Lecture 23: Convex Relaxations; The Polynomial Perspective

December 5, 2023

In the previous lecture on the naïve mean-field approximation, we approximated the free energy by *restricting* the class of probability measures in Gibbs Variational Principle. This gives us a lower bound on the free energy whose approximation error was controlled via measure decompositions. In this lecture, we do we opposite. We *relax* the optimization problem to include objects which are not bonafide distributions globally. We show how to do this in a principled way using convex programming hierarchies. In the second half of the lecture, we take the dual perspective to the Gibbs Variational Principle, and design deterministic algorithms for estimating the number of bases of a matroid.

1 Relaxing to Pseudo-Distributions

Again, we focus on distributions over the Boolean cube $\{\pm 1\}^n$ for convenience. Let $f : \{\pm 1\}^n \to \mathbb{R}$ be a function, which is the *Hamiltonian* defining the associated Gibbs measure $\mu(\sigma) \propto e^{f(\sigma)}$. We recall the following standard fact from analysis of Boolean functions.

Fact 1.1. *Let $f : \{\pm 1\}^n \to \mathbb{R}$ be an arbitrary function. Then there is a unique multiaffine polynomial $\sum_{S \subseteq [n]} \hat{f}(S) \prod_{i \in S} x_i$ which agrees with f on $\{\pm 1\}^n$. The numbers $\hat{f}(S)$ are the* Fourier coefficients *of f. We write $\mathrm{supp}(f) \overset{\mathsf{def}}{=} \{S \subseteq [n] : \hat{f}(S) \neq 0\}$ for the* support *of f, and $\deg(f) \overset{\mathsf{def}}{=} \max_{S \in \mathrm{supp}(f)} |S|$ for the* degree *of f.*

We typically think of f as being a *low-degree function*, i.e. it is a linear combination of low-degree monomials (e.g. f is quadratic, corresponding to Ising models). This is the most relevant setting for statistical physics applications, since f should be "local" in some sense.

Now recall that by Gibbs Variational Principle, the free energy can be written as the convex program

$$\mathcal{F} = \mathcal{F}(f) \overset{\mathsf{def}}{=} \log \sum_{\sigma \in \{\pm 1\}^n} e^{f(\sigma)} = \sup_{\nu} \left\{ \mathbb{E}_{\nu}[f] + H(\nu) \right\}, \tag{1}$$

where the supremum is over all probability measures ν over $\{\pm 1\}^n$. In general, it is intractable write down ν. One way to make things tractable is to relax the constraint that ν be a genuine probability measure. However, in order for this to have any chance of working, we need to ensure that both quantities $\mathbb{E}_{\nu}[f]$ and $H(\nu)$ still make sense in the relaxation; if they do not, we need to find "good" surrogates for them. Ignoring $H(\nu)$ for the moment, "locality/low-degreeness" of f already suggests a natural relaxation, since the expectation of a low-degree function only requires knowledge of the low-degree moments of ν.

Definition 1 (Pseudo-Distribution)**.** *For a downwards closed*[1] *family of subsets $\mathscr{F} \subseteq 2^{[n]}$, a $\mathscr{F}$-*pseudo-distribution *over $\{\pm 1\}^n$ is a collection $\tilde{\boldsymbol{p}} = \{\tilde{\boldsymbol{p}}_S\}_{S \in \mathscr{F}}$ of probability distributions $\tilde{\boldsymbol{p}}_S$ over $\{\pm 1\}^S$ satisfying the following* local consistency relations:

$$\tilde{\boldsymbol{p}}_S(\tau) = \Pr_{\sigma \sim \tilde{\boldsymbol{p}}_T}[\sigma_S = \tau], \qquad \forall S, T \in \mathscr{F} \textit{ s.t. } S \subseteq T, \forall \tau : S \to \{\pm 1\}. \tag{2}$$

The degree *of the pseudo-distribution is defined as $\max_{S \in \mathscr{F}} |S|$.*

[1] This means that if $T \in \mathscr{F}$ and $S \subseteq T$, then $S \in \mathscr{F}$.

Remark 1. This notion of locally consistent collections of local probability distributions makes sense for any product space, e.g. $[q]^n$.

The benefit of Definition 1 is that a $\mathscr{F}$-pseudo-distribution of degree-k only has $\sum_{S\in\mathscr{F}} 2^{|S|} \leq n^{O(k)}$ parameters, which is polynomial for constant k. Furthermore, the set of $\mathscr{F}$-pseudo-distributions is a polytope whose constraints are specified by Eq. (2) (plus nonnegativity and normalization for each marginal $\tilde{\boldsymbol{p}}_S$). If $k \leq O(1)$, there are only polynomially many such constraints, and so we can efficiently optimize convex/concave functions over degree-$O(1)$ pseudo-distributions.

