

- **Hubbard-Stratonovich transform** [Hubbard '58]. **Quadratic to linear** reduction:

$$e^{\langle x, J_{LR} x\rangle/2} = \frac{1}{(2\pi)^{d/2}} \int_{span(J_{LR})} e^{\langle J_{LR}^{1/2} x, z\rangle}\, e^{-\|z\|^2/2}\, dz$$

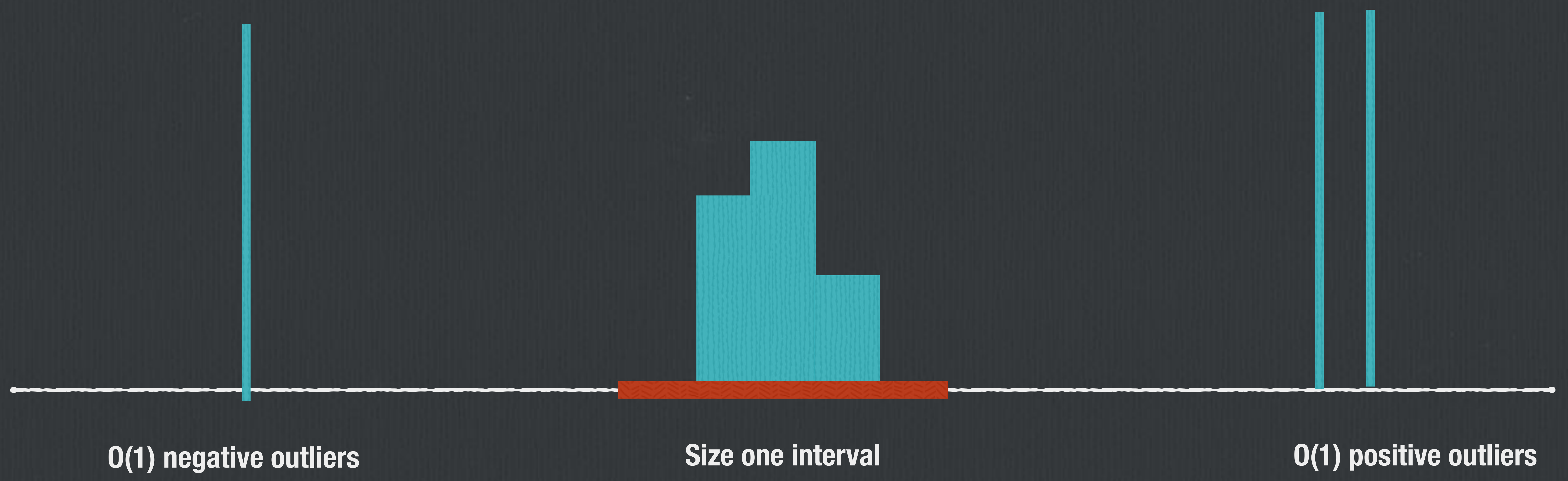
- Using $J = J' + J_{LR}$ yields mixture decomposition:

$$e^{\langle x, Jx\rangle/2} \propto \int_{span(J_{LR})} e^{\langle x, J'x\rangle/2 + \langle J_{LR}^{1/2} x, z\rangle}\, e^{-\|z\|^2/2}\, dz$$

$$\propto \int_{span(J_{LR})} p_{J', J_{LR}^{1/2} z}(x)\, \underbrace{Z_{J', J_{LR}^{1/2} z}\, e^{-\|z\|^2/2}}_{q(z)}\, dz$$

