CS 59300 – Algorithms for Data Science

Classical and Quantum approaches

Lecture 11 (10/09)

Sampling and variational inference

https://ruizhezhang.com/course_fall_2025.html

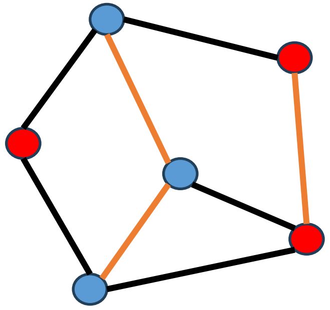
Example

Ising model

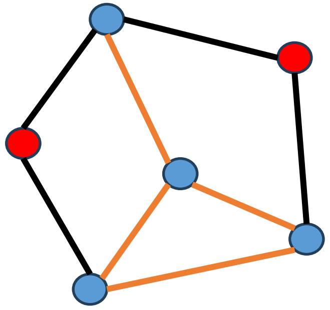
- Graph $G = (V, E)$
- Parameter $\beta \in \mathbb{R}$
- Configuration $\sigma \in \{+1, -1\}^V$ with weight:

$$wt(\sigma) = \exp(\beta \cdot \#\text{monochromatic edges})$$

- Gibbs distribution: $\pi_{\text{Ising}}(\sigma) = \frac{wt(\sigma)}{Z_{\text{Ising}}(\beta)} = \frac{wt(\sigma)}{\sum_{\tau \in \{-1,1\}^V} wt(\tau)}$



$e^{3\beta}$



$e^{4\beta}$

Example

Ising model

$$\pi_{\text{Ising}}(\sigma) = \frac{wt(\sigma)}{Z_{\text{Ising}}(\beta)} = \frac{wt(\sigma)}{\sum_{\tau \in \{0,1\}^V} wt(\tau)}$$
$$\propto \exp\left(\beta \sum_{ij \in E} \frac{1 + \sigma_i \sigma_j}{2}\right)$$

- **Sampling:** can you efficiently draw samples from the Gibbs distribution π_{Ising}?
- **Optimization:** can you minimize the Hamiltonian $H(\sigma) \coloneqq \sum_{ij \in E} \frac{1+\sigma_i \sigma_j}{2}$ for $\sigma \in \{\pm 1\}^V$?
- **Partition function estimation:** can you approximate $Z_{\text{Ising}}(\beta)$ to within ϵ error?

Applications

Statistical inference

$$\Pr[\Theta \mid X] = \frac{\Pr[X \mid \Theta] \cdot \Pr[\Theta]}{\Pr[X]}$$

$$\Pr[X] = \int \Pr[X \mid \Theta] \cdot \Pr[\Theta]\, d\Theta$$

Volume estimation

- Given access to a high-dim convex body $\mathcal{K}$ (via membership oracle or constraints)
- Estimate $\mathrm{Vol}(\mathcal{K})$

Statistical mechanics and phase transitions

- Graphical model:

$$H(\sigma) \coloneqq -\sum_{ij \in E} \psi_{ij}(\sigma_i, \sigma_j) \quad \forall \sigma \in [q]^V$$

$$\mu_\beta(\sigma) \propto \exp(-\beta H(\sigma))$$

- Independent sets, matchings, colorings …

Fairness and Differential privacy

- Detecting gerrymandering: randomly sample redistricting plans from an appropriate distribution
- Exponential mechanism for ϵ-DP:

$$\pi(x) \propto \exp(\beta u(x, D))$$

where u is the utility function

Today's plan

- A canonical approach for sampling is via Markov chains
  - Design a Markov chain such that the target distribution π is its fixed point
  - Simulate the Markov chain on any initial point for T steps
  - Prove the mixing time of this Markov chain

- In the past few lectures, we've introduced various convex relaxations and rounding algorithms

- Today, we'll see another approach that uses relaxation+rounding: Variational Inference and Mean-field approximation

Starting point

Gibbs variational principle. Let Ω be a finite state space. Then the Shannon entropy function

$$\mu \mapsto -H(\mu) = \sum_{x\in\Omega} \mu(x)\log\mu(x)$$

on probability measures over Ω is smooth and strictly convex

Furthermore, for every function $f:\Omega \to \mathbb{R}$,

"free energy" $$\mathcal{F} \coloneqq \log \sum_{x\in\Omega} e^{f(x)} = \sup_{\nu}\{\mathbb{E}_{x\sim\nu}[f(x)] + H(\nu)\}$$

and the supremum is uniquely attained at the Gibbs measure $\mu(x) \propto e^{f(x)}$

Proof of the Gibbs variational principle

- For any distributions p and q over Ω, the KL divergence is defined as:

$$D_{\mathrm{KL}}(p\|q) \coloneqq \sum_{x\in\Omega} p(x)\log\left(\frac{p(x)}{q(x)}\right) = -H(p) - \mathbb{E}_{x\sim p}[\log q(x)]$$

- $D_{\mathrm{KL}}(p\|q) \geq 0$ with equality iff $p = q$
- Let $p \coloneqq \nu$ and $q \coloneqq \mu = e^f/Z_f$ where $Z_f = \sum_{x\in\Omega} e^{f(x)}$ is the partition function
- Then, we have

$$0 \leq D_{\mathrm{KL}}(\nu\|\mu) = -H(\nu) - \mathbb{E}_{x\sim\nu}\left[\log\left(e^{f(x)}/Z_f\right)\right] = -H(\nu) - \mathbb{E}_{x\sim\nu}[f(x)] + \log Z_f$$

$\blacksquare$

Starting point

Gibbs variational principle. Let Ω be a finite state space. Then the Shannon entropy function

$$\mu \mapsto -H(\mu) = \sum_{x \in \Omega} \mu(x) \log \mu(x)$$

on probability measures over Ω is smooth and strictly convex

Furthermore, for every function $f : \Omega \to \mathbb{R}$,

"free energy" $$\mathcal{F} \coloneqq \log \sum_{x \in \Omega} e^{f(x)} = \sup_{\nu} \{\mathbb{E}_{x \sim \nu}[f(x)] + H(\nu)\}$$

Issue: Ω is **exp** large!

and the supremum is uniquely attained at the Gibbs measure $\mu(x) \propto e^{f(x)}$

Estimating log-partition function → Maximizing a concave function

The naïve mean-field approximation

- **Idea:** restrict the class of probability measure ν in the optimization
- Product measure over $\{\pm 1\}^n$ where $n \coloneqq |\Omega|$:

$$\mathcal{F}_{\mathrm{NMF}} \coloneqq \sup_{\nu \text{ product}} \{\mathbb{E}_{x\sim\nu}[f(x)] + H(\nu)\}$$
$$= \sup_{\boldsymbol{m}\in[-1,1]^n} \{\mathbb{E}_{x\sim\pi(\boldsymbol{m})}[f(x)] + H(\pi(\boldsymbol{m}))\} \xleftrightarrow{?} \mathcal{F}$$

- Every product measure is uniquely identified by its mean vector $\boldsymbol{m} \in [-1,1]^n$
- The entropy can be explicitly calculated:

$$H(\pi(\boldsymbol{m})) = -\sum_{i=1}^{n}\left(\frac{1+\boldsymbol{m}_i}{2}\log\frac{1+\boldsymbol{m}_i}{2} + \frac{1-\boldsymbol{m}_i}{2}\log\frac{1-\boldsymbol{m}_i}{2}\right)$$

- For many natural f (e.g., quadratic form), $\mathbb{E}_{x\sim\pi(\boldsymbol{m})}[f(x)]$ is also easy to compute

How good is the mean-field approximation?

The Gibbs measure $\mu \propto e^f$ exhibits mean-field behavior (as $n \to \infty$) if

$$\frac{\mathcal{F} - \mathcal{F}_{\text{NMF}}}{n} = o(1)$$

- $o(n)$-additive approximation to $\mathcal{F}$ $\quad\Longleftrightarrow\quad$ $e^{o(n)}$-multiplicative approximation to Z_f
- Related to the asymptotic free energy density:

$$\lim_{n\to\infty} \frac{1}{n}\mathcal{F} \approx_{o(1)} \lim_{n\to\infty} \frac{1}{n}\mathcal{F}_{\text{NMF}}$$

  - Can derive many physically interesting quantities, e.g. magnetization, specific heat, susceptibility
  - Can predict phase transitions by the differentiability/continuity/smoothness of the asymptotic free energy density in the model parameters (e.g. β)

NMF approximation error for Ising models

Theorem 1 (Jain-Koehler-Risteski '19).

For a symmetric interaction matrix $A \in \mathbb{R}^{n \times n}$, and consider the Ising Gibbs measure

$$\mu(\sigma) \propto e^{f(\sigma)} \quad \text{where} \quad f(\sigma) = \frac{1}{2}\sigma^\top A \sigma$$

Then, $\mathcal{F} - \mathcal{F}_{\mathrm{NMF}} = \mathcal{O}\left(n^{2/3}\|A\|_F^{2/3}\right)$

Example 1:

- Consider $A = \frac{\beta}{d} A_G$ where G is a d-regular graph and A_G is the adjacency matrix
- $\|A\|_F^{2/3} = (\beta/d)^{2/3}(dn)^{1/3}$
- $\mathcal{F} - \mathcal{F}_{\mathrm{NMF}} = \mathcal{O}\left(n\beta^{2/3}d^{-1/3}\right)$ "NMF works better on dense problem"

NMF approximation error for Ising models

Theorem 2 (Eldan '20).