Clearly, given a genuine probability distribution μ over $\{\pm 1\}^n$, the collection of marginals $\{\mu_S\}_{S\in\mathscr{F}}$ is a $\mathscr{F}$-pseudo-distribution over $\{\pm 1\}^n$ for every downwards closed family $\mathscr{F} \subseteq 2^{[n]}$. Hence, the set of $\mathscr{F}$-pseudo-distributions fully contains the set of valid probability distribution over $\{\pm 1\}^n$. In general, this containment is strict; see e.g. Example 1.

Now to each $\mathscr{F}$-pseudo-distribution $\tilde{p}$ comes with an associated *pseudo-expectation operator* $\tilde{\mathbb{E}} = \tilde{\mathbb{E}}_{\tilde{\boldsymbol{p}}}$ which acts on (multiaffine) polynomials (see Fact 1.1) $f(x) = \sum_{S\subseteq[n]} c_S \prod_{i\in S} x_i$ satisfying $\mathrm{supp}(f) \subseteq \mathscr{F}$:

$$\tilde{\mathbb{E}}\left[f\right] = \sum_{S\in\mathscr{F}} c_S \cdot \mathbb{E}_{\sigma_S\sim\tilde{\boldsymbol{p}}_S}\left[\prod_{i\in S}\sigma_i\right]. \tag{3}$$

Theme 1.2. *Relax Eq. (1) to optimizing over $\mathscr{F}$-pseudo-distributions for some "nice/low-degree" (and downwards closed) $\mathscr{F} \subseteq 2^{[n]}$.*

Of course, we have neglected the entropy term $H(\nu)$, which is no longer well-defined for $\mathscr{F}$-pseudo-distributions. We will see a couple ways of remedying this.

1.1 Pairwise Interactions and the Bethe Approximation

A natural approach to defining a "pseudo-entropy" functional for $\mathscr{F}$-pseudo-distributions is to sum up the entropies of each of the marginals. This is not quite correct due to "double counting" issues, but morally, this approach is what gives rise to the *Bethe approximation* in the context of models with pairwise interactions. To illustrate it, we consider the very special case where f is the quadratic form w.r.t. the adjacency matrix of some graph $G = (V, E)$, and μ is the corresponding Ising model; note that by Remark 1, everything we will say here extends to any q-state spin system. Since f is quadratic, let us consider degree-2 pseudo-distributions.

Definition 2 (Bethe Approximation (Very Special Case))**.** *Let $G = (V, E)$ be a graph, and let $f(\sigma) = \frac{1}{2}\sigma^\top A\sigma$ where A is the adjacency matrix of G. Let $\mathscr{F}$ be the downwards closure of the set of edges E; note that* $\mathrm{supp}(f) \subseteq \mathscr{F}$. *Define the* Bethe free energy *by*

$$\mathcal{F}_{\mathsf{Bethe}} = \mathcal{F}_{\mathsf{Bethe}}(f) \overset{\mathsf{def}}{=} \sup_{\mathscr{F}\text{-pseudo-distribution } \tilde{\boldsymbol{p}}} \left\{\tilde{\mathbb{E}}[f] + H_{\mathsf{Bethe}}\left(\tilde{\boldsymbol{p}}\right)\right\}, \tag{4}$$

where $H_{\mathsf{Bethe}}\left(\tilde{\boldsymbol{p}}\right)$ is the Bethe entropy

$$\begin{aligned} H_{\mathsf{Bethe}}\left(\tilde{\boldsymbol{p}}\right) &\overset{\mathsf{def}}{=} \sum_{e\in E} H\left(\tilde{\boldsymbol{p}}_e\right) - \sum_{v\in V}\left(\deg(v) - 1\right) H\left(\tilde{\boldsymbol{p}}_v\right) \\ &= \sum_{v\in V} H\left(\tilde{\boldsymbol{p}}_v\right) - \sum_{e\in E} \mathscr{I}\left(\tilde{\boldsymbol{p}}_e\right). \end{aligned} \tag{5}$$

Again, the correction $-\sum_{v\in V}\left(\deg(v) - 1\right) H\left(\tilde{\boldsymbol{p}}_v\right)$ is just to avoid "double counting", since each single-vertex marginal distribution $\tilde{\boldsymbol{p}}_v$ participates in $\deg(v)$-many $H\left(\tilde{\boldsymbol{p}}_e\right)$. It is not difficult to show that if G is a tree, then there is a unique global distribution whose pairwise marginals are described by such a $\mathscr{F}$-pseudo-distribution. Furthermore, in this special case, H_{Bethe} gives precisely the Shannon entropy of this unique global distribution.