Two key principles

$$\text{spec}(J)$$



II. Eliminate negative eigenvalues: replace by an “effective field”.

I. Eliminate positive eigenvalues: low-dimensional decomposition (sketch done!)

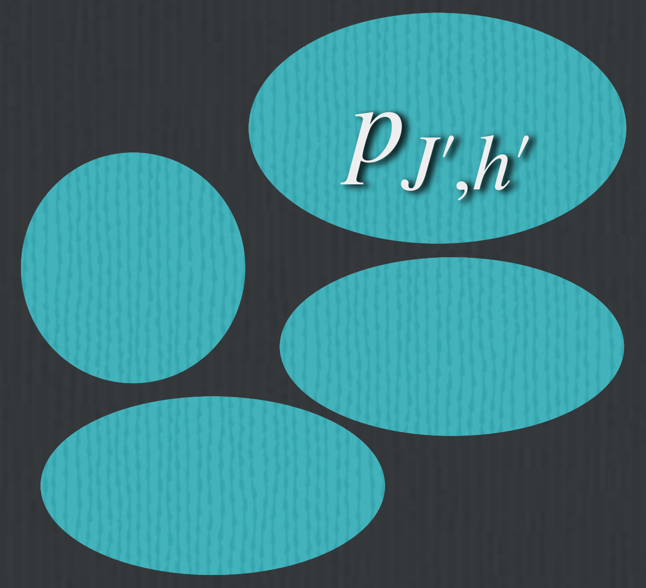
Step II: handling negative eigenvalues

- **Trick:** write $J = J_+ - J\_$ with $J_+, J\_ \geq 0$ and run SGD on

$$G(\mu) := \log \mathbb{E}_{P_{J_+,h}}[e^{\langle \mu, -J\_x \rangle}] + \langle \mu, J\_\mu \rangle / 2$$

- **Why? Postulate** that J_x concentrates near deterministic quantity J_μ .
  - Then $\langle x, Jx \rangle \approx \langle x, J_+ x \rangle - \langle x, J\_\mu \rangle$ so $P_{J,h} \approx P_{J_+, h - J\_\mu}$ (kill negative eigenvalues)
  - In particular $J\_\mu \approx \mathbb{E}_{P_{J,h}}[J\_x] \approx \mathbb{E}_{P_{J_+, h-J\_\mu}}[J\_x]$ (so $\nabla G(\mu) \approx 0$)
- **FACTS:** (1) G has a critical point, (2) at any critical point $P_{J,h} \approx P_{J_+, h-J\_\mu}$ for rejection sampling

The Simplified Algorithm



- Now we know there exists an **approximate mixture decomposition into "hot" models:**

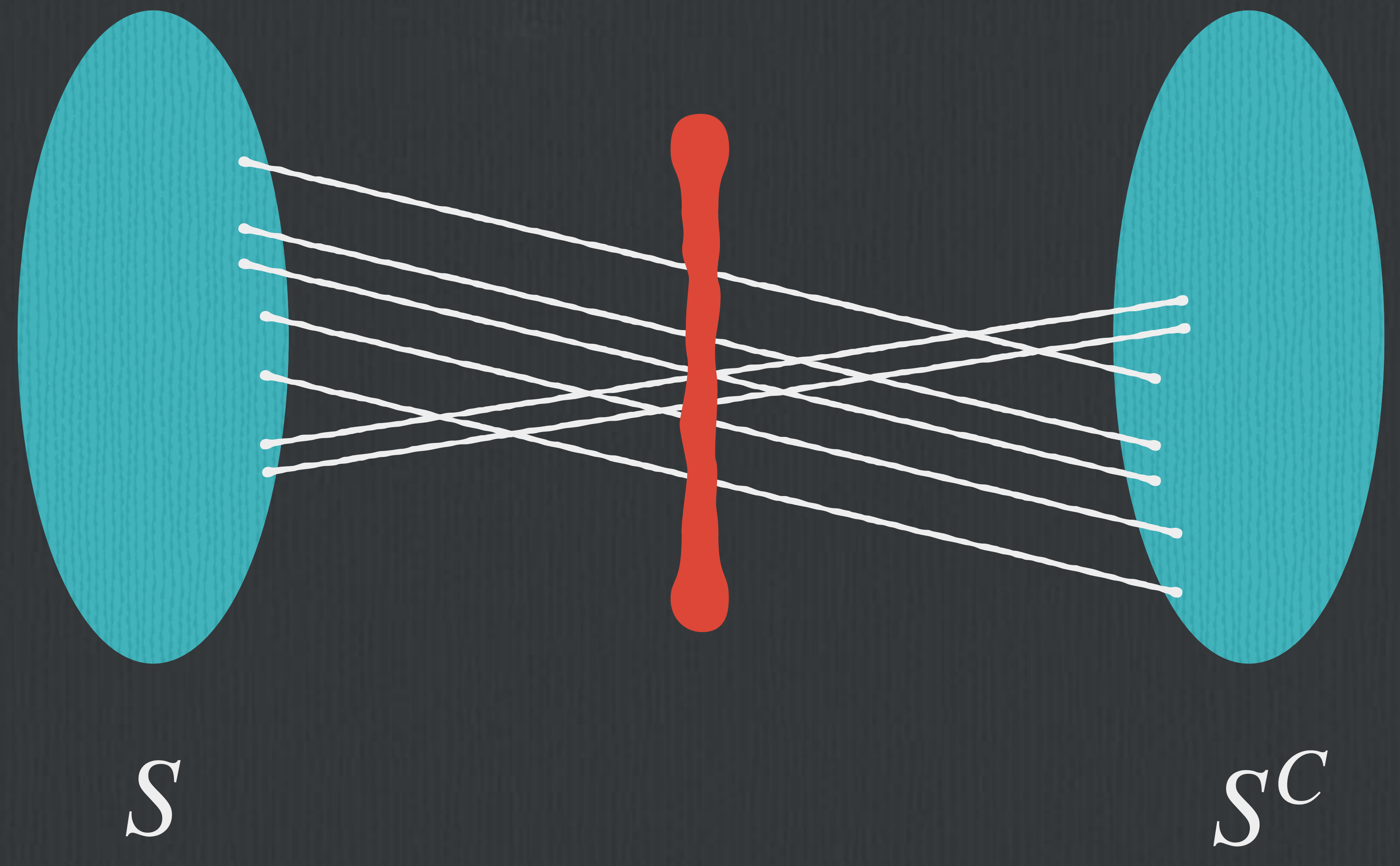
$$p_{J,h}(x) \approx \int p_{J',h+b-J\_\mu(b)}(x)\, q(b)\, db \qquad \|J'\|_{OP} \leq 1$$