For a symmetric interaction matrix $A \in \mathbb{R}^{n\times n}$, and consider the Ising Gibbs measure

$$\mu(\sigma) \propto e^{f(\sigma)} \quad \text{where} \quad f(\sigma) = \frac{1}{2}\sigma^\top A\sigma$$

Then, $\mathcal{F} - \mathcal{F}_{\mathrm{NMF}} \leq 3\log\det\left(I + L^{1/2}\mathrm{Cov}(\mu)L^{1/2}\right)$, where $L \coloneqq (A^2)^{1/2}$

Example 2:

- Consider $A = \frac{\beta}{n}\mathbf{1}\mathbf{1}^\top$
- $\mathcal{F} - \mathcal{F}_{\mathrm{NMF}} \leq 3\log(n\beta)$ instead of $\beta n^{2/3}$

Sherali-Adams hierarchy

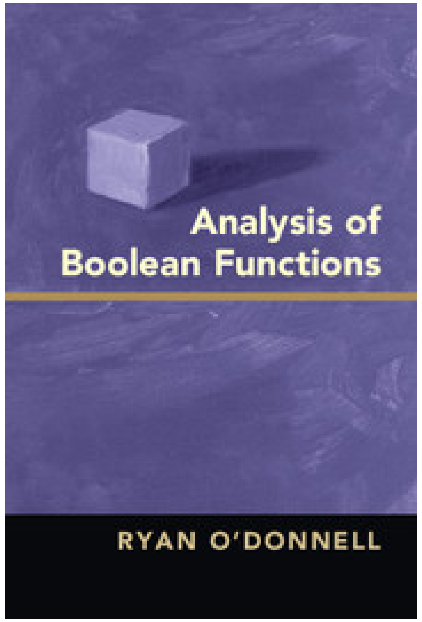
Low-degree function

Let $f : \{\pm 1\}^n \to \mathbb{R}$ be an arbitrary function. Then there is a unique multi-affine polynomial

$$\sum_{S \subseteq [n]} \hat{f}(S) \prod_{i \in S} x_i$$

which agrees with f on $\{\pm 1\}^n$

- $\hat{f}(S)$ are the Fourier coefficients of f
- $\mathrm{supp}(f) \coloneqq \{S \subseteq [n] : \hat{f}(S) \neq 0\}$ is the support of f
- $\deg(f) \coloneqq \max_{S \in \mathrm{supp}(f)} |S|$ is the degree of f



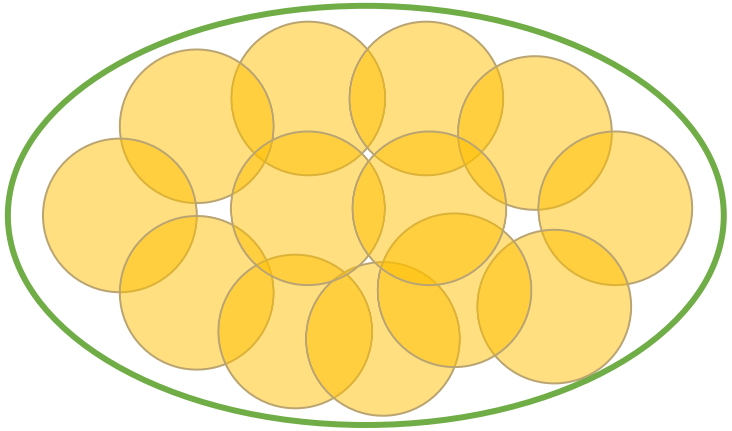
Sherali-Adams pseudo-distribution

Pseudo-distribution

Let $\mathfrak{F} \subseteq 2^{[n]}$ be a downwards closed family of subsets (i.e. if $T \in \mathfrak{F}$ and $S \subseteq T$, then $S \in \mathfrak{F}$)
An $\mathfrak{F}$-pseudo-distribution over $\{\pm 1\}^n$ is a collection $\tilde{\boldsymbol{p}} = \{\tilde{\boldsymbol{p}}_S\}_{S \in \mathfrak{F}}$ of probability distributions $\tilde{p}_S$ over $\{\pm 1\}^S$ satisfying the following local consistency relations:

$$\tilde{\boldsymbol{p}}_S[\tau] = \Pr_{\sigma \sim \tilde{\boldsymbol{p}}_T}[\sigma_S = \tau], \qquad \forall S, T \in \mathfrak{F} \ s.t.\ S \subseteq T, \qquad \forall \tau \in \{\pm 1\}^S$$

- The degree of the pseudo-distribution is $\max_{S \in \mathfrak{F}} |S|$



- For a degree-k pseudo-distribution,

$$\#\mathrm{para} = \sum_{S \subseteq \mathfrak{F}} 2^{|S|} \leq n^{\mathcal{O}(k)}$$

- Every genuine distribution μ is a pseudo-distribution $\{\mu_S\}$, where μ_S is the marginal distribution on S

Reverse direction?

Sherali-Adams pseudo-distribution

Pseudo-distribution

Let $\mathfrak{F} \subseteq 2^{[n]}$ be a downwards closed family of subsets (i.e. if $T \in \mathfrak{F}$ and $S \subseteq T$, then $S \in \mathfrak{F}$)

An $\mathfrak{F}$-pseudo-distribution over $\{\pm 1\}^n$ is a collection $\tilde{\boldsymbol{p}} = \{\tilde{\boldsymbol{p}}_S\}_{S \in \mathfrak{F}}$ of probability distributions $\tilde{p}_S$ over $\{\pm 1\}^S$ satisfying the following local consistency relations:

$$\tilde{\boldsymbol{p}}_S[\tau] = \Pr_{\sigma \sim \tilde{\boldsymbol{p}}_T}[\sigma_S = \tau], \qquad \forall S, T \in \mathfrak{F} \;\; s.t. \;\; S \subseteq T, \qquad \forall \tau \in \{\pm 1\}^S$$

- The degree of the pseudo-distribution is $\max_{S \in \mathfrak{F}} |S|$

Counterexample

$$\left.\begin{array}{l} \bullet\;\; n = 3 \text{ and } \mathfrak{F} = \big\{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}\big\} \\ \bullet\;\; \tilde{\boldsymbol{p}}_i[i = \pm 1] = 1/2 \\ \bullet\;\; \tilde{\boldsymbol{p}}_{ij}[i = 1, j = -1] = \tilde{\boldsymbol{p}}_{ij}[i = -1, j = 1] = 1/2 \end{array}\right\} \text{a degree-2 pseudo-distribution}$$

- No global distribution $\boldsymbol{p}_{123}$ can exist since $\{1,2,3\}$ cannot be all distinct

Sherali-Adams pseudo-distribution

Pseudo-distribution

Let $\mathfrak{F} \subseteq 2^{[n]}$ be a downwards closed family of subsets (i.e. if $T \in \mathfrak{F}$ and $S \subseteq T$, then $S \in \mathfrak{F}$)

An $\mathfrak{F}$-pseudo-distribution over $\{\pm 1\}^n$ is a collection $\tilde{\boldsymbol{p}} = \{\tilde{\boldsymbol{p}}_S\}_{S \in \mathfrak{F}}$ of probability distributions $\tilde{\boldsymbol{p}}_S$ over $\{\pm 1\}^S$ satisfying the following local consistency relations:

$$\tilde{\boldsymbol{p}}_S[\tau] = \Pr_{\sigma \sim \tilde{\boldsymbol{p}}_T}[\sigma_S = \tau], \qquad \forall S, T \in \mathfrak{F} \;\; s.t. \;\; S \subseteq T, \qquad \forall \tau \in \{\pm 1\}^S$$

- The degree of the pseudo-distribution is $\max_{S \in \mathfrak{F}} |S|$
- Pseudo-expectation: for any $f(x) = \sum_S c_S \prod_{i \in S} x_i$ with $\mathrm{supp}(f) \subseteq \mathfrak{F}$,

$$\tilde{\mathbb{E}}[f] \coloneqq \sum_S c_S \mathbb{E}_{\sigma_S \sim \tilde{\boldsymbol{p}}_S}\left[\prod_{i \in S} \sigma_i\right]$$

The Bethe approximation (level-2 Sherali-Adams)

Let $G = (V, E)$ with adjacency matrix A, and consider the Ising Gibbs measure

$$\mu(\sigma) \propto e^{f(\sigma)} \quad \text{where} \quad f(\sigma) = \frac{1}{2}\sigma^\top A\sigma$$

Let $\mathfrak{F}$ be the downwards closure of the set of edges E, i.e. $\mathfrak{F} = \{\emptyset\} \cup \big\{\{v\} : v \in \mathrm{V}\big\} \cup \big\{\{u, v\} : uv \in E\big\}$

Define the Bethe free energy by

$$\mathcal{F}_{\text{Bethe}} \coloneqq \sup_{\mathfrak{F}-\text{pseudo}-\text{dist.}\,\tilde{\boldsymbol{p}}} \big\{\tilde{\mathbb{E}}[f] + H_{\text{Bethe}}(\tilde{\boldsymbol{p}})\big\}$$

where H_{Bethe} is the Bethe entropy:

$$H_{\text{Bethe}}(\tilde{\boldsymbol{p}}) \coloneqq \sum_{e \in E} H(\tilde{\mathrm{p}}_e) - \sum_{v \in V} (\deg(v) - 1) H(\tilde{\boldsymbol{p}}_v)$$

$$= \sum_{v \in V} H(\tilde{\boldsymbol{p}}_v) - \sum_{uv \in E} I(u; v)$$

“correct double-counting”

The Bethe entropy

$$H_{\text{Bethe}}(\widetilde{\boldsymbol{p}}) := \sum_{e \in E} H(\widetilde{\boldsymbol{p}}_e) - \sum_{v \in V} (\deg(v) - 1) H(\widetilde{\boldsymbol{p}}_v)$$

Fact. Let T be a tree and $\boldsymbol{p}$ be any probability distribution defined on T. Then,

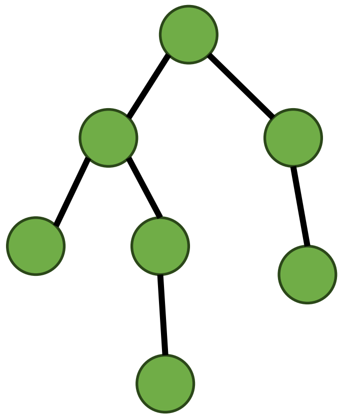
$$\boldsymbol{p}(\sigma) = \frac{\prod_{uv \in E} \boldsymbol{p}_{uv}(\sigma_u, \sigma_v)}{\prod_{v \in \mathrm{V}} \left(\boldsymbol{p}_v(\sigma_v)\right)^{\deg(v)-1}} \qquad \forall \sigma \in \{\pm 1\}^V$$

$$H(\boldsymbol{p}) = -\sum_{\sigma \in \{\pm 1\}^V} \boldsymbol{p}(\sigma) \log \boldsymbol{p}(\sigma) = \sum_{\sigma \in \{\pm 1\}^V} \boldsymbol{p}(\sigma) \left(\sum_{v \in V} (\deg(v) - 1) \log \boldsymbol{p}_v(\sigma_v) - \sum_{uv \in E} \log \boldsymbol{p}_{uv}(\sigma_u, \sigma_v) \right)$$

$$= \sum_{v \in V} (\deg(v) - 1) \sum_{\sigma_v} \boldsymbol{p}_v(\sigma_v) \log \boldsymbol{p}_v(\sigma_v) - \sum_{uv \in E} \sum_{\sigma_u, \sigma_v} \boldsymbol{p}_{uv}(\sigma_u, \sigma_v) \log \boldsymbol{p}_{uv}(\sigma_u, \sigma_v)$$

$$= \sum_{e \in E} H(\boldsymbol{p}_e) - \sum_{v \in V} (\deg(v) - 1) H(\boldsymbol{p}_v)$$

$$= H_{\text{Bethe}}(\boldsymbol{p})$$



The Bethe approximation (level-2 Sherali-Adams)

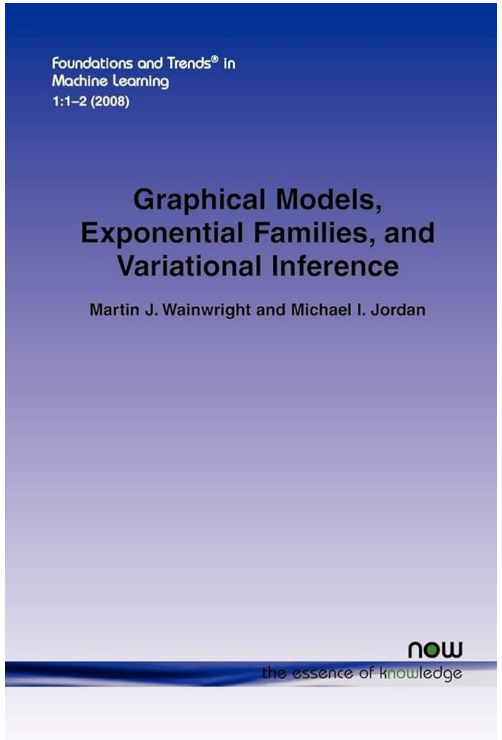
Let $G = (V, E)$ with adjacency matrix A, and consider the Ising Gibbs measure

$$\mu(\sigma) \propto e^{f(\sigma)} \quad \text{where} \quad f(\sigma) = \frac{1}{2}\sigma^\top A\sigma$$

Let $\mathfrak{F}$ be the downwards closure of the set of edges E, i.e. $\mathfrak{F} = \{\emptyset\} \cup \big\{\{v\} : v \in \mathrm{V}\big\} \cup \big\{\{u,v\} : uv \in E\big\}$

Define the Bethe free energy by

$$\mathcal{F}_{\mathrm{Bethe}} := \sup_{\mathfrak{F}-\text{pseudo}-\text{dist. }\tilde{\boldsymbol{p}}} \left\{\widetilde{\mathbb{E}}[f] + H_{\mathrm{Bethe}}(\tilde{\boldsymbol{p}})\right\}$$



- Widely used for approximating the free energy of sparse graphical models
- The optimizer of $\mathcal{F}_{\mathrm{Bethe}}$ gives the belief propagation equations

Higher-level Sherali-Adams

- Define $\mathfrak{F}_k \coloneqq \binom{[n]}{\leq k}$, and $\mathrm{SA}(k;[n])$ be the set of all $\mathfrak{F}_k$-pseudo-distributions

Conditioning a pseudo-distribution

Let $\widetilde{\boldsymbol{p}} \in \mathrm{SA}(k;[n])$. For any $S \in \mathfrak{F}_{k-1}$, and any $\tau \in \{\pm 1\}^S$, define the conditional pseudo-distribution:

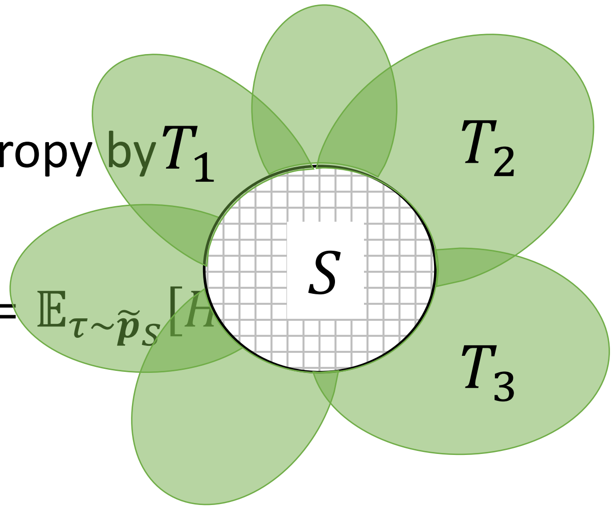
$$\widetilde{\boldsymbol{p}}_T^\tau(\sigma) \coloneqq \widetilde{\boldsymbol{p}}_{S\cup T}(\tau \circ \sigma) \qquad \forall T \in \binom{[n]\backslash S}{\leq k - |S|}, \forall \sigma \in \{\pm 1\}^T$$

Then, $\widetilde{\boldsymbol{p}}^\tau \in \mathrm{SA}(k - |S|; [n]\backslash S)$

Augmented pseudo-entropy

Let $\widetilde{\boldsymbol{p}} \in \mathrm{SA}(k;[n])$. For $0 \leq j \leq k-1$, define the j-th augmented pseudo-entropy by

$$\widetilde{H}_j(\widetilde{\boldsymbol{p}}) \coloneqq \min_{|S|\leq j}\left\{H(\widetilde{\boldsymbol{p}}_S) + \sum_{i\notin S} H(\widetilde{\boldsymbol{p}}_i|\widetilde{\boldsymbol{p}}_S)\right\} \quad \text{where} \quad H(\widetilde{\boldsymbol{p}}_i|\widetilde{\boldsymbol{p}}_S) \coloneqq \mathbb{E}_{\tau\sim\widetilde{\boldsymbol{p}}_S}[H$$



Higher-level Sherali-Adams

- Define $\mathfrak{F}_k \coloneqq \binom{[n]}{\le k}$, and $\mathrm{SA}(k;[n])$ be the set of all $\mathfrak{F}_k$-pseudo-distributions

Augmented pseudo-entropy

Let $\widetilde{\boldsymbol{p}} \in \mathrm{SA}(k;[n])$. For $0 \le j \le k-1$, define the j-th augmented pseudo-entropy by

$$\widetilde{H}_j(\widetilde{\boldsymbol{p}}) \coloneqq \min_{|S| \le j} \left\{ H(\widetilde{\boldsymbol{p}}_S) + \sum_{i \notin S} H(\widetilde{\boldsymbol{p}}_i | \widetilde{\boldsymbol{p}}_S) \right\} \quad \text{where} \quad H(\widetilde{\boldsymbol{p}}_i | \widetilde{\boldsymbol{p}}_S) \coloneqq \mathbb{E}_{\tau \sim \widetilde{p}_S}[H(\widetilde{\boldsymbol{p}}_i^\tau)]$$

Sherali-Adams free energy

Let $f: \{\pm 1\}^n \to \mathbb{R}$ with $\deg(f) \le k$. For $0 \le j \le k-1$, define

$$\mathcal{F}_{\mathrm{SA}(k;[n]),j} \coloneqq \sup_{\widetilde{\boldsymbol{p}} \in \mathrm{SA}(k;[n])} \{\widetilde{\mathbb{E}}[f] + \widetilde{H}_j(\widetilde{\boldsymbol{p}})\}$$

Augmented pseudo-entropy

$$\widetilde{H}_j(\widetilde{\boldsymbol{p}}) \coloneqq \min_{|S|\le j}\left\{H(\widetilde{\boldsymbol{p}}_S) + \sum_{i\notin S} H(\widetilde{\boldsymbol{p}}_i|\widetilde{\boldsymbol{p}}_S)\right\} \quad \text{where} \quad H(\widetilde{\boldsymbol{p}}_i|\widetilde{\boldsymbol{p}}_S) \coloneqq \mathbb{E}_{\tau\sim\widetilde{\boldsymbol{p}}_S}[H(\widetilde{\boldsymbol{p}}_i^{\tau})]$$

Lemma. For every $0 \le j \le k-1$, the function $\widetilde{\boldsymbol{p}} \mapsto H_j(\widetilde{\boldsymbol{p}})$ over $\mathrm{SA}(k;[n])$ satisfies:

1) For every genuine probability distribution μ, $H(\mu) \le \widetilde{H}_j(\mu)$

Proof.

- Let $\boldsymbol{X} \sim \mu$. By the chain rule of Shannon entropy,

$$\begin{aligned} H(\boldsymbol{X}) &= H(\boldsymbol{X}_S) + H\left(\boldsymbol{X}_{[n]\setminus S} \mid \boldsymbol{X}_S\right) \\ &\le H(\boldsymbol{X}_S) + \sum_{i\in[n]\setminus S} H(\boldsymbol{X}_i \mid \boldsymbol{X}_S) \end{aligned}$$

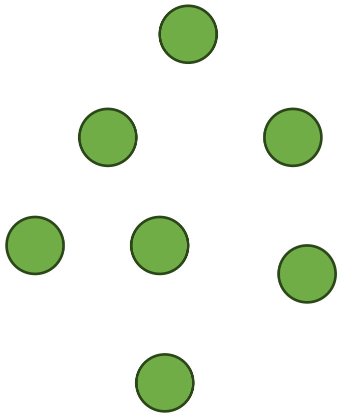
"Maximum Entropy Principle"

Augmented pseudo-entropy

$$\widetilde{H}_j(\widetilde{\boldsymbol{p}}) \coloneqq \min_{|S|\le j} \left\{ H(\widetilde{\boldsymbol{p}}_S) + \sum_{i\notin S} H(\widetilde{\boldsymbol{p}}_i | \widetilde{\boldsymbol{p}}_S) \right\} \quad \text{where} \quad H(\widetilde{\boldsymbol{p}}_i|\widetilde{\boldsymbol{p}}_S) \coloneqq \mathbb{E}_{\tau\sim\widetilde{\boldsymbol{p}}_S}[H(\widetilde{\boldsymbol{p}}_i^{\tau})]$$

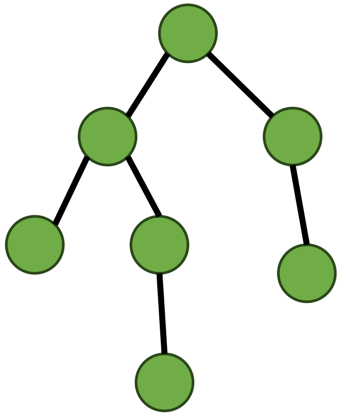
Lemma. For every $0 \le j \le k-1$, the function $\widetilde{\boldsymbol{p}} \mapsto H_j(\widetilde{\boldsymbol{p}})$ over $\mathrm{SA}(k;[n])$ satisfies:

1) For every genuine probability distribution μ, $H(\mu) \le \widetilde{H}_j(\mu)$



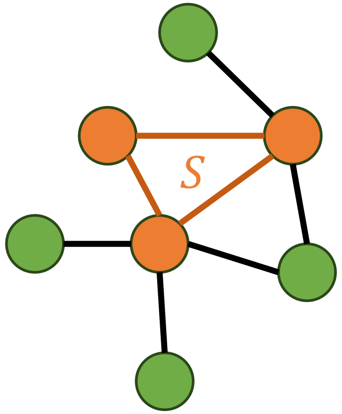
Mean-field entropy

$$\sum_v H(\boldsymbol{p}_v)$$



Bethe entropy

$$\sum_e H(\boldsymbol{p}_e) - \sum_v (\deg(v)-1)H(\boldsymbol{p}_v)$$



Augmented pseudo-entropy

$$H(\boldsymbol{p}_S) + \sum_{i\notin S} H(\boldsymbol{p}_i|\boldsymbol{p}_S)$$

Augmented pseudo-entropy

$$\widetilde{H}_j(\widetilde{\boldsymbol{p}}) \coloneqq \min_{|S|\leq j}\left\{H(\widetilde{\boldsymbol{p}}_S) + \sum_{i\notin S} H(\widetilde{\boldsymbol{p}}_i|\widetilde{\boldsymbol{p}}_S)\right\} \quad \text{where} \quad H(\widetilde{\boldsymbol{p}}_i|\widetilde{\boldsymbol{p}}_S) \coloneqq \mathbb{E}_{\tau\sim\widetilde{\boldsymbol{p}}_S}[H(\widetilde{\boldsymbol{p}}_i^{\tau})]$$

Lemma. For every $0 \leq j \leq k-1$, the function $\widetilde{\boldsymbol{p}} \mapsto H_j(\widetilde{\boldsymbol{p}})$ over $\mathrm{SA}(k;[n])$ satisfies:

1) For every genuine probability distribution μ, $H(\mu) \leq \widetilde{H}_j(\mu)$
2) The function is concave over $\mathrm{SA}(k;[n])$

Proof.

- $\mathrm{SA}(k;[n])$ is convex: for $\widetilde{\boldsymbol{p}}, \widetilde{\boldsymbol{q}} \in \mathrm{SA}(k;[n])$, $\lambda\widetilde{\boldsymbol{p}} + (1-\lambda)\widetilde{\boldsymbol{q}} \in \mathrm{SA}(k;[n])$
- Concavity is preserved under $\sum$ and min $\implies$ It suffices to show that $H(\widetilde{\boldsymbol{p}}_S)$ and $H(\widetilde{\boldsymbol{p}}_i|\widetilde{\boldsymbol{p}}_S)$ are concave
- Follows from the standard proof of concavity of Shannon entropy

Augmented pseudo-entropy

$$\widetilde{H}_j(\widetilde{\boldsymbol{p}}) \coloneqq \min_{|S|\le j}\left\{H(\widetilde{\boldsymbol{p}}_S) + \sum_{i\notin S} H(\widetilde{\boldsymbol{p}}_i|\widetilde{\boldsymbol{p}}_S)\right\} \quad \text{where} \quad H(\widetilde{\boldsymbol{p}}_i|\widetilde{\boldsymbol{p}}_S) \coloneqq \mathbb{E}_{\tau\sim\widetilde{\boldsymbol{p}}_S}[H(\widetilde{\boldsymbol{p}}_i^{\tau})]$$

Lemma. For every $0 \le j \le k-1$, the function $\widetilde{\boldsymbol{p}} \mapsto H_j(\widetilde{\boldsymbol{p}})$ over $\mathrm{SA}(k;[n])$ satisfies:

1) For every genuine probability distribution μ, $H(\mu) \le \widetilde{H}_j(\mu)$

2) The function is concave over $\mathrm{SA}(k;[n])$

- By 1), $\mathcal{F}_{\mathrm{SA}(k;[n]),j} \coloneqq \sup_{\widetilde{\boldsymbol{p}}\in\mathrm{SA}(k;[n])}\{\widetilde{\mathbb{E}}[f] + \widetilde{H}_j(\widetilde{\boldsymbol{p}})\} \ge \mathcal{F}$
- By 2), $\mathcal{F}_{\mathrm{SA}(k;[n]),j}$ is a constrained convex optimization problem of size $n^{\mathcal{O}(k)}$, which can be solved in $n^{\mathcal{O}(k)}$-time

SA approximation error

Theorem 3 (Risteski ’16).

For a symmetric interaction matrix $A \in \mathbb{R}^{n\times n}$, and consider the Ising Gibbs measure

$$\mu(\sigma) \propto e^{f(\sigma)} \quad \text{where} \quad f(\sigma) = \frac{1}{2}\sigma^\top A\sigma$$

For $0 \le k \le n-2$,

$$0 \le \mathcal{F}_{\mathrm{SA}(k+2;[n]),k} - \mathcal{F} \le \mathcal{O}\left(n\|A\|_F/\sqrt{k}\right)$$

Moreover, if $\tilde{\boldsymbol{p}}$ is the optimal pseudo-distribution, then we can round it into a product measure π satisfying

$$\mathcal{F} - \mathcal{F}_{\mathrm{NMF}} \le \quad \mathcal{F} - \left(\mathbb{E}_\pi[f] + H(\pi)\right) \le \mathcal{O}\left(n\|A\|_F/\sqrt{k} + k\right)$$

$(n\|A\|_F)^{2/3}$ by balancing the two terms

Rounding the pseudo-distribution

- Let $\tilde{p}$ be the optimal pseudo-distribution. Fix $S \subseteq [n]$ with $|S| \leq k$
- Define a mixture of product distributions:
  - Sample $\tau \sim \tilde{p}_S$
  - Sample $\sigma \in \{\pm 1\}^n$ according to a product measure π^τ defined by:

$$\pi_i^\tau = \begin{cases} \delta_{\tau_i} & \forall i \in S \\ \tilde{p}_i^\tau & \forall i \notin S \end{cases} \qquad \sigma_S := \tau$$

- We'll prove that for the optimal $S^\star \subseteq [n]$ with $|S^\star| \leq k$,

$$\mathcal{F}_{\mathrm{SA}(k+2;[n]),k} - \left(\mathbb{E}_\nu[f] + H(\nu)\right) \leq \mathcal{O}\left(n\|A\|_F/\sqrt{k}\right) \quad \text{where } \nu = \mathbb{E}_{\tau\sim\tilde{p}_{S^\star}}[\pi^\tau]$$

- Since $\mathbb{E}_\nu[f] + H(\nu) \leq \mathcal{F}$, it implies that $\mathcal{F}_{\mathrm{SA}(k+2;[n]),k} - \mathcal{F} \leq \mathcal{O}\left(n\|A\|_F/\sqrt{k}\right)$

Rounding the pseudo-distribution

- Let $\tilde{p}$ be the optimal pseudo-distribution. Fix $S \subseteq [n]$ with $|S| \leq k$
- Define a mixture of product distributions:
  - Sample $\tau \sim \tilde{p}_S$
  - Sample $\sigma \in \{\pm 1\}^n$ according to a product measure π^τ defined by:

$$\pi_i^\tau = \begin{cases} \delta_{\tau_i} & \forall i \in S \\ \tilde{p}_i^\tau & \forall i \notin S \end{cases} \qquad \sigma_S \coloneqq \tau$$

- We'll prove that for the optimal $S^\star \subseteq [n]$ with $|S^\star| \leq k$,

$$\mathcal{F}_{\mathrm{SA}(k+2;[n]),k} - \left(\mathbb{E}_\nu[f] + H(\nu)\right) \leq \mathcal{O}\left(n\|A\|_F/\sqrt{k}\right) \quad \text{where } \nu = \mathbb{E}_{\tau\sim\tilde{p}_{S^\star}}[\pi^\tau]$$

- For rounding, notice that $H(\nu) = H(\tilde{p}_{S^\star}) + \mathbb{E}_{\tau\sim\tilde{p}_{S^\star}}[H(\pi^\tau)] \leq |S^\star| + \mathbb{E}_{\tau\sim\tilde{p}_{S^\star}}[H(\pi^\tau)] \leq \mathcal{O}(k) + \mathbb{E}_{\tau\sim\tilde{p}_{S^\star}}[H(\pi^\tau)]$. We can take $\tau^\star$ that maximizes $\mathbb{E}_{\pi^\tau}[f] + H(\pi^\tau)$:

$$\mathcal{F} - \left(\mathbb{E}_{\pi^{\tau^\star}}[f] + H\left(\pi^{\tau^\star}\right)\right) \leq \mathcal{F}_{\mathrm{SA}(k+2;[n]),k} - \left(\mathbb{E}_{\pi^{\tau^\star}}[f] + H\left(\pi^{\tau^\star}\right)\right) \leq \mathcal{O}\left(n\|A\|_F/\sqrt{k} + k\right)$$

Rounding the pseudo-distribution

- Let $\tilde{p}$ be the optimal pseudo-distribution. Fix $S \subseteq [n]$ with $|S| \leq k$
- Define a mixture of product distributions:
  - Sample $\tau \sim \tilde{p}_S$
  - Sample $\sigma \in \{\pm 1\}^n$ according to a product measure π^τ defined by:

$$\pi_i^\tau = \begin{cases} \delta_{\tau_i} & \forall i \in S \\ \tilde{p}_i^\tau & \forall i \notin S \end{cases} \qquad \sigma_S \coloneqq \tau$$

$\exists\, S^\star \subseteq [n]$ with $|S^\star| \leq k$,

$$\mathcal{F}_{\mathrm{SA}(k+2;[n]),k} - \left(\mathbb{E}_\nu[f] + H(\nu)\right) \leq \mathcal{O}\left(n\|A\|_F/\sqrt{k}\right) \quad \text{where } \nu = \mathbb{E}_{\tau\sim\tilde{p}_{S^\star}}[\pi^\tau]$$

We postpone the proof to the end, since it builds upon the techniques for proving the NMF error bounds (Theorems 1 and 2).

NMF approximation error for Ising models (Proofs)

Theorem 1 (Jain-Koehler-Risteski '19).

For a symmetric interaction matrix $A \in \mathbb{R}^{n \times n}$, and consider the Ising Gibbs measure

$$\mu(\sigma) \propto e^{f(\sigma)} \quad \text{where} \quad f(\sigma) = \frac{1}{2}\sigma^\top A \sigma$$

Then, $\mathcal{F} - \mathcal{F}_{\mathrm{NMF}} = \mathcal{O}\left(n^{2/3} \|A\|_F^{2/3}\right)$

Theorem 2 (Eldan '20).

For a symmetric interaction matrix $A \in \mathbb{R}^{n \times n}$, and consider the Ising Gibbs measure

$$\mu(\sigma) \propto e^{f(\sigma)} \quad \text{where} \quad f(\sigma) = \frac{1}{2}\sigma^\top A \sigma$$

Then, $\mathcal{F} - \mathcal{F}_{\mathrm{NMF}} \leq 3 \log \det\left(I + L^{1/2} \mathrm{Cov}(\mu) L^{1/2}\right)$, where $L \coloneqq (A^2)^{1/2}$

Measure decomposition

Lemma. Suppose we can decompose $\mu(\sigma) \propto e^{f(\sigma)}$ as a mixture $\mathbb{E}_{\theta\sim\xi}\left[\mu^{(\theta)}\right]$, where ξ is a distribution over some auxiliary state space $\mathcal{I}$, and each component measure $\mu^{(\theta)}$ is again a distribution over $\{\pm 1\}^n$. Assume this decomposition admits the following properties:

- "Low-entropy" mixture:

$$H(\mu) - \mathbb{E}_{\theta\sim\xi}\left[H\left(\mu^{(\theta)}\right)\right] \leq \alpha$$

- "Near-product" components:

$$\mathbb{E}_{\theta\sim\xi}\left[\mathbb{E}_{\mu^{(\theta)}}[f] - \mathbb{E}_{\pi\left(\mu^{(\theta)}\right)}[f]\right] \leq \eta$$

Then $\mathcal{F} - \mathcal{F}_{\mathrm{NMF}} \leq \alpha + \eta$

$\pi\left(\mu^{(\theta)}\right)$ the unique product measure with the same marginals as μ

Proof of the measure decomposition lemma

- According to the Gibbs Variational Principle,

$$\mathcal{F} = \mathbb{E}_{\sigma\sim\mu}[f(\sigma)] + H(\mu) = \mathbb{E}_{\theta\sim\xi}\left[\mathbb{E}_{\sigma\sim\mu^{(\theta)}}[f(\sigma)] + H\left(\mu^{(\theta)}\right)\right] + \left(H(\mu) - \mathbb{E}_{\theta\sim\xi}\left[H\left(\mu^{(\theta)}\right)\right]\right)$$
$$\leq \mathbb{E}_{\theta\sim\xi}\left[\mathbb{E}_{\sigma\sim\mu^{(\theta)}}[f(\sigma)] + H\left(\mu^{(\theta)}\right)\right] + \alpha$$

- According to the Maximum Entropy Principle, $H\left(\mu^{(\theta)}\right) \leq H\left(\pi\left(\mu^{(\theta)}\right)\right)$

- Therefore,

$$\mathcal{F} \leq \mathbb{E}_{\theta\sim\xi}\left[\mathbb{E}_{\sigma\sim\mu^{(\theta)}}[f(\sigma)] + H\left(\pi\left(\mu^{(\theta)}\right)\right)\right] + \alpha$$
$$\leq \mathbb{E}_{\theta\sim\xi}\left[\mathbb{E}_{\sigma\sim\pi\left(\mu^{(\theta)}\right)}[f(\sigma)] + H\left(\pi\left(\mu^{(\theta)}\right)\right)\right] + \alpha + \left(\mathbb{E}_{\mu^{(\theta)}}[f] - \mathbb{E}_{\pi\left(\mu^{(\theta)}\right)}[f]\right)$$
$$\leq \mathcal{F}_{\mathrm{NMF}} + \alpha + \eta$$

$\blacksquare$

Decomposition via Pinning

Pinning Lemma. Let μ be any probability measure over $\{\pm 1\}^n$. Then for every $\ell \in [n]$, there exists $S \subseteq [n]$ with $|S| \leq \ell - 1$ such that

$$\mathbb{E}_{\tau \sim \mu_S}\left[\mathbb{E}_{\{i,j\} \sim \mathrm{Unif}\binom{[n]}{2}}\left[\mathrm{Cov}_{\sigma \sim \mu^\tau}(\sigma_i, \sigma_j)^2\right]\right] \leq \frac{2}{\ell}$$

$$\mathrm{Cov}(X, Y) \coloneqq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Theorem 1 (Jain-Koehler-Risteski '19).

For a symmetric interaction matrix $A \in \mathbb{R}^{n \times n}$, and consider the Ising Gibbs measure

$$\mu(\sigma) \propto e^{f(\sigma)} \quad \text{where} \quad f(\sigma) = \frac{1}{2}\sigma^\top A \sigma$$

Then, $\mathcal{F} - \mathcal{F}_{\mathrm{NMF}} = \mathcal{O}\left(n^{2/3}\|A\|_F^{2/3}\right)$

Decomposition via Pinning

Pinning Lemma. Let μ be any probability measure over $\{\pm 1\}^n$. Then for every $\ell \in [n]$, there exists $S \subseteq [n]$ with $|S| \leq \ell - 1$ such that

$$\mathbb{E}_{\tau \sim \mu_S}\left[\mathbb{E}_{\{i,j\} \sim \mathrm{Unif}\binom{[n]}{2}}\left[\mathrm{Cov}_{\sigma \sim \mu^\tau}(\sigma_i, \sigma_j)^2\right]\right] \leq \frac{2}{\ell}$$

$\mathrm{Cov}(X, Y) \coloneqq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$

Proof of Theorem 1.

- Let $\ell = \mathcal{O}(1/\epsilon^2)$ and apply the Pinning Lemma, which gives a subset S of size $\mathcal{O}(1/\epsilon^2)$
- Let the mixture distribution $\xi \coloneqq \mu_S$
- μ_S is supported on a set of size $2^{|S|}$. Thus, $H(\xi) \leq |S| = \mathcal{O}(1/\epsilon^2)$
- $H(\mu) - \mathbb{E}_{\theta \sim \xi}\left[H\left(\mu^{(\theta)}\right)\right] \leq H(\xi)$ (by the chain rule of conditional entropy)
- Hence, $\alpha = \mathcal{O}(1/\epsilon^2)$

"Entropy-covariance trade-off"

Decomposition via Pinning

Pinning Lemma. Let μ be any probability measure over $\{\pm 1\}^n$. Then for every $\ell \in [n]$, there exists $S \subseteq [n]$ with $|S| \leq \ell - 1$ such that

$$\mathbb{E}_{\tau \sim \mu_S}\left[\mathbb{E}_{\{i,j\} \sim \mathrm{Unif}\binom{[n]}{2}}\left[\mathrm{Cov}_{\sigma \sim \mu^\tau}(\sigma_i, \sigma_j)^2\right]\right] \leq \frac{2}{\ell}$$

$\mathrm{Cov}(X, Y) \coloneqq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$

Proof of Theorem 1.

- Recall that $f(\sigma) = \frac{1}{2}\sigma^\top A \sigma$
- $\mathbb{E}_{\sigma \sim \mu^\tau}[f(\sigma)] = \frac{1}{2}\sum_{ij} A_{ij} \mathbb{E}_{\sigma \sim \mu^\tau}[\sigma_i \sigma_j]$ and $\mathbb{E}_{\sigma \sim \pi(\mu^\tau)}[f(\sigma)] = \frac{1}{2}\sum_{ij} A_{ij} \mathbb{E}_{\sigma \sim \mu^\tau}[\sigma_i] \mathbb{E}_{\sigma \sim \mu^\tau}[\sigma_j]$
- $\mathbb{E}_{\sigma \sim \mu^\tau}[f(\sigma)] - \mathbb{E}_{\sigma \sim \pi(\mu^\tau)}[f(\sigma)] = \frac{1}{2}\sum_{ij} A_{ij} \mathbb{E}_{\sigma \sim \mu^\tau}\left[\mathrm{Cov}_{\sigma \sim \mu^\tau}(\sigma_i \sigma_j)\right] = \frac{1}{2}\mathrm{tr}[A \cdot \mathrm{Cov}(\mu^\tau)]$

Decomposition via Pinning

Pinning Lemma. Let μ be any probability measure over $\{\pm 1\}^n$. Then for every $\ell \in [n]$, there exists $S \subseteq [n]$ with $|S| \leq \ell - 1$ such that

$$\mathbb{E}_{\tau \sim \mu_S}\left[\mathbb{E}_{\{i,j\} \sim \mathrm{Unif}\binom{[n]}{2}}\left[\mathrm{Cov}_{\sigma \sim \mu^\tau}\left(\sigma_i, \sigma_j\right)^2\right]\right] \leq \frac{2}{\ell}$$

$\mathrm{Cov}(X, Y) \coloneqq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$

Proof of Theorem 1.

$$\begin{aligned}
2\mathbb{E}_{\tau \sim \mu_S}\left[\mathbb{E}_{\sigma \sim \mu^\tau}[f(\sigma)] - \mathbb{E}_{\sigma \sim \pi(\mu^\tau)}[f(\sigma)]\right] &= \mathrm{tr}\left[A \cdot \mathbb{E}_{\tau \sim \mu_S}[\mathrm{Cov}(\mu^\tau)]\right] \\
&\leq \|A\|_F \cdot \left\|\mathbb{E}_{\tau \sim \mu_S}[\mathrm{Cov}(\mu^\tau)]\right\|_F \\
&\leq \|A\|_F \cdot \mathbb{E}_{\tau \sim \mu_S}[\|\mathrm{Cov}(\mu^\tau)\|_F^2]^{1/2} \\
&= \mathcal{O}(\epsilon n \|A\|_F)
\end{aligned}$$

- Thus, $\eta = \mathcal{O}(\epsilon n \|A\|_F)$. We have $\mathcal{F} - \mathcal{F}_{NMF} \leq \mathcal{O}(1/\epsilon^2 + \epsilon n \|A\|_F) = \mathcal{O}\left(n^{2/3} \|A\|_F^{2/3}\right)$

$\blacksquare$

Proof of the Pinning Lemma

- Recall that the mutual information $I(X;Y)$ is defined by:
$$I(X;Y) \coloneqq D_{\mathrm{KL}}\big(\mathrm{Law}(X,Y) || \mathrm{Law}(X) \otimes \mathrm{Law}(Y)\big) = H(X) - H(X|Y)$$
- **Fact.** Let X, Y be $\{\pm 1\}$-valued random variables. Then $\mathrm{Cov}(X,Y)^2 \le 2I(X;Y)$
- We'll prove that $\exists S$, $\mathbb{E}_{\{i,j\}\sim\mathrm{Unif}\binom{[n]}{2}}\big[I\big(\sigma_i;\sigma_j\big|\sigma_S\big)\big] \le \frac{1}{\ell}$

$I\big(\sigma_i;\sigma_j\big|\sigma_S\big) = H\big(\sigma_j\big|\sigma_S\big) - H\big(\sigma_j\big|\sigma_{S\cup\{i\}}\big)$

- For any $i_1, \ldots, i_\ell, j \in [n]$,

$$\frac{1}{\ell}\sum_{t=1}^{\ell} I\big(\sigma_{i_t};\sigma_j\big|\sigma_{i_1},\ldots,\sigma_{i_{t-1}}\big) = \frac{1}{\ell}\sum_{t=1}^{\ell}\Big(H\big(\sigma_j\big|\sigma_{i_1},\ldots,\sigma_{i_{t-1}}\big) - H\big(\sigma_j\big|\sigma_{i_1},\ldots,\sigma_{i_{t-1}},\sigma_{i_t}\big)\Big)$$

telescoping sum

$$= \frac{1}{\ell}\Big(H\big(\sigma_j\big) - H\big(\sigma_j\big|\sigma_{i_1},\ldots,\sigma_{i_\ell}\big)\Big) \le \frac{1}{\ell}$$

Proof of the Pinning Lemma

- We'll prove that $\exists S,\ \mathbb{E}_{\{i,j\}\sim\mathrm{Unif}\binom{[n]}{2}}\left[I\left(\sigma_i;\sigma_j \middle| \sigma_S\right)\right] \leq \frac{1}{\ell}$
- For any $i_1, \dots, i_\ell, j \in [n]$,

$$\frac{1}{\ell}\sum_{t=1}^{\ell} I\left(\sigma_{i_t};\sigma_j \middle| \sigma_{i_1},\dots,\sigma_{i_{t-1}}\right) = \frac{1}{\ell}\sum_{t=1}^{\ell}\left(H\left(\sigma_j \middle| \sigma_{i_1},\dots,\sigma_{i_{t-1}}\right) - H\left(\sigma_j \middle| \sigma_{i_1},\dots,\sigma_{i_{t-1}},\sigma_{i_t}\right)\right)$$

$$= \frac{1}{\ell}\left(H\left(\sigma_j\right) - H\left(\sigma_j \middle| \sigma_{i_1},\dots,\sigma_{i_\ell}\right)\right) \leq \frac{1}{\ell}$$

- Averaging over $i_1, \dots, i_\ell, j$, we get that

$$\frac{1}{\ell}\sum_{t=1}^{\ell} \mathbb{E}_{i_1,\dots,i_{t-1}\sim[n]}\left[\mathbb{E}_{i_t,j\sim[n]}\left[I\left(\sigma_{i_t};\sigma_j \middle| \sigma_{i_1},\dots,\sigma_{i_{t-1}}\right)\right]\right] \leq \frac{1}{\ell}$$

- Therefore, there must be an $S = \{i_1, \dots, i_{t-1}\}$ for some $t \leq \ell$ that satisfies the condition

$\blacksquare$

SA approximation error

Theorem 3 (Risteski ’16).

For a symmetric interaction matrix $A \in \mathbb{R}^{n\times n}$, and consider the Ising Gibbs measure

$$\mu(\sigma) \propto e^{f(\sigma)} \quad \text{where} \quad f(\sigma) = \frac{1}{2}\sigma^\top A\sigma$$

For $0 \le k \le n-2$,

$$0 \le \mathcal{F}_{\mathrm{SA}(k+2;[n]),k} - \mathcal{F} \le \mathcal{O}\big(n\|A\|_F/\sqrt{k}\big)$$

Moreover, if $\tilde{\boldsymbol{p}}$ is the optimal pseudo-distribution, then we can round it into a product measure π satisfying

$$\mathcal{F} - \big(\mathbb{E}_\pi[f] + H(\pi)\big) \le \mathcal{O}\big(n\|A\|_F/\sqrt{k} + k\big)$$

Rounding the pseudo-distribution

- Let $\tilde{p}$ be the optimal pseudo-distribution. Fix $S \subseteq [n]$ with $|S| \leq k$
- Define a mixture of distributions:
  - Sample $\tau \sim \tilde{p}_S$
  - Sample $\sigma \in \{\pm 1\}^n$ according to a product measure π^τ defined by:

$$\pi_i^\tau = \begin{cases} \delta_{\tau_i} & \forall i \in S \\ \tilde{p}_i^\tau & \forall i \notin S \end{cases} \qquad \sigma_S \coloneqq \tau$$

$\exists\, S^\star \subseteq [n]$ with $|S^\star| \leq k$,

$$\mathcal{F}_{\mathrm{SA}(k+2;[n]),k} - \big(\mathbb{E}_\nu[f] + H(\nu)\big) \leq \mathcal{O}\big(n\|A\|_F/\sqrt{k}\big) \quad \text{where } \nu = \mathbb{E}_{\tau\sim\tilde{p}_{S^\star}}[\pi^\tau]$$

Proof of SA approximation error

- The Pinning Lemma also works for $\mathfrak{F}_{k+2}$-pseudo-distributions when pinning up to k coordinates
- There exists $S^\star \subseteq [n]$ with $|S^\star| \leq k$ such that

$$\mathbb{E}_{\tau \sim \widetilde{\boldsymbol{p}}_{S^\star}}\left[\mathbb{E}_{\{i,j\} \sim \mathrm{Unif}\binom{[n]}{2}}\left[\widetilde{\mathrm{Cov}}_{\sigma \sim \widetilde{\boldsymbol{p}}^\tau}(\sigma_i, \sigma_j)^2\right]\right] \leq \frac{2}{k}$$

where $\widetilde{\mathrm{Cov}}_{\sigma \sim \widetilde{\boldsymbol{p}}^\tau}(\sigma_i, \sigma_j) \coloneqq \widetilde{\mathbb{E}}_{\widetilde{\boldsymbol{p}}_{ij}^\tau}[\sigma_i \sigma_j] - \widetilde{\mathbb{E}}_{\widetilde{\boldsymbol{p}}_i^\tau}[\sigma_i] \cdot \widetilde{\mathbb{E}}_{\widetilde{\boldsymbol{p}}_j^\tau}[\sigma_j]$ is the pseudo-covariance

- Using the same argument in the proof of Theorem 1, we get that

$$\widetilde{\mathbb{E}}[f] - \mathbb{E}_\nu[f] \leq \mathcal{O}\left(n\|A\|_F/\sqrt{k}\right)$$

- By the definitio
- Combining them

Proof of Theorem 1.

$$\begin{aligned} 2 \cdot \mathbb{E}_{\tau \sim \mu_S}\left[\mathbb{E}_{\sigma \sim \mu^\tau}[f(\sigma)] - \mathbb{E}_{\sigma \sim \pi(\mu^\tau)}[f(\sigma)]\right] &= \mathrm{tr}\left[A \cdot \mathbb{E}_{\tau \sim \mu_S}[\mathrm{Cov}(\mu^\tau)]\right] \\ &\leq \|A\|_F \cdot \left\|\mathbb{E}_{\tau \sim \mu_S}[\mathrm{Cov}(\mu^\tau)]\right\|_F \\ &\leq \|A\|_F \cdot \mathbb{E}_{\tau \sim \mu_S}[\|\mathrm{Cov}(\mu^\tau)\|_F^2]^{1/2} \\ &= \mathcal{O}(\epsilon n \|A\|_F) \end{aligned}$$

$$\widetilde{H}_j(\widetilde{\boldsymbol{p}}) \coloneqq \min_{|S| \leq j}\left\{H(\widetilde{\boldsymbol{p}}_S) + \sum_{i \notin S} H(\widetilde{\boldsymbol{p}}_i | \widetilde{\boldsymbol{p}}_S)\right\}$$

$\blacksquare$

Sherali-Adams vs. Sum-of-Squares

Counterexample

- $n = 3$ and $\mathfrak{F} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}\}$
- $\tilde{p}_i[i = \pm 1] = 1/2$
- $\tilde{p}_{ij}[i = 1, j = -1] = \tilde{p}_{ij}[i = -1, j = 1] = 1/2$

- Level-2 Sherali-Adams cannot refute it
- Degree-2 SoS can refute it:

$$\mathcal{M}_2 \coloneqq \begin{array}{c c} & \begin{array}{cccc} 1 & x_1 & x_2 & x_3 \end{array} \\ \begin{array}{c} 1 \\ x_1 \\ x_2 \\ x_3 \end{array} & \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & -1 \\ 0 & -1 & 1 & -1 \\ 0 & -1 & -1 & 1 \end{pmatrix} \end{array} \not\succcurlyeq 0$$

Eigenvalues: $2, 2, 1, -1$

Sherali-Adams vs. Sum-of-Squares

- Level-k Sherali-Adams
  - $n^{\mathcal{O}(k)}$ linear constraints

- Degree-k Sum-of-Squares
  - $\mathcal{M}_k \succcurlyeq 0 \quad \Longleftrightarrow \quad u^\top \mathcal{M}_k u \geq 0 \quad \forall u \in \mathbb{R}^{n^{\mathcal{O}(k)}}$

infinitely many linear constraints

Proof of Theorem 2

Theorem 2 (Eldan '20).

For a symmetric interaction matrix $A \in \mathbb{R}^{n \times n}$, and consider the Ising Gibbs measure

$$\mu(\sigma) \propto e^{f(\sigma)} \quad \text{where} \quad f(\sigma) = \frac{1}{2}\sigma^\top A \sigma$$

Then, $\mathcal{F} - \mathcal{F}_{\mathrm{NMF}} \leq 3 \log\det\left(I + L^{1/2}\mathrm{Cov}(\mu)L^{1/2}\right)$, where $L \coloneqq (A^2)^{1/2}$

- Technical tool: stochastic localization (SL)

Refined Decompositions via SL

Theorem (Eldan '20).

Let μ be any probability measure over $\{\pm 1\}^n$. Then for every symmetric positive definite matrix $L \succ 0$, there exists a decomposition of $\mu = \mathbb{E}_{\theta\sim\xi}\left[\mu^{(\theta)}\right]$ enjoying the following properties:

- $H(\mu) - \mathbb{E}_{\theta\sim\xi}\left[H\left(\mu^{(\theta)}\right)\right] \leq \log\det\left(I + L^{1/2}\mathrm{Cov}(\mu)L^{1/2}\right)$
- $\mathbb{E}_{\theta\sim\xi}\left[\mathrm{Cov}\left(\mu^{(\theta)}\right)\right] \preccurlyeq L^{-1}$
- $\mathbb{E}_{\theta\sim\xi}\left[\mathrm{Cov}\left(\mu^{(\theta)}\right)L\mathrm{Cov}\left(\mu^{(\theta)}\right)\right] \preccurlyeq \mathrm{Cov}(\mu)$

Proof of Theorem 2

We need to check the two conditions in the measure decomposition lemma:

"Low-entropy" mixture:

$$H(\mu) - \mathbb{E}_{\theta\sim\xi}\left[H\left(\mu^{(\theta)}\right)\right] \leq \alpha$$

- $\alpha = \log\det\left(I + L^{1/2}\mathrm{Cov}(\mu)L^{1/2}\right)$

"Near-product" components:

$$\mathbb{E}_{\theta\sim\xi}\left[\mathbb{E}_{\mu^{(\theta)}}[f] - \mathbb{E}_{\pi\left(\mu^{(\theta)}\right)}[f]\right] \leq \eta$$

- Following the proof of Theorem 1,

$$\mathbb{E}_{\theta\sim\xi}\left[\mathbb{E}_{\mu^{(\theta)}}[f] - \mathbb{E}_{\pi\left(\mu^{(\theta)}\right)}[f]\right] = \frac{1}{2}\mathrm{tr}\left[A\cdot\mathbb{E}_{\theta\sim\xi}\left[\mathrm{Cov}\left(\mu^{(\theta)}\right)\right]\right] \leq \frac{1}{2}\mathrm{tr}\left[\mathbb{E}_{\theta\sim\xi}\left[L^{1/2}\mathrm{Cov}\left(\mu^{(\theta)}\right)L^{1/2}\right]\right]$$

$(L \succcurlyeq A)$

- $\mathbb{E}_{\theta\sim\xi}\left[L^{1/2}\mathrm{Cov}\left(\mu^{(\theta)}\right)L^{1/2}\right] \preccurlyeq I$ (by Eldan's decomposition)
- $\mathbb{E}_{\theta\sim\xi}\left[\mathrm{Cov}\left(\mu^{(\theta)}\right)\right] \preccurlyeq \mathrm{Cov}(\mu)$ (by the Law of Total Covariance)

Proof of Theorem 2

We need to check the two conditions in the measure decomposition lemma:

"Low-entropy" mixture:

$$H(\mu) - \mathbb{E}_{\theta\sim\xi}\left[H\left(\mu^{(\theta)}\right)\right] \le \alpha$$

- $\alpha = \log\det\left(I + L^{1/2}\mathrm{Cov}(\mu)L^{1/2}\right)$

"Near-product" components:

$$\mathbb{E}_{\theta\sim\xi}\left[\mathbb{E}_{\mu^{(\theta)}}[f] - \mathbb{E}_{\pi\left(\mu^{(\theta)}\right)}[f]\right] \le \eta$$

- Following the proof of Theorem 1,

$$\mathbb{E}_{\theta\sim\xi}\left[\mathbb{E}_{\mu^{(\theta)}}[f] - \mathbb{E}_{\pi\left(\mu^{(\theta)}\right)}[f]\right] = \frac{1}{2}\mathrm{tr}\left[A\cdot\mathbb{E}_{\theta\sim\xi}\left[\mathrm{Cov}\left(\mu^{(\theta)}\right)\right]\right] \le \frac{1}{2}\mathrm{tr}\left[\mathbb{E}_{\theta\sim\xi}\left[L^{1/2}\mathrm{Cov}\left(\mu^{(\theta)}\right)L^{1/2}\right]\right]$$

- $\lambda_i\left(\mathbb{E}_{\theta\sim\xi}\left[L^{1/2}\mathrm{Cov}\left(\mu^{(\theta)}\right)L^{1/2}\right]\right) \le \min\left\{1, \lambda_i\left(L^{1/2}\mathrm{Cov}(\mu)L^{1/2}\right)\right\} \le 2\log\left(1 + \lambda_i\left(L^{1/2}\mathrm{Cov}(\mu)L^{1/2}\right)\right)$

$(\mathrm{Cov}(\mu) \succcurlyeq 0)$

$\blacksquare$

Proof of Eldan's Decomposition Theorem

We only prove the first two properties following the presentation in (Alaoui-Montanari '22)

Gaussian channel localization

$$\sigma \sim \mu \longrightarrow \boxed{\text{Gaussian channel}} \longrightarrow \theta = \sigma + L^{-1/2} g$$

$$g \sim \mathcal{N}(0, I)$$

- $\mu^{(\theta)} := \mathrm{Law}(\sigma \mid \theta)$ and $\xi(\theta) \propto \mathbb{E}_{\sigma\sim\mu}\left[\mathbb{E}_{g\sim\mathcal{N}(0,I)}\left[\mathbf{1}_{\theta=\sigma+L^{-1/2}g}\right]\right]$
- For the first property,

$$H(\mu) - \mathbb{E}_{\theta\sim\xi}\left[H\left(\mu^{(\theta)}\right)\right] = H(\sigma) - H(\sigma \mid \theta) = I(\sigma;\theta) = H(\theta) - H(\theta \mid \sigma)$$

- For $H(\theta)$, by another version of Maximum Entropy Principle,

$$H(\theta) \le H\left(\mathcal{N}\left(0, \mathrm{Cov}(\xi)\right)\right) = \frac{n}{2}\log(2\pi e) + \frac{1}{2}\mathrm{tr}[\log \mathrm{Cov}(\xi)] = \frac{n}{2}\log(2\pi e) + \frac{1}{2}\mathrm{tr}\left[\log\left(L^{-1} + \mathrm{Cov}(\mu)\right)\right]$$

Proof of Eldan's Decomposition Theorem

We only prove the first two properties following the presentation in (Alaoui-Montanari '22)

Gaussian channel localization

$$\sigma \sim \mu \longrightarrow \boxed{\text{Gaussian channel}} \longrightarrow \theta = \sigma + L^{-1/2} g$$

$$g \sim \mathcal{N}(0, I)$$

- $\mu^{(\theta)} := \mathrm{Law}(\sigma \mid \theta)$ and $\xi(\theta) \propto \mathbb{E}_{\sigma\sim\mu}\left[\mathbb{E}_{g\sim\mathcal{N}(0,I)}\left[\mathbf{1}_{\theta=\sigma+L^{-1/2}g}\right]\right]$
- For $H(\theta \mid \sigma)$,

$$H(\theta \mid \sigma) = H\left(L^{-1/2} g\right) = \frac{n}{2}\log(2\pi e) + \frac{1}{2}\mathrm{tr}[\log L^{-1}]$$

- Hence, $I(\sigma; \theta) \leq \frac{1}{2}\mathrm{tr}\left[\log\left(L^{-1} + \mathrm{Cov}(\mu)\right)\right] - \frac{1}{2}\mathrm{tr}[\log L^{-1}] \leq \frac{1}{2}\log\det\left(I + L^{1/2}\mathrm{Cov}(\mu)L^{1/2}\right)$

Proof of Eldan's Decomposition Theorem

We only prove the first two properties following the presentation in (Alaoui-Montanari '22)

Gaussian channel localization

$$\sigma \sim \mu \longrightarrow \boxed{\text{Gaussian channel}} \longrightarrow \theta = \sigma + L^{-1/2} g$$

$$g \sim \mathcal{N}(0, I)$$

- $\mu^{(\theta)} \coloneqq \mathrm{Law}(\sigma \mid \theta)$ and $\xi(\theta) \propto \mathbb{E}_{\sigma \sim \mu}\left[\mathbb{E}_{g \sim \mathcal{N}(0,I)}\left[\mathbf{1}_{\theta = \sigma + L^{-1/2} g}\right]\right]$
- For the second property, our goal is to show that

$$\mathbb{E}_{\theta \sim \xi}\left[\mathrm{Cov}\left(\mu^{(\theta)}\right)\right] \preccurlyeq L^{-1} = \mathrm{Cov}\left(-L^{-1/2} g\right) = \mathrm{Cov}(\sigma - \theta)$$

Proof of Eldan's Decomposition Theorem

We only prove the first two properties following the presentation in (Alaoui-Montanari '22)

Gaussian channel localization

$$\sigma \sim \mu \longrightarrow \boxed{\text{Gaussian channel}} \longrightarrow \theta = \sigma + L^{-1/2} g$$

$$g \sim \mathcal{N}(0, I)$$

- $\mu^{(\theta)} \coloneqq \mathrm{Law}(\sigma \mid \theta)$ and $\xi(\theta) \propto \mathbb{E}_{\sigma\sim\mu}\left[\mathbb{E}_{g\sim\mathcal{N}(0,I)}\left[\mathbf{1}_{\theta=\sigma+L^{-1/2}g}\right]\right]$
- For the second property, our goal is to show that
$$\mathrm{tr}\big[\mathbb{E}_{\theta\sim\xi}\big[\mathrm{Cov}\big(\mu^{(\theta)}\big)\big]\cdot B\big] \le \mathrm{tr}[\mathrm{Cov}(\sigma-\theta)\cdot B] \qquad \forall\, B \succcurlyeq 0$$
which is further equivalent to
$$\mathbb{E}_{\theta,\sigma}[(\sigma - \mathbb{E}[\sigma \mid \theta])^\top B(\sigma - \mathbb{E}[\sigma \mid \theta])] \le \mathbb{E}_{\theta,\sigma}[(\sigma-\theta)^\top B(\sigma-\theta)]$$

Proof of Eldan's Decomposition Theorem

We only prove the first two properties following the presentation in (Alaoui-Montanari '22)

Gaussian channel localization

$$\sigma \sim \mu \longrightarrow \boxed{\text{Gaussian channel}} \longrightarrow \theta = \sigma + L^{-1/2} g$$

$$g \sim \mathcal{N}(0, I)$$

- For the second property,

$$\mathbb{E}_{\theta,\sigma}[(\sigma - \mathbb{E}[\sigma \mid \theta])^\top B(\sigma - \mathbb{E}[\sigma \mid \theta])] \leq \mathbb{E}_{\theta,\sigma}[(\sigma - \theta)^\top B(\sigma - \theta)]$$

- Given $\theta = \sigma + L^{-1/2} g$, how to estimate σ?
  - **Maximum likelihood estimator:** $\hat{\sigma} = \theta$
  - **Bayes estimator:** $\hat{\sigma}_{\text{Bayes}} = \mathbb{E}[\sigma \mid \theta]$

Fact. Bayes estimator is optimal under mean-squared error

∎

Bayesian estimation theory

- The loss function is the mean-squared error weighted by B:

$$\mathbb{E}_{\theta,\sigma}[(\sigma - \hat{\sigma})^\top B(\sigma - \hat{\sigma})] = \mathbb{E}_{\theta,\sigma}[\|\sigma - \hat{\sigma}\|_B^2]$$

- For any estimator $\hat{\sigma}(\theta)$,

$$\begin{aligned} r_\theta(\hat{\sigma}) &:= \mathbb{E}_{\sigma\mid\theta}[\|\sigma - \hat{\sigma}\|_B^2] = \mathbb{E}_{\sigma\mid\theta}\left[\left\|\sigma - \hat{\sigma}_{\text{Bayes}} + \hat{\sigma}_{\text{Bayes}} - \hat{\sigma}\right\|_B^2\right] \\ &= \mathbb{E}_{\sigma\mid\theta}\left[\left\|\sigma - \hat{\sigma}_{\text{Bayes}}\right\|_B^2\right] + \left\|\hat{\sigma}_{\text{Bayes}} - \hat{\sigma}\right\|_B^2 \qquad (\mathbb{E}_{\sigma\mid\theta}[\sigma - \hat{\sigma}_{\text{Bayes}}] = 0) \\ &\geq r_\theta(\hat{\sigma}_{\text{Bayes}}) \end{aligned}$$

- Therefore,

$$\mathbb{E}_{\theta,\sigma}\left[\left\|\sigma - \hat{\sigma}_{\text{Bayes}}\right\|_B^2\right] \leq \mathbb{E}_{\theta,\sigma}[\|\sigma - \hat{\sigma}\|_B^2]$$

for any estimaror $\hat{\sigma}$

$\blacksquare$

Corollary of Theorem 2

Corollary.

For a symmetric interaction matrix $A \in \mathbb{R}^{n\times n}$, and consider the Ising Gibbs measure

$$\mu(\sigma) \propto e^{f(\sigma)} \quad \text{where} \quad f(\sigma) = \frac{1}{2}\sigma^\top A\sigma$$

Then, $\mathcal{F} - \mathcal{F}_{\mathrm{NMF}} \leq 3 \cdot \mathrm{rank}(A) \cdot \log(\|A\|n + 1)$

Example 2:

- Consider $A = \frac{\beta}{n}\mathbf{1}\mathbf{1}^\top$
- $\mathrm{rank}(A) = 1$ and $\|A\| = \beta$
- According to the corollary, $\mathcal{F} - \mathcal{F}_{\mathrm{NMF}} \leq 3\log(n\beta + 1)$

Corollary of Theorem 2

Corollary.

For a symmetric interaction matrix $A \in \mathbb{R}^{n\times n}$, and consider the Ising Gibbs measure

$$\mu(\sigma) \propto e^{f(\sigma)} \quad \text{where} \quad f(\sigma) = \frac{1}{2}\sigma^\top A\sigma$$

Then, $\mathcal{F} - \mathcal{F}_{\mathrm{NMF}} \leq 3 \cdot \mathrm{rank}(A) \cdot \log(\|A\|n + 1)$

Proof.

$$\begin{aligned}\log\det\left(I + L^{1/2}\mathrm{Cov}(\mu)L^{1/2}\right) &= \sum_{i\in[n]} \log\left(\lambda_i\left(L^{1/2}\mathrm{Cov}(\mu)L^{1/2}\right) + 1\right) \\ &\leq \mathrm{rank}(A) \cdot \log\left(\left\|L^{1/2}\mathrm{Cov}(\mu)L^{1/2}\right\| + 1\right) \\ &\leq \mathrm{rank}(A) \cdot \log(\|A\| \cdot \|\mathrm{Cov}(\mu)\| + 1)\end{aligned}$$

Corollary of Theorem 2

Corollary.

For a symmetric interaction matrix $A \in \mathbb{R}^{n\times n}$, and consider the Ising Gibbs measure

$$\mu(\sigma) \propto e^{f(\sigma)} \quad \text{where} \quad f(\sigma) = \frac{1}{2}\sigma^\top A\sigma$$

Then, $\mathcal{F} - \mathcal{F}_{\mathrm{NMF}} \leq 3 \cdot \mathrm{rank}(A) \cdot \log(\|A\|n + 1)$

Proof.

$$\log\det\left(I + L^{1/2}\mathrm{Cov}(\mu)L^{1/2}\right) \leq \mathrm{rank}(A) \cdot \log(\|A\| \cdot \|\mathrm{Cov}(\mu)\| + 1)$$

- $\|\mathrm{Cov}(\mu)\| \leq \mathrm{tr}[\mathrm{Cov}(\mu)] \leq \sum_{\sigma\in\{\pm1\}^n} \|\sigma\|^2 \mu(\sigma) \leq n$

$\blacksquare$