The Bethe free energy $\mathcal{F}_{\mathsf{Bethe}}$ is a heavily studied approximation to the true free energy, particularly in sparse graphs satisfying suitable girth conditions; see e.g. the monograph [WJ08] and references therein. A well-known calculation due to [YFW00; YFW05] shows that the optimizers of Eq. (4) satisfy a certain fixed point equation known as the *belief propagation equations*. This is precisely the tree recursion we previously derived combinatorially for spin systems on trees using conditional independence. It gives a variational perspective on this.

2 The Sherali–Adams Hierarchy

Now, we turn to relaxations of Eq. (1) based on convex programming *hierarchies*. For this, we let $\mathscr{F}_k = \bigcup_{j=0}^{k} \binom{[n]}{j}$ for some $k \in [n]$, and let $\mathsf{SA}(k;[n])$ denote the polytope of $\mathscr{F}_k$-pseudo-distributions. This is the *level-k Sherali–Adams relaxation* for the polytope of globally consistent probability distributions. By increasing k, we increase the accuracy (and computational complexity) of our approximations, giving us a *hierarchy* of optimization problems.

Now let us define an appropriate "pseudo-entropy" functional for $\mathscr{F}_k$-pseudo-distributions. For this, we first need to know how to condition a pseudo-distribution.

Fact 2.1 (Conditioning a Pseudo-Distribution)**.** *Let $\tilde{\boldsymbol{p}} \in \mathsf{SA}(k;[n])$. For any subset of coordinates $S \subseteq [n]$ with $|S| \leq k-1$, and any pinning $\tau : S \to \{\pm 1\}$, define $\tilde{\boldsymbol{p}}^{\tau}$ by*

$$\tilde{\boldsymbol{p}}^{\tau}_T(\sigma) \overset{\mathsf{def}}{=} \tilde{\boldsymbol{p}}_{S \sqcup T}(\tau, \sigma) \qquad \forall T \subseteq [n] \setminus S \text{ s.t. } |T| \leq k - |S|, \forall \sigma : T \to \{\pm 1\}.$$

Then $\tilde{\boldsymbol{p}}^{\tau} \in \mathsf{SA}(k - |S|; [n] \setminus S)$.

Definition 3 (("Augmented Mean-Field") Pseudo-Entropy; [Ris16])**.** *Let $\tilde{\boldsymbol{p}} \in \mathsf{SA}(k;[n])$. For each $0 \leq j \leq k-1$, define its j*th ("augmented mean-field") pseudo-entropy *by the quantity*

$$\tilde{H}_j(\tilde{\boldsymbol{p}}) \overset{\mathsf{def}}{=} \min_{S:|S| \leq j} \left\{ H(\tilde{\boldsymbol{p}}_S) + \sum_{i \notin S} H(\tilde{\boldsymbol{p}}_i \mid \tilde{\boldsymbol{p}}_S) \right\},$$

where the second conditional entropy term is w.r.t. $\tilde{\boldsymbol{p}}_{S \cup \{i\}}$ and given by

$$H(\tilde{\boldsymbol{p}}_i \mid \tilde{\boldsymbol{p}}_S) \overset{\mathsf{def}}{=} \mathbb{E}_{\tau \sim \tilde{\boldsymbol{p}}_S}\left[H(\tilde{\boldsymbol{p}}^{\tau}_i)\right].$$

If μ is a genuine probability distribution over $\{\pm 1\}^n$, we simply write $\tilde{H}_j(\mu)$ for the pseudo-entropy of its collection of marginal distributions.