- **Yields a natural sampling algorithm:**
  - (1) Riemann integration for integral, (2) SGD+Glauber to compute critical point $\mu(b)$, (3) Glauber dynamics for $p_{J'}$, (4) rejection sampling.
  - Using this gives suboptimal $poly(1/\epsilon)$ runtime in error ϵ. We can get $\log(1/\epsilon)$ runtime by designing a "**simulated tempering**" chain. (see paper)

Example Application

Task: Dense MAX-CUT

- ☐ Input : adjacency matrix A of a graph on n vertices with $\Theta(n^2)$ edges
- ☐ Goal: find a set $S \subseteq [n]$ to maximize size of cut $\#\{(u,v) \in E : u \in S, v \in S^C\}$



- ☐ NP hard but admits a PTAS [AKK ’92, dlV '92].

“Statistical Physicsy” Approach

- Let OPT $:= \dfrac{|E|}{2n} + \max_{x \in \{\pm 1\}^n} \dfrac{-1}{4n} \langle x, Ax \rangle$.   ((1/n) * Dense MAX-CUT so OPT is $\Theta(n)$)

- Gibbs measure at inverse temperature $\beta \geq 0$:

$$p_\beta(x) \propto \exp\left(\frac{-\beta}{4n}\langle x, Ax\rangle\right)$$

- Exercise (typical sample is a $(1 + 1/\beta)$-apx to OPT):

$$\text{OPT} \geq \frac{|E|}{2\beta n} + \mathbb{E}_{p_\beta}\left[\frac{-1}{4n}\langle x, Ax\rangle\right] \geq \text{OPT} - \frac{n \log 2}{\beta}$$

- Exercise (few large eigenvalues): $\dfrac{1}{n^2}\displaystyle\sum_{i=1}^{n} \lambda_i(A)^2 = O(1)$

Approximation from Sampling

- CORR: for any $\beta \geq 0$, can sample Gibbs measure $p_\beta(x) \propto \exp\left(\frac{-\beta}{4n}\langle x, Ax\rangle\right)$ in time $n^{O(\beta^2)}$
    - Yields a $(1 + 1/\beta)$ approximation to Dense MAX-CUT.
- Matches $n^{O(1/\epsilon^2)}$ runtime [AKK '92] for Dense MAX-CUT but now get all near-optimal cuts!
- No computational phase transition in β ! Sampling *easier* than optimization !
- Q: what other PTAS's can be found by "just" sampling the Gibbs measure?

Conclusion

- We developed a new algorithm which provably samples from all "approximately low rank" Ising models.
  - Many other interesting applications: Hopfield networks, Ferromagnetic SK, Contextual SBM, mixture models, Ising models on expanders, …
- Variational inference ideas help us predict which problems are tractable.
- What aspects of this story extend beyond Ising ?

Thanks!