Lecture 22: Introduction to Variational & Entropy-Based Methods, Naïve Mean-Field Approximation

November 30, 2023

In the remaining few lectures, we turn to methods for approximate counting based on optimization and entropy. The main workhorse behind all of this is the following variational characterization of the (log-)partition function. For convenience, we specialize the statement to finite state spaces Ω endowed with the uniform measure (although this is not necessary); a far more general statement, and its proof, were already provided in the previous lecture on entropic independence. Recall that $H(\mu) \stackrel{\mathsf{def}}{=} -\sum_{x\in\Omega} \mu(x) \log \mu(x)$ denotes the *Shannon entropy* of μ.

Theorem 0.1 (Specialization of Gibbs Variational Principle)**.** *Let Ω be a finite state space. Then the function $\mu \mapsto -H(\mu)$ on probability measures over Ω is smooth and strictly convex. Furthermore, for every function $f : \Omega \to \mathbb{R}$,*

$$\log \sum_{x\in\Omega} e^{f(x)} = \sup_{\nu} \left\{ \mathbb{E}_{x\sim\nu} \left[f(x) \right] + H(\nu) \right\}, \tag{1}$$

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

The convex program Eq. (1) shows how one can in principle compute the (log-)partition function via optimization. The glaring issue, of course, is that the number of variables is $|\Omega|$, which is often exponentially large.

Theme 0.2. *Find a "tractable"* relaxation/restriction *of the program Eq. (1), and solve the new (ideally convex) program to approximate the (log-)partition function.*

The main challenge with this approach is finding an efficiently computable relaxation/restriction for which we can provide two-sided guarantees on the approximation error. For the remainder of this lecture, we focus on the product space $\Omega = \{\pm 1\}^n$ for simplicity.

1 The Naïve Mean-Field Approximation

One way to produce a "tractable" optimization problem from Eq. (1) is to restrict the class of probability measures ν. The *(naïve) mean-field approximation* does exactly this. More specifically, the approximation restricts ν in Eq. (1) to be a *product measure* over $\{\pm 1\}^n$:

$$\begin{aligned} \mathcal{F}_{\mathsf{NMF}} &\stackrel{\mathsf{def}}{=} \sup_{\nu \text{ product}} \left\{ \mathbb{E}_{\sigma\sim\nu} \left[f(\sigma) \right] + H(\nu) \right\} \\ &= \sup_{\boldsymbol{m}\in[-1,1]^n} \left\{ \mathbb{E}_{\sigma\sim\pi(\boldsymbol{m})} [f(\sigma)] + H(\pi(\boldsymbol{m})) \right\}, \end{aligned} \tag{2}$$

where $\pi(\boldsymbol{m})$ is the unique product measure on $\{\pm 1\}^n$ with $\boldsymbol{m}$ as its mean vector. Note that by independence of the coordinates, $H(\pi(\boldsymbol{m})) = \sum_{i=1}^n \left(\frac{1+\boldsymbol{m}_i}{2} \log \frac{1+\boldsymbol{m}_i}{2} + \frac{1-\boldsymbol{m}_i}{2} \log \frac{1-\boldsymbol{m}_i}{2} \right)$. For a given $\boldsymbol{m} \in [-1,1]^n$, the expectation $\mathbb{E}_{\sigma\sim\pi(\boldsymbol{m})}[f(\sigma)]$ is also easy to compute for many natural functions f (e.g. quadratic forms, which correspond to Ising model partition functions) just from independence of the coordinates. Hence, at least from the perspective of representation size, Eq. (2) is tractable to write down.

We say a (sequence of) probability measures $\mu(\sigma) \propto e^{f(\sigma)}$ on $\{\pm 1\}^n$ (with $n \to \infty$) exhibits "mean-field behavior" if

$$\frac{\mathcal{F} - \mathcal{F}_{\mathsf{NMF}}}{n} \leq o(1), \tag{3}$$

where $\mathcal{F} \stackrel{\text{def}}{=} \log \sum_{\sigma \in \{\pm 1\}^n} e^{f(\sigma)}$ is the log-partition function, otherwise known as the *free energy*. Note that $\mathcal{F}_{\mathsf{NMF}} \leq \mathcal{F}$ trivially. This $o(n)$-additive approximation to $\mathcal{F}$ is equivalent to $e^{o(n)}$-multiplicative approximation to the true partition function itself. On the one hand, this is a much weaker notion of approximation than an FPRAS. On the other hand, we will see this phenomenon occurs in many models far beyond the regime in which local Markov chains (e.g. Glauber dynamics) mix rapidly.

One of the main uses for this method is the fact that Eq. (2) at least has some chance of admitting a closed-form formula in the large-n limit, which can then be used to obtain predictions on the behavior of μ e.g. phase transitions. Indeed, the notion of "phase transition" doesn't make sense for fixed finite n, since everything becomes continuous/differentiable/smooth in the parameters of the model (e.g. β). These are only revealed by looking at a quantity like the *asymptotic free energy density* $\lim_{n\to\infty} \frac{1}{n}\mathcal{F}$. Eq. (3) implies that this is also just $\lim_{n\to\infty} \frac{1}{n}\mathcal{F}_{\mathsf{NMF}}$, which is more amenable.

Note, however, that convexity is lost in Eq. (2) because the convex combination of two product measures is no longer a product measure. Nonetheless, fixed-point iteration (derived from first-order optimality conditions) and gradient descent are natural methods for solving this in practice. Provable guarantees for these are often confined to the high-temperature setting (e.g. when Dobrushin's condition is satisfied); see [Koe19] for results going beyond this. We discuss issues of computing the optimum of Eq. (2), or approximations of a similar quality to $\mathcal{F}_{\mathsf{NMF}}$, in a future lecture. For the moment, we only focus on the problem of bounding the approximation error, which in itself already has significant consequences for the structure of the measure μ; see e.g. [BM17]. For convenience, we focus on Ising-type models in this lecture.

Theorem 1.1 ([JKR19]; building on [Ris16])**.** *Fix a symmetric interaction matrix $A \in \mathbb{R}^{n\times n}$, and consider the Ising Gibbs measure $\mu(\sigma) \propto e^{f(\sigma)}$ where $f(\sigma) = \frac{1}{2}\sigma^\top A\sigma$. Then*

$$\mathcal{F} - \mathcal{F}_{\mathsf{NMF}} \leq O\left(n^{2/3}\,\|A\|_F^{2/3}\right).$$

Further extensions of this result to systems with higher-order interactions (i.e. hypergraphs) are available in [JKR19]. We also establish a more flexible bound, due to Eldan [Eld20].

Theorem 1.2 ([Eld20])**.** *Fix a symmetric interaction matrix $A \in \mathbb{R}^{n\times n}$, and consider the Ising Gibbs measure $\mu(\sigma) \propto e^{f(\sigma)}$ where $f(\sigma) = \frac{1}{2}\sigma^\top A\sigma$. Then*

$$\mathcal{F} - \mathcal{F}_{\mathsf{NMF}} \leq 3 \log\det\left(\mathsf{Id} + L^{1/2}\,\mathrm{Cov}(\mu)L^{1/2}\right),$$

where $L = \left(A^2\right)^{1/2}$.

One can show that Theorem 1.2 recovers Theorem 1.1. The point is the right-hand side has a more favorable dependence on the eigenvalues (e.g. if they decay in absolute value). For details of the calculations, see [Eld20]. We prove both results to illustrate how two distinct but related techniques, both based on measure decompositions/localization schemes, can be used to study the naïve mean-field approximation. Before we do this, let us give an example application of Theorems 1.1 and 1.2.

Example 1 (Ferromagnetic Ising Model on a d-Regular Graph). Let $d \in \mathbb{N}$, and let $A = \frac{\beta}{d}\mathsf{Adj}_G$, where G is an arbitrary d-regular graph. Note the $\frac{1}{d}$ normalization is natural so that the phase transition in β between fast vs. slow mixing of Glauber dynamics is of constant order. We have $\|A\|_F^{2/3} = \left(\frac{\beta}{d}\right)^{2/3} \cdot (dn)^{1/3} = \left(\frac{\beta^2 n}{d}\right)^{1/3}$. Applying Theorem 1.1, we see that the error in the naïve mean-field approximation is at most $\beta^{2/3} n \cdot d^{-1/3}$. Hence, the ferromagnetic Ising model on G exhibits mean-field phenomena if $d \geq \omega_n(1)$, even if β is a large constant far exceeding the phase transition threshold. In other words, we typically expect the mean-field approximation to work better on *dense* problem instances rather than sparse ones.

Using Theorem 1.2, one can obtain much better bounds on the approximation error, e.g. if one assumes the graph is a strong expander and its eigenvalues decay. For instance, in the extreme case where $A = \frac{\beta}{n}\mathbf{1}\mathbf{1}^\top$, corresponding to $G = K_n$ (the *Curie–Weiss model*), Theorem 1.2 gives a $O(\log n)$ bound on the error for all constant β, as opposed to $O(n^{1-\epsilon})$.

2 Error Bounds via Measure Decompositions

Building off of our previous discussion of localization schemes, the first main result of this lecture is that a naïve mean-field approximation holds if one can find a decomposition such that the mixture measure has "low" entropy, and each component measure is "close" to a product measure. Following our $\pi(\boldsymbol{m})$ notation earlier, if μ is any probability measure on $\{\pm 1\}^n$, then we write $\pi(\mu)$ for the unique product measure on $\{\pm 1\}^n$ with the same marginals as that of μ. This insight was elucidated in a number of works [BC16; Coj+18; JKR19].

Theorem 2.1. *Let $f : \{\pm 1\}^n \to \mathbb{R}$ be some Hamiltonian, and 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:***

$$H(\mu) - \mathbb{E}_{\theta\sim\xi}\left[H\left(\mu^{(\theta)}\right)\right] \le \alpha \tag{4}$$

  for some $\alpha > 0$ (possibly depending on n).[1]

- ***"Near-Product" Components (on Average):***

$$\mathbb{E}_{\theta\sim\xi}\left[\mathbb{E}_{\mu^{(\theta)}}[f] - \mathbb{E}_{\pi\left(\mu^{(\theta)}\right)}[f]\right] \le \eta \tag{5}$$

  for some $\eta > 0$ (possibly depending on n).

Then $\mathcal{F} - \mathcal{F}_{\mathsf{NMF}} \le \alpha + \eta$.

Remark 1. It is perhaps interesting to compare this result with what is typically required for mixing of local Markov chains. For instance, in the spectral/entropic independence framework, we require that *all* component measures (i.e. conditionals) to be "close" to a product measure, rather than "on average". Furthermore, "closeness" is measured by the magnitude of the covariance matrix, instead of the deviation in expectation of a single specific test function. As in the third problem set, we also typically require the mixture measure to satisfy some kind of mixing condition, which is technically incomparable with the "low entropy" criterion in Eq. (4), but seems more stringent at a conceptual level.

Proof. Rewriting $\mathcal{F}$ and decomposing the terms, we have

$$\begin{aligned}
\mathcal{F} &= \mathbb{E}_{\sigma\sim\mu}[f(\sigma)] + H(\mu) && \text{(Optimality of } \mu \text{ in Eq. (1))}\\
&= \mathbb{E}_{\theta\sim\xi}\left[\mathbb{E}_{\sigma\sim\mu^{(\theta)}}[f(\sigma)] + H\left(\mu^{(\theta)}\right)\right] + \left(H(\mu) - \mathbb{E}_{\theta\sim\xi}\left[H\left(\mu^{(\theta)}\right)\right]\right)\\
&\le \mathbb{E}_{\theta\sim\xi}\left[\mathbb{E}_{\sigma\sim\mu^{(\theta)}}[f(\sigma)] + H\left(\mu^{(\theta)}\right)\right] + \alpha. && \text{(Using Eq. (4))}
\end{aligned}$$

Now let us examine the terms $\mathbb{E}_{\sigma\sim\mu^{(\theta)}}[f(\sigma)] + H\left(\mu^{(\theta)}\right)$ for each $\theta \in \mathrm{supp}(\xi)$. Observe that since product measures maximize entropy for a prescribed marginal vector[2], we have $H\left(\mu^{(\theta)}\right) \le H\left(\pi\left(\mu^{(\theta)}\right)\right)$. Combining this with Eq. (5), we obtain

$$\begin{aligned}
\mathbb{E}_{\theta\sim\xi}\left[\mathbb{E}_{\sigma\sim\mu^{(\theta)}}[f(\sigma)] + H\left(\mu^{(\theta)}\right)\right] &\le \mathbb{E}_{\theta\sim\xi}\left[\mathbb{E}_{\sigma\sim\pi\left(\mu^{(\theta)}\right)}[f(\sigma)] + H\left(\pi\left(\mu^{(\theta)}\right)\right)\right] + \eta\\
&&\text{(Using Eq. (5))}\\
&\le \mathcal{F}_{\mathsf{NMF}} + \eta. && \text{(Definition of } \mathcal{F}_{\mathsf{NMF}})
\end{aligned}$$

Putting these inequalities together yields the claim. □

[1] This condition is actually weaker than the more natural condition $H(\xi) \le \alpha$.

[2] This is the "Maximum Entropy Principle" we saw in the previous lecture on entropic independence.

2.1 Decomposition via Pinning

In light of Theorem 2.1, our goal is now to find such a "good" measure decomposition. One way to find such a decomposition is to randomly pin a subset of coordinates. This is essentially the coordinate-by-coordinate localization scheme.

Lemma 2.2 (The "Pinning Lemma"; [Mon08; RT12; MR17])**.** *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*[3]

$$\mathbb{E}_{\tau \sim \mu_S} \left[\mathbb{E}_{\{i,j\} \sim \mathsf{Unif}\binom{[n]}{2}} \left[\mathrm{Cov}_{\sigma \sim \mu^\tau} (\sigma_i, \sigma_j)^2 \right] \right] \leq \frac{2 \log 2}{\ell}. \tag{6}$$

Remark 2. We note that this is also a heavily used lemma in the context of rounding semidefinite programming hierarchies; see [BRS11; RT12; MR17] and references therein.

Let us first combine the "Pinning Lemma" with Theorem 2.1 to complete the proof of Theorem 1.1.

Proof of Theorem 1.1. Let $\epsilon > 0$ be a parameter to be determined later. Applying Lemma 2.2 with $\ell = O(1/\epsilon^2)$, we obtain a set $S \subseteq [n]$ of size $O(1/\epsilon^2)$ such that the average squared covariance between a random pair of coordinates under a random pinning $\tau \sim \mu_S$ is at most $O(\epsilon^2)$. We take our mixture distribution ξ to be μ_S, which is over pinnings $\tau : S \to \{\pm 1\}$; our component measures will be the conditionals μ^τ. Let us now verify the conditions of Theorem 2.1 for this decomposition.

- Since S has size at most $O(1/\epsilon^2)$ and $\xi = \mu_S$ is supported on a set of size $2^{|S|}$, we have $H(\xi) \leq |S| \leq O(1/\epsilon^2)$ where the first inequality follows from the fact that the uniform measure maximizes Shannon entropy. But $H(\mu) - \mathbb{E}_{\theta \sim \xi} \left[H\left(\mu^{(\theta)}\right) \right] \leq H(\xi)$ (e.g. by using the Chain Rule for conditional entropies), and so we may take $\alpha = O(1/\epsilon^2)$.
- For the components, recall that since $f(\sigma) = \frac{1}{2}\sigma^\top A \sigma$, for every $\tau : S \to \{\pm 1\}$,

$$\mathbb{E}_{\sigma \sim \mu^\tau}[f(\sigma)] = \frac{1}{2} \sum_{i,j=1}^n A_{ij} \mathbb{E}_{\sigma \sim \mu^\tau} [\sigma_i \sigma_j]$$

$$\mathbb{E}_{\sigma \sim \pi(\mu^\tau)}[f(\sigma)] = \frac{1}{2} \sum_{i,j=1}^n A_{ij} \mathbb{E}_{\sigma \sim \mu^\tau} [\sigma_i] \mathbb{E}_{\sigma \sim \mu^\tau} [\sigma_j] .$$

  It follows that

$$\mathbb{E}_{\sigma \sim \mu^\tau}[f(\sigma)] - \mathbb{E}_{\sigma \sim \pi(\mu^\tau)}[f(\sigma)] \leq \frac{1}{2} \sum_{i,j=1}^n A_{ij} \mathrm{Cov}_{\sigma \sim \mu^\tau}(\sigma_i, \sigma_j) = \frac{1}{2} \mathrm{Tr}\left(A \, \mathrm{Cov}\left(\mu^\tau\right)\right).$$

  Averaging over a random choice of $\tau \sim \xi = \mu_S$, we obtain that

$$\begin{aligned}
\mathbb{E}_{\tau \sim \mu_S} \left[\mathbb{E}_{\sigma \sim \mu^\tau}[f(\sigma)] - \mathbb{E}_{\sigma \sim \pi(\mu^\tau)}[f(\sigma)] \right] &\leq \frac{1}{2} \mathrm{Tr}\left(A \cdot \mathbb{E}_{\tau \sim \mu_S} \left[\mathrm{Cov}\left(\mu^\tau\right)\right]\right) \\
&\leq \frac{1}{2} \|A\|_F \cdot \left\| \mathbb{E}_{\tau \sim \mu_S} \left[\mathrm{Cov}\left(\mu^\tau\right)\right] \right\|_F \quad \text{(Cauchy–Schwarz)} \\
&\leq \frac{1}{2} \|A\|_F \cdot \mathbb{E}_{\tau \sim \mu_S} \left[\left\| \mathrm{Cov}\left(\mu^\tau\right) \right\|_F^2 \right]^{1/2} \\
&\qquad \text{(Cauchy–Schwarz)} \\
&\leq O(\epsilon) \cdot n \cdot \|A\|_F . \qquad \text{(Eq. (6))}
\end{aligned}$$

  Hence, we take $\eta = O(\epsilon) \cdot n \cdot \|A\|_F$.

Combining these calculations with Theorem 2.1, we obtain that the error in the naïve mean-field approximation is upper bounded as

$$\mathcal{F} - \mathcal{F}_{\mathsf{NMF}} \leq O(1/\epsilon^2) + O(\epsilon) \cdot n \cdot \|A\|_F .$$

Choosing $\epsilon \approx n^{-1/3} \|A\|_F^{-1/3}$ to balance both terms in the right-hand side, we obtain the desired bound. □

[3] It might look slightly strange that the inner expectation is over a random pair of coordinates $\{i,j\} \in \binom{[n]}{2}$, not a random pair of *unpinned* coordinates $\{i,j\} \in \binom{[n] \setminus S}{2}$. Indeed, if $i \in S$ or $j \in S$, then $\mathrm{Cov}_{\sigma \sim \mu^\tau}(\sigma_i, \sigma_j) = 0$. However, in the primary regime of interest, i.e. $\ell \leq cn$ for some (typically small) constant $0 < c < 1$, the two are interchangeable up to small multiplicative losses in the right-hand side.

2.1.1 Proof of Lemma 2.2

It will be more convenient to use information-theoretic quantities, since we can then use the chain rule for entropy. Recall that if X, Y are two random variables, then the *mutual information* between them (or their laws) is defined as

$$\mathscr{I}(X;Y) \stackrel{\text{def}}{=} \mathscr{D}_{\mathrm{KL}}\left(\mathsf{Law}(X,Y) \,\|\, \mathsf{Law}(X) \otimes \mathsf{Law}(Y)\right).$$

It can also be expressed as

$$\mathscr{I}(X;Y) = H(X) - H(X \mid Y),$$

where $H(X \mid Y) \stackrel{\text{def}}{=} \mathbb{E}_{y \sim \mathsf{Law}(Y)}\left[H(X \mid Y = y)\right]$ is the *conditional entropy*. $\mathscr{I}(X;Y)$ is another way of measuring the correlation between X and Y using the KL-divergence.

Lemma 2.3 ([BRS11])**.** *Let X, Y be $\{\pm 1\}$-valued random variables. Then $\mathrm{Cov}(X,Y)^2 \le 2 \cdot \mathscr{I}(X;Y)$.*

A proof is provided in Appendix A. Using this, it suffices to find $S \subseteq [n]$ with $|S| \le \ell - 1$ and

$$\mathbb{E}_{\{i,j\} \sim \mathsf{Unif}\binom{[n]}{2}}\left[\mathscr{I}\left(\sigma_i; \sigma_j \mid \sigma_S\right)\right] \le \frac{\log 2}{\ell}. \tag{7}$$

Now, observe that for any $S \subseteq [n]$ and any $i, j \in [n]$,

$$\mathscr{I}\left(\sigma_i; \sigma_j \mid \sigma_S\right) = H\left(\sigma_j \mid \sigma_S\right) - H\left(\sigma_j \mid \sigma_{S \cup \{i\}}\right). \tag{8}$$

Our goal is to upper bound the mutual information. The trick is that the left-hand side can be averaged, while the right-hand side is something we can telescope to keep small. More specifically, if $i_1, \ldots, i_\ell, j$ is any sequence of coordinates (e.g. by ordering the elements of $S \cup \{i\} = \{i_1, \ldots, i_\ell\}$ from Eq. (8)), then

$$\begin{aligned}
\frac{1}{\ell}\sum_{t=1}^{\ell} \mathscr{I}\left(\sigma_{i_t}; \sigma_j \mid \sigma_{i_1}, \ldots, \sigma_{i_{t-1}}\right) &= \frac{1}{\ell}\sum_{t=1}^{\ell}\left[H(\sigma_j \mid \sigma_{i_1}, \ldots, \sigma_{i_{t-1}}) - H(\sigma_j \mid \sigma_{i_1}, \ldots, \sigma_{i_t})\right] && \text{(Using Eq. (8))} \\
&= \frac{H(\sigma_j) - H(\sigma_j \mid \sigma_{i_1}, \ldots, \sigma_{i_\ell})}{\ell} && \text{(Telescoping)} \\
&\le \frac{\log 2}{\ell}. && \text{(Nonnegative of entropy, and } \sigma_j \in \{\pm 1\})
\end{aligned}$$

Averaging over random coordinates $i_1, \ldots, i_\ell, j \sim \mathsf{Unif}[n]$ drawn independently, we see that

$$\frac{1}{\ell}\sum_{t=1}^{\ell} \mathop{\mathbb{E}}_{i_1, \ldots, i_{t-1} \sim \mathsf{Unif}[n]}\left[\mathop{\mathbb{E}}_{i_t, j \sim \mathsf{Unif}[n]}\left[\mathscr{I}\left(\sigma_{i_t}; \sigma_j \mid \sigma_{i_1}, \ldots, \sigma_{i_{t-1}}\right)\right]\right] \le \frac{\log 2}{\ell}.$$

Since this holds for a randomly chosen subset of coordinates $\{i_1, \ldots, i_{t-1}\}$ for a randomly chosen $1 \le t \le \ell$, there must be some specific subset $S \subseteq [n]$ of size $\le \ell - 1$ such that Eq. (7) holds.

2.2 Refined Decompositions via Stochastic Localization

We now show how to prove Theorem 1.2 by giving a more refined decomposition based on stochastic localization. The main decomposition result is the following.

Theorem 2.4 ([Eld20])**.** *Let μ be any probability measure over $\{\pm 1\}^n$.*[4] *Then for every symmetric positive definite matrix $L \succ 0$, there exists a decomposition of μ as $\mu = \mathbb{E}_{\theta \sim \xi}\left[\mu^{(\theta)}\right]$ enjoying the following properties:*

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

[4] Eldan's Theorem extends to any measure over $\mathbb{R}^n$.

- $\mathbb{E}_{\theta\sim\xi}\left[\mathrm{Cov}\left(\mu^{(\theta)}\right) L \,\mathrm{Cov}\left(\mu^{(\theta)}\right)\right] \preceq \mathrm{Cov}(\mu)$

Let us first use this to complete the proof of Theorem 1.2.

Proof of Theorem 1.2. Using the decomposition furnished by Theorem 2.4 with $L = \left(A^2\right)^{1/2}$, we may take $\alpha = \log\det\left(\mathsf{Id} + L^{1/2}\,\mathrm{Cov}(\mu)L^{1/2}\right)$ in Theorem 2.1 just using the first property of the decomposition. Now, let us verify Eq. (5). Following the proof of Theorem 1.1, since $f(\sigma) = \frac{1}{2}\sigma^\top A\sigma$ is quadratic,

$$\begin{aligned}\mathbb{E}_{\mu^{(\theta)}}[f] - \mathbb{E}_{\pi\left(\mu^{(\theta)}\right)}[f] &= \frac{1}{2}\,\mathrm{Tr}\left(A\,\mathrm{Cov}\left(\mu^{(\theta)}\right)\right)\\ &\le \frac{1}{2}\,\mathrm{Tr}\left(L^{1/2}\,\mathrm{Cov}\left(\mu^{(\theta)}\right)L^{1/2}\right). \qquad \text{(Using } L \succeq A\text{)}\end{aligned}$$

Averaging over $\theta\sim\xi$, obtain that

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

We claim the right-hand side is at most $2\log\det\left(\mathsf{Id} + L^{1/2}\,\mathrm{Cov}(\mu)L^{1/2}\right)$. Indeed, the second item of Theorem 2.4 ensures that

$$\mathbb{E}_{\theta\sim\xi}\left[L^{1/2}\,\mathrm{Cov}\left(\mu^{(\theta)}\right)L^{1/2}\right] \preceq \mathsf{Id}.$$

At the same time, the Law of Total Covariance ensures that $\mathbb{E}_\theta\left[\mathrm{Cov}\left(\mu^{(\theta)}\right)\right] \preceq \mathrm{Cov}(\mu)$, whence

$$\mathbb{E}_{\theta\sim\xi}\left[L^{1/2}\,\mathrm{Cov}\left(\mu^{(\theta)}\right)L^{1/2}\right] \preceq L^{1/2}\,\mathrm{Cov}(\mu)L^{1/2}.$$

It follows that each individual eigenvalue satisfies

$$\begin{aligned}\lambda_i\left(\mathbb{E}_{\theta\sim\xi}\left[L^{1/2}\,\mathrm{Cov}\left(\mu^{(\theta)}\right)L^{1/2}\right]\right) &\le \min\left\{1, \lambda_i\left(L^{1/2}\,\mathrm{Cov}(\mu)L^{1/2}\right)\right\}\\ &\le 2\log\left(1 + \lambda_i\left(L^{1/2}\,\mathrm{Cov}(\mu)L^{1/2}\right)\right). \qquad \text{(Using } \mathrm{Cov}(\mu) \succeq 0\text{)}\end{aligned}$$

Summing over all i yields the claim. □

2.2.1 Proof of Eldan's Decomposition Theorem

For brevity, we only prove the first two items. The third property $\mathbb{E}_{\theta\sim\xi}\left[\mathrm{Cov}\left(\mu^{(\theta)}\right) L \,\mathrm{Cov}\left(\mu^{(\theta)}\right)\right] \preceq \mathrm{Cov}(\mu)$ requires a little more work, and was not used in the proof of Theorem 1.2; we refer interested readers to [Eld20; AM22] for its proof.

We follow the presentation in [AM22]. We use the Gaussian channel localization (or stochastic localization), which we recall is given by the following statistical inference formulation: For $L \succ 0$ as in the theorem statement, we draw $\sigma \sim \mu$ (the "signal") and $g \sim \mathcal{N}(0, \mathsf{Id})$ (the "noise") independently, and observe

$$\theta = \sigma + L^{-1/2}g.$$

In other words, we pass σ through a noisy channel which additively corrupts σ with Gaussian noise.[5] The component measures in the decomposition are then given by the law of σ conditioned on observing θ, i.e. $\mu^{(\theta)} \stackrel{\mathsf{def}}{=} \mathsf{Law}(\sigma \mid \theta)$. The mixture measure ξ is described by convolution

$$\xi(\theta) \propto \mathbb{E}_{\sigma\sim\mu}\left[\mathbb{E}_{g\sim\mathcal{N}(0,\mathsf{Id})}\left[\mathbf{1}_{\theta=\sigma+L^{-1/2}g}\right]\right]. \tag{9}$$

on $\mathbb{R}^n$. With this in hand, we now prove the first two items in turn.

- Observe that

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

[5] Technically, at least for the first two items, we do not need the full power of stochastic localization, only a "snapshot" at a specific time.

where the mutual information is computed w.r.t. the above coupling between θ and σ. This identity has corresponds to how much "information" we have about σ knowing θ. However, we can go the other way by symmetry, and view $\mathscr{I}(\sigma;\theta)$ as expressing how much "information" we have about θ knowing σ. In particular,

$$\mathscr{I}(\sigma;\theta) = H(\theta) - H(\theta \mid \sigma).$$

To bound the first term, recall that for a fixed positive definite matrix Σ, the Gaussian $\mathcal{N}(0,\Sigma)$ maximizes (differential) entropy out of all (centered) distributions with covariance Σ. Hence,

$$\begin{aligned} H(\theta) \le H\left(\mathcal{N}\left(0, \mathrm{Cov}\left(\xi\right)\right)\right) &= \frac{n}{2}\log(2\pi e) + \frac{1}{2}\operatorname{Tr}\log \mathrm{Cov}\left(\xi\right) \\ &= \frac{n}{2}\log(2\pi e) + \frac{1}{2}\operatorname{Tr}\log\left(L^{-1} + \mathrm{Cov}(\mu)\right). \qquad \text{(Using Eq. (9))} \end{aligned}$$

For the second term, since $H(\theta \mid \sigma) = H(L^{-1/2}g)$, we obtain

$$H(\theta \mid \sigma) = \frac{n}{2}\log(2\pi e) + \frac{1}{2}\operatorname{Tr}\log L^{-1}.$$

It follows that

$$\begin{aligned} \mathscr{I}(\sigma;\theta) &\le \frac{1}{2}\operatorname{Tr}\log \mathrm{Cov}\left(\mu^{(\theta)}\right) - \frac{1}{2}\operatorname{Tr}\log L^{-1} \\ &\le \frac{1}{2}\log\det\left(\mathsf{Id} + L^{1/2}\,\mathrm{Cov}(\mu)L^{1/2}\right). \end{aligned}$$

This establishes the first item.

- Our goal is to establish that $\mathbb{E}_{\theta\sim\xi}\left[\mathrm{Cov}\left(\mu^{(\theta)}\right)\right] \preceq L^{-1} = \mathrm{Cov}\left(-L^{-1/2}g\right) = \mathrm{Cov}\left(\sigma - \theta\right)$, i.e. for every other positive semidefinite matrix $B \succeq 0$,

$$\operatorname{Tr}\left(\mathbb{E}_{\theta\sim\xi}\left[\mathrm{Cov}\left(\mu^{(\theta)}\right)\right]\cdot B\right) \le \operatorname{Tr}\left(\mathrm{Cov}(\sigma-\theta)\cdot B\right).$$

Applying cyclicity of trace, this is equivalent to

$$\mathbb{E}\left[\left(\sigma - \boldsymbol{m}\left(\mu^{(\theta)}\right)\right)^{\top} B\left(\sigma - \boldsymbol{m}\left(\mu^{(\theta)}\right)\right)\right] \le \mathbb{E}\left[(\sigma-\theta)^{\top} B\,(\sigma-\theta)\right], \tag{10}$$

where the expectations on both sides are both w.r.t. a draw of θ, σ from the above process. Eq. (10) has a natural interpretation from *estimation theory*. Consider the problem of approximating σ given knowledge of θ. There are two natural estimators for σ. One is simply to output θ, since we know the noise is centered. This is the "maximum likelihood estimator", and is what constitutes the right-hand side of Eq. (10). The other is to use the *Bayes estimator*, namely the mean $\boldsymbol{m}\left(\mu^{(\theta)}\right)$ of $\mu^{(\theta)} = \mathsf{Law}(\sigma \mid \theta)$. This gives the left-hand side of Eq. (10). Framed this way, the inequality in Eq. (10) is just saying that the Bayes estimator is indeed *optimal*, in the sense that it minimizes the expected Euclidean distance to σ (reweighted by B) out of all possible estimators. This is a standard result in Bayesian statistics, and we leave it as an exercise.

2.3 "Low-Complexity" Conditions

A related line of work shows that under a suitable "low-complexity" condition on the Hamiltonian f defining μ, much stronger decompositions are achievable. For instance, one can guarantee that the most component measures are close in transportation distance to a product measure, which is significantly stronger than Eq. (5). We refer interested readers to [CD16; BM17; Eld18; EG18; Aus19; Aug21].

References

[AM22] Ahmed El Alaoui and Andrea Montanari. "An Information-Theoretic View of Stochastic Localization". In: *IEEE Transactions on Information Theory* 68.11 (2022) (cit. on p. 6).

[Aug21] Fanny Augeri. "A transportation approach to the mean-field approximation". In: *Probability Theory and Related Fields* 180 (2021), pp. 1–32 (cit. on p. 7).

[Aus19] Tim Austin. "The structure of low-complexity Gibbs measures on product spaces". In: *The Annals of Probability* 47.6 (2019), pp. 4002–4023 (cit. on p. 7).

[BC16] Victor Bapst and Amin Coja-Oghlan. "Harnessing the Bethe free energy". In: *Random Structures & Algorithms* 49.4 (2016), pp. 694–741. DOI: https://doi.org/10.1002/rsa.20692. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/rsa.20692 (cit. on p. 3).

[BM17] Anirban Basak and Sumit Mukherjee. "Universality of the mean-field for the Potts model". In: *Probability Theory and Related Fields* 168 (2017), pp. 557–600 (cit. on pp. 2, 7).

[BRS11] Boaz Barak, Prasad Raghavendra, and David Steurer. "Rounding Semidefinite Programming Hierarchies via Global Correlation". In: *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2011, pp. 472–481. DOI: 10.1109/FOCS.2011.95 (cit. on pp. 4, 5).

[CD16] Sourav Chatterjee and Amir Dembo. "Nonlinear large deviations". In: *Advances in Mathematics* 299 (2016), pp. 396–450. ISSN: 0001-8708. DOI: https://doi.org/10.1016/j.aim.2016.05.017 (cit. on p. 7).

[Coj+18] Amin Coja-Oghlan, Florent Krzakala, Will Perkins, and Lenka Zdeborová. "Information-theoretic thresholds from the cavity method". In: *Advances in Mathematics* 333 (2018), pp. 694–795. ISSN: 0001-8708. DOI: https://doi.org/10.1016/j.aim.2018.05.029 (cit. on p. 3).

[EG18] Ronen Eldan and Renan Gross. "Decomposition of mean-field Gibbs distributions into product measures". In: *Electronic Journal of Probability* 23 (2018), pp. 1–24 (cit. on p. 7).

[Eld18] Ronen Eldan. "Gaussian-width gradient complexity, reverse log-Sobolev inequalities and nonlinear large deviations". In: *Geometric and Functional Analysis* 28 (2018), pp. 1548–1596 (cit. on p. 7).

[Eld20] Ronen Eldan. "Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation". In: *Probability Theory and Related Fields* 176 (2020), pp. 737–755 (cit. on pp. 2, 5, 6).

[JKR19] Vishesh Jain, Frederic Koehler, and Andrej Risteski. "Mean-Field Approximation, Convex Hierarchies, and the Optimality of Correlation Rounding: A Unified Perspective". In: *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. STOC 2019. Phoenix, AZ, USA: Association for Computing Machinery, 2019, pp. 1226–1236. ISBN: 9781450367059. DOI: 10.1145/3313276.3316299 (cit. on pp. 2, 3).

[Koe19] Frederic Koehler. "Fast Convergence of Belief Propagation to Global Optima: Beyond Correlation Decay". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019 (cit. on p. 2).

[Mon08] Andrea Montanari. "Estimating random variables from random sparse observations". In: *European Transactions on Telecommunications* 19.4 (2008), pp. 385–403. DOI: https://doi.org/10.1002/ett.1289. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ett.1289 (cit. on p. 4).

[MR17] Pasin Manurangsi and Prasad Raghavendra. "A Birthday Repetition Theorem and Complexity of Approximating Dense CSPs". In: *44th International Colloquium on Automata, Languages, and Programming (ICALP 2017)*. Ed. by Ioannis Chatzigiannakis, Piotr Indyk, Fabian Kuhn, and Anca Muscholl. Vol. 80. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2017, 78:1–78:15. ISBN: 978-3-95977-041-5. DOI: 10.4230/LIPIcs.ICALP.2017.78 (cit. on p. 4).

[Ris16] Andrej Risteski. "How to calculate partition functions using convex programming hierarchies: provable bounds for variational methods". In: *29th Annual Conference on Learning Theory*. Ed. by Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir. Vol. 49. Proceedings of Machine Learning Research. Columbia University, New York, New York, USA: PMLR, 23–26 Jun 2016, pp. 1402–1416 (cit. on p. 2).

[RT12] Prasad Raghavendra and Ning Tan. "Approximating CSPs with Global Cardinality Constraints Using SDP Hierarchies". In: *Proceedings of the 2012 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. 2012, pp. 373–387. DOI: 10.1137/1.9781611973099.33. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9781611973099.33 (cit. on p. 4).

A Unfinished Proofs

Proof of Lemma 2.3. The main idea is to write $\mathrm{Cov}(X, Y)$ using total variation distance, and then apply Pinsker's Inequality. Observe that

$$\begin{aligned}
&\mathscr{D}_{\mathsf{TV}}\left(\mathsf{Law}(X,Y), \mathsf{Law}(X) \otimes \mathsf{Law}(Y)\right) \\
&= \frac{1}{2} \sum_{x,y \in \{\pm 1\}} \left|\Pr[X = x, Y = y] - \Pr[X = x]\Pr[Y = y]\right| \\
&= \frac{1}{2} \sum_{x,y \in \{\pm 1\}} |x| \cdot |y| \cdot \left|\Pr[X = x, Y = y] - \Pr[X = x]\Pr[Y = y]\right| && \text{(Since } x, y \in \{\pm 1\}\text{)} \\
&\geq \frac{1}{2} \left| \sum_{x,y \in \{\pm 1\}} xy \Pr[X = x, Y = y] - \sum_{x,y \in \{\pm 1\}} x \Pr[X = x] \cdot y \Pr[Y = y] \right| && \text{(Triangle Inequality)} \\
&= \frac{1}{2} \left|\mathrm{Cov}(X, Y)\right| .
\end{aligned}$$

Applying Pinsker's Inequality to upper bound the left-hand side yields the claim. □