Lemma 2.2 ([Ris16])**.** *For every $0 \leq j \leq k-1$, the function $\tilde{\boldsymbol{p}} \mapsto \tilde{H}_j(\tilde{\boldsymbol{p}})$ over $\mathsf{SA}(k;[n])$ enjoys the following properties:*

- *For every genuine probability distribution μ over $\{\pm 1\}^n$, $H(\mu) \leq \tilde{H}_j(\mu)$.*
- *The function is concave over $\mathsf{SA}(k;[n])$.*

Proof Sketch. The first property just follows from subadditivity of entropy, since for every $S \subseteq [n]$, letting $\sigma \sim \mu$, we have

$$H(\mu) = H(\sigma_S) + H(\sigma_{[n] \setminus S} \mid \sigma_S) \leq H(\sigma_S) + \sum_{i \notin S} H(\sigma_i \mid \sigma_S).$$

Picking the best possible S such that $|S| \leq j$ gives the first claim. For the second claim, note that concavity is closed under taking sums and minima. Hence, it suffices to establish concavity of $H(\tilde{\boldsymbol{p}}_S)$ and $H(\tilde{\boldsymbol{p}}_i \mid \tilde{\boldsymbol{p}}_S)$ for each S. Both just follow from the standard proof of concavity of Shannon entropy (or convexity of KL-divergence). □

Definition 4 ("Sherali–Adams" Free Energy; [Ris16])**.** *Fix a function $f : \{\pm 1\}^n \to \mathbb{R}$. In light of Definition 1 and Definition 3, we define the "Sherali–Adams" free energy for each $k \geq \deg(f)$ and $0 \leq j \leq k-1$ by*

$$\mathcal{F}_{\mathsf{SA}(k;[n]),j} = \mathcal{F}_{\mathsf{SA}(k;[n]),j}(f) \overset{\mathsf{def}}{=} \sup_{\tilde{\boldsymbol{p}} \in \mathsf{SA}(k;[n])} \left\{ \tilde{\mathbb{E}}[f] + \tilde{H}_j(\tilde{\boldsymbol{p}}) \right\}. \tag{6}$$

Theorem 2.3 ([Ris16]; see also [JKR19])**.** *Fix a symmetric interaction matrix $A \in \mathbb{R}^{n \times n}$, and let $f(\sigma) = \frac{1}{2}\sigma^{\top} A \sigma$. For $0 \leq k \leq n-2$, $\mathcal{F}_{\mathsf{SA}(k+2;[n]),k}(f)$ can be computed in $n^{O(k)}$-time, and satisfies*

$$0 \leq \mathcal{F}_{\mathsf{SA}(k+2;[n]),k}(f) - \mathcal{F}(f) \leq O\left(\frac{n \|A\|_F}{\sqrt{k}}\right).$$

Furthermore, if $\tilde{\boldsymbol{p}}$ is the optimal pseudo-distribution attaining $\mathcal{F}_{\mathsf{SA}(k+2;[n]),k}(f)$, then we can round *$\tilde{\boldsymbol{p}}$ into a product measure π over $\{\pm 1\}^n$ satisfying*

$$\mathcal{F}(f) - \{\mathbb{E}_{\pi}[f] + H(\pi)\} \leq O\left(\frac{n \|A\|_F}{\sqrt{k}} + k\right).$$

(Note that the left-hand side upper bounds $\mathcal{F}(f) - \mathcal{F}_{\mathsf{NMF}}(f)$.)

One should think of $k = \Theta(1/\epsilon^2)$ where $\epsilon \approx n^{-1/3} \|A\|_F^{-1/3}$. Hence, this gives a subexponential-time algorithm in the regime in which the Gibbs measure $\mu(\sigma) \propto e^{f(\sigma)}$ exhibits "mean-field behavior" (i.e. $\|A\|_F^2 \leq o(n)$).

Proof. The objective in Eq. (6) is concave since $\tilde{\mathbb{E}}[f]$ is a linear function in the variables $\tilde{\boldsymbol{p}}$, and $\tilde{H}_k(\tilde{\boldsymbol{p}})$ is concave by Lemma 2.2. These functions can be individually computed in $n^{O(k)}$-time. Since there are $n^{O(k)}$ linear constraints on $\tilde{\boldsymbol{p}}$, Eq. (6) is a convex program which can be solved in $n^{O(k)}$-time (e.g. via ellipsoid method). Finally, the inequality $\mathcal{F}_{\mathsf{SA}(k+2;[n]),k}(f) - \mathcal{F}(f) \geq 0$ follows trivially from Lemma 2.2 and the fact that $\mathsf{SA}(k+2;[n])$ contains the collection of marginals arising from genuine probability distributions.

For the remaining claims, the key is that all of the arguments we used previously to study the mean-field approximation to $\mathcal{F}(f)$ (e.g. low-entropy measure decompositions and the Pinning Lemma) also go through verbatim if we replace $\mathcal{F}(f)$ by $\mathcal{F}_{\mathsf{SA}(k+2;[n]),k}(f)$. To formalize this, let $\tilde{\boldsymbol{p}}$ be the optimal pseudo-distribution attaining $\mathcal{F}_{\mathsf{SA}(k+2;[n]),k}(f)$. From this, we build a mixture of product measures as follows: Fix $S \subseteq [n]$ satisfying $|S| \leq k$, and for each pinning $\tau : S \to \{\pm 1\}$, we let the component measure $\pi^{(\tau)}$ be the unique product measure over $\{\pm 1\}^n$ such that

- $\pi_i^{(\tau)} = \delta_{\tau(i)}$ for all $i \in S$, and
- $\pi_i^{(\tau)} = \tilde{\boldsymbol{p}}_i^{\tau}$ for all $i \notin S$.

In other words, the marginals inside S are specified by the pinning τ, and the marginals outside S are specified by the conditionals of the pseudo-distribution $\tilde{\boldsymbol{p}}$. The mixture measure is simply $\tilde{\boldsymbol{p}}_S$. This is essentially coordinate-by-coordinate localization, adapted to the pseudo-distribution $\tilde{\boldsymbol{p}}$.

Rather than comparing $\mathcal{F}_{\mathsf{SA}(k+2;[n]),k}(f)$ directly with the true free energy $\mathcal{F}(f)$, we instead compare it with $\mathcal{F}_{\mathsf{NMF}}(f)$ (more or less), since these two quantities sandwich $\mathcal{F}(f)$. More specifically, we establish that for the optimal choice $S^* \subseteq [n]$ with $|S^*| \leq k$,

$$\mathcal{F}_{\mathsf{SA}(k+2;[n]),k}(f) - \{\mathbb{E}_{\nu}[f] + H(\nu)\} \leq O\left(\frac{n\|A\|_F}{\sqrt{k}}\right) \qquad \text{where} \qquad \nu = \mathbb{E}_{\tau\sim\tilde{\boldsymbol{p}}_{S^*}}\left[\pi^{(\tau)}\right]. \tag{7}$$

Eq. (7) implies the first claim since $\mathbb{E}_{\nu}[f] + H(\nu) \leq \mathcal{F}(f)$. Eq. (7) also implies the second claim since $\mathcal{F}_{\mathsf{SA}(k+2;[n]),k}(f) \geq \mathcal{F}(f)$, and

$$H(\nu) = H(\tilde{\boldsymbol{p}}_{S^*}) + \mathbb{E}_{\tau\sim\tilde{\boldsymbol{p}}_{S^*}}[\nu^{\tau}] \leq |S^*|\log 2 + \mathbb{E}_{\tau\sim\tilde{\boldsymbol{p}}_{S^*}}\left[H\left(\pi^{(\tau)}\right)\right]$$

by standard maximum entropy considerations, so we can use the rounding $\pi = \pi^{(\tau)}$ for $\tau : S \to \{\pm 1\}$ maximizing $\mathbb{E}_{\pi^{(\tau)}}[f] + H\left(\pi^{(\tau)}\right)$. All that remains is to prove Eq. (7) for a good choice of S^*.

The key observation is that the Pinning Lemma also makes sense for $\mathscr{F}_{k+2}$-pseudo-distributions when pinning up to $k+1$ many coordinates. In particular, by running exactly the same proof, there exists $S^* \subseteq [n]$ of size $|S^*| \leq k$ such that

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

where

$$\widetilde{\mathrm{Cov}}_{\tilde{\boldsymbol{p}}^{\tau}}(\sigma_i,\sigma_j) = \mathbb{E}_{\tilde{\boldsymbol{p}}_{ij}^{\tau}}[\sigma_i\sigma_j] - \mathbb{E}_{\tilde{\boldsymbol{p}}_i^{\tau}}[\sigma_i] \cdot \mathbb{E}_{\tilde{\boldsymbol{p}}_j^{\tau}}[\sigma_j]$$

is the *pseudo-covariance* w.r.t. the pseudo-distribution $\tilde{\boldsymbol{p}}^{\tau}$, which has degree at least 2. Hence, via the same arguments as we did previously in the context of the mean-field approximation (e.g. using the Pinning Lemma, Cauchy–Schwarz, and the fact that f is quadratic) we obtain that

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

for this specific S^*. At the same time, for this choice of S^*, we also have

$$\tilde{H}_k(\tilde{\boldsymbol{p}}) \leq H(\tilde{\boldsymbol{p}}_{S^*}) + \sum_{i\notin S^*} H(\tilde{\boldsymbol{p}}_i \mid \tilde{\boldsymbol{p}}_{S^*}) = H(\nu)$$

just because we minimize over the choice of subset in the definition of $\tilde{H}_k$. Note the second equality holds just by the the Chain Rule for conditional entropy and the fact that each component measure in ν is a product measure with the correct marginals. These two inequalities combined yield Eq. (7) as desired. □

3 Deterministic Algorithm for Bases of Matroids

We now switch to the dual perspective, and take advantage of the following variational representation of the KL-divergence. In the next lecture, we give a more unified presentation of this and the various optimization problems we've already built around the Gibbs Variational Principle.

Theorem 3.1 (Donsker–Varadhan Variational Represetation)**.** *Let μ be a base probability measure on a (finite) state space Ω. Then the function $f \mapsto \log \mathbb{E}_{x\sim\mu}\left[e^{f(x)}\right]$ is smooth and strictly convex (up to shifting by an additive constant). Furthermore, for every probability measure ν on Ω,*

$$\mathscr{D}_{\mathrm{KL}}\left(\nu \,\|\, \mu\right) = \sup_f \left\{ \mathbb{E}_{x\sim\nu}\left[f(x)\right] - \log \mathbb{E}_{x\sim\mu}\left[e^{f(x)}\right] \right\}, \tag{8}$$

and the supremum is uniquely attained at the function $f(x) = \log \frac{\nu(x)}{\mu(x)}$ (up to shifting by an additive constant).

The main result of this section is the following.

Theorem 3.2 ([AOV21])**.** *There is a deterministic algorithm which outputs a 2^r-multiplicative approximation to the number of bases of any matroid $\mathcal{M}$ of rank-r over $[n]$ given by an independence oracle.*

Remark 2. [AOV21] also designed a deterministic $2^{O(r)}$-approximation algorithm for the number of *common* bases in two matroids $\mathcal{M}, \mathcal{N}$ over the same ground set. Again, the algorithm is just to compute

$$\sup_{\boldsymbol{p}\in P_{\mathcal{M}\cap\mathcal{N}}} \left\{ \sum_{i=1}^{n} H\left(\boldsymbol{p}\right)_i \right\},$$

where $P_{\mathcal{M}\cap\mathcal{N}} \overset{\mathsf{def}}{=} \operatorname{conv}\left\{\mathbf{1}_B : B \in \mathscr{B}_{\mathcal{M}} \cap \mathscr{B}_{\mathcal{N}}\right\}$. This can be done because $P_{\mathcal{M}\cap\mathcal{N}} = P_{\mathcal{M}} \cap P_{\mathcal{N}}$, a theorem due to Edmonds, see e.g. [Sch03].

Remark 3. This result is sharp in the sense that no polynomial-time deterministic algorithm can achieve better than $2^{O(n/\log^2(n))}$-multiplicative approximation if it is only allowed access the matroid through an independence oracle [ABF94].

One fruitful way to use Eq. (8), especially towards counting discrete objects satisfying various *hard* combinatorial constraints, is to let μ be some "simple" probability measure over a "nicer" *enlarged* state space, and let ν be the restriction of μ. For instance, for Theorem 3.2, we will let μ be a product measure over all of $2^{[n]}$, and let ν be the induced distribution over the collection $\mathscr{B}$ of bases of the matroid $\mathcal{M}$. This ν is an exponential tilt of the uniform measure over $\mathscr{B}$. With this in mind, we have the following lemma.

Lemma 3.3. *Let $P_{\mathcal{M}} \overset{\mathsf{def}}{=} \operatorname{conv}\{\mathbf{1}_B : B \in \mathscr{B}\}$ be the basis polytope of the matroid $\mathcal{M}$. Then*

$$\log|\mathscr{B}| \le \sup_{\boldsymbol{p}\in P_{\mathcal{M}}} \left\{ \sum_{i=1}^{n} H\left(\boldsymbol{p}_i\right) \right\}.$$

Proof. By subadditivity of entropy, we have

$$\log|\mathscr{B}| = H\left(\nu\right) \le \sum_{i=1}^{n} H\left(\nu_i\right).$$

This can also be derived from Eq. (8) by restricting to linear functions $f(S) = \langle \boldsymbol{v}, \mathbf{1}_S \rangle$ and solving the resulting optimization problem. Unfortunately, we do not know how to compute the marginals ν_i (e.g. by self-reducibility, this is just as hard as computing $|\mathscr{B}|$ itself). However, since $(\mu_i)_{i\in[n]} = \mathbb{E}_{B\sim\mathsf{Unif}\mathscr{B}}[\mathbf{1}_B] \in P_{\mathcal{M}}$, $\sum_{i=1}^{n} H\left(\mu_i\right)$ is upper bounded by $\sup_{\boldsymbol{p}\in P_{\mathcal{M}}} \left\{\sum_{i=1}^{n} H\left(\boldsymbol{p}_i\right)\right\}$ as desired. □

The real heart of the matter is establishing an approximate converse.

Theorem 3.4 ([AOV21])**.** *For every matroid $\mathcal{M}$ of rank-r over $[n]$,*

$$\sup_{\boldsymbol{p}\in P_{\mathcal{M}}} \left\{ \sum_{i=1}^{n} H\left(\boldsymbol{p}_i\right) \right\} \le r + \log|\mathscr{B}|.$$

Proof of Theorem 3.2. Lemma 3.3 and Theorem 3.2 together immediately implies that exponentiating the result of $(*) = \sup_{\boldsymbol{p} \in P_{\mathcal{M}}} \left\{\sum_{i=1}^{n} H(\boldsymbol{p}_i)\right\}$ gives a 2^r-multiplicative approximation to the number of bases $|\mathscr{B}|$. Since the objective is a sum of concave functions, it itself is concave. Furthermore, it is well-known that one can design a separation oracle for the polytope $P_{\mathcal{M}}$ given an independence oracle for $\mathcal{M}$ itself. Hence, we can optimize convex/concave functions over $P_{\mathcal{M}}$, and in particular, compute $(*)$. □

All that remains is to prove Theorem 3.4.

3.1 Entropy Inequalities from Log-Concavity

The key is the following entropy inequality for exponential tilts of the uniform measure over bases of matroids. We will prove it in a moment via log-concavity of the bases generating polynomial.

Proposition 3.5 ([AOV21])**.** *As before, let ν be uniform over $\mathscr{B}$, and let $\boldsymbol{v} \in \mathbb{R}^n$ be any vector. Then*

$$\sum_{i=1}^{n} (\mathcal{T}_{\boldsymbol{v}}\nu)_i \log \frac{1}{(\mathcal{T}_{\boldsymbol{v}}\nu)_i} \leq H(\mathcal{T}_{\boldsymbol{v}}\nu),$$

where recall $(\mathcal{T}_{\boldsymbol{v}}\nu)(B) \propto \prod_{i \in B} e^{v_i}$ over $\mathscr{B}$.

Remark 4. As we will see, this inequality applies to *any* distribution μ over $2^{[n]}$ with a log-concave generating polynomial $\sum_{S \subseteq [n]} \mu(S)\boldsymbol{z}^S$. See [Ali+21] for a generalization to *fractionally log-concave* generating polynomials, which is equivalent to uniform spectral independence under all exponential tilts.

Proof of Theorem 3.4. Observe that by definition by $P_{\mathcal{M}}$, for every $\boldsymbol{p} \in P_{\mathcal{M}}$, there exists some distribution ν' over $\mathscr{B}$ that has $\boldsymbol{p}$ as its marginals. By the Maximum Entropy Principle, i.e. by choosing the ν' maximizing Shannon entropy, there is a unique $\boldsymbol{v} \in \mathbb{R}^n$ such that $(\mathcal{T}_{\boldsymbol{v}}\nu)_i = \boldsymbol{p}_i$ for all $i \in [n]$. It follows that

$$\begin{aligned}
\sup_{\boldsymbol{p} \in P_{\mathcal{M}}} \left\{\sum_{i=1}^{n} H(\boldsymbol{p}_i)\right\} &= \sup_{\boldsymbol{v} \in \mathbb{R}^n} \left\{\sum_{i=1}^{n} H((\mathcal{T}_{\boldsymbol{v}}\nu)_i)\right\} && \text{(Maximum Entropy Principle)} \\
&\leq \sup_{\boldsymbol{v} \in \mathbb{R}^n} \left\{\sum_{i=1}^{n} (\mathcal{T}_{\boldsymbol{v}}\nu)_i \log \frac{1}{(\mathcal{T}_{\boldsymbol{v}}\nu)_i} + \sum_{i=1}^{n} (\mathcal{T}_{\boldsymbol{v}}\nu)_i\right\} && \text{(Using } (1-p)\log \tfrac{1}{1-p} \leq p\text{)} \\
&= r + \sup_{\boldsymbol{v} \in \mathbb{R}^n} \left\{\sum_{i=1}^{n} (\mathcal{T}_{\boldsymbol{v}}\nu)_i \log \frac{1}{(\mathcal{T}_{\boldsymbol{v}}\nu)_i}\right\} && (r\text{-Homogeneity of } \mathscr{B}) \\
&\leq r + \sup_{\boldsymbol{v}} \{H(\mathcal{T}_{\boldsymbol{v}}\nu)\} && \text{(Proposition 3.5)} \\
&= r + \log |\mathscr{B}|. && \text{(Uniform measure maximizes Shannon entropy)}
\end{aligned}$$

□

Proof of Proposition 3.5. Let $g_{\mathcal{M}}(\boldsymbol{z}) \overset{\mathsf{def}}{=} \sum_{B \in \mathscr{B}} \boldsymbol{z}^B$ be the bases generating polynomial of $\mathcal{M}$. Up to some normalization, this is precisely the cumulant generating function of the uniform measure over $\mathscr{B}$ when viewed as a polynomial. Previously, we established that $\log g_{\mathcal{M}}(\boldsymbol{z})$ is a concave function over the entire nonnegative orthant $\mathbb{R}^n_{\geq 0}$, which is equivalent to 0-spectral/entropic independence for $\mathcal{T}_{\boldsymbol{v}}\nu$ for all $\boldsymbol{v} \in \mathbb{R}^n$. Let us now use this log-concavity to prove Proposition 3.5.

For notational convenience, we consider the special case $\boldsymbol{v} = \mathbf{1}$; everything we say will generalize to arbitrary $\boldsymbol{v} \in \mathbb{R}^n$. Letting $\ell_{\mathcal{M}}(\boldsymbol{z}) \overset{\mathsf{def}}{=} \log g_{\mathcal{M}}\left(\frac{z_1}{\nu_1}, \ldots, \frac{z_n}{\nu_n}\right)$, by Jensen's Inequality and log-concavity of $g_{\mathcal{M}}$,

$$\ell_{\mathcal{M}}(\mathbb{E}_{B \sim \nu}[\mathbf{1}_B]) \geq \mathbb{E}_{B \sim \nu}[\ell_{\mathcal{M}}(\mathbf{1}_B)].$$

Since $\mathbb{E}_{B\sim\nu}[\mathbf{1}_B]$ is just the marginal vector $(\nu_i)_{i\in[n]}$, the left-hand side is zero. For the right-hand side, we have

$$\begin{aligned}\mathbb{E}_{B\sim\nu}\left[\ell_{\mathcal{M}}(\mathbf{1}_B)\right] &= \sum_{B\in\mathscr{B}} \nu(B)\log\left(\nu(B)\prod_{i\in B}\frac{1}{\nu_i}\right)\\ &= -H(\nu) + \sum_{i=1}^{n}\left(\sum_{B\ni i}\nu(B)\right)\cdot\log\frac{1}{\nu_i}\\ &= -H(\nu) + \sum_{i=1}^{n}\nu_i\log\frac{1}{\nu_i}.\end{aligned}$$

Putting these inequalities together yields the claim. □

References

[ABF94] Y. Azar, A.Z. Broder, and A.M. Frieze. “On the problem of approximating the number of bases of a matroid”. In: *Information Processing Letters* 50.1 (1994), pp. 9–11. ISSN: 0020-0190 (cit. on p. 5).

[Ali+21] Yeganeh Alimohammadi, Nima Anari, Kirankumar Shiragur, and Thuy-Duong Vuong. “Fractionally Log-Concave and Sector-Stable Polynomials: Counting Planar Matchings and More”. In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. STOC 2021. Virtual, Italy: Association for Computing Machinery, 2021, pp. 433–446 (cit. on p. 6).

[AOV21] Nima Anari, Shayan Oveis Gharan, and Cynthia Vinzant. “Log-concave polynomials, I: Entropy and a deterministic approximation algorithm for counting bases of matroids”. In: *Duke Mathematical Journal* (2021), pp. 1–46 (cit. on pp. 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 p. 3).

[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. 3).

[Sch03] Alexander Schrijver. “Matroid intersection”. In: *Combinatorial Optimization: Polyhedra and Efficiency*. Vol. B. Springer, 2003 (cit. on p. 5).

[WJ08] Martin J. Wainwright and Michael I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Vol. 1. Foundations and Trends in Machine Learning 1–2. Now Publishers Inc, 2008, pp. 1–305 (cit. on p. 2).

[YFW00] Jonathan S Yedidia, William Freeman, and Yair Weiss. “Generalized Belief Propagation”. In: *Advances in Neural Information Processing Systems*. Ed. by T. Leen, T. Dietterich, and V. Tresp. Vol. 13. MIT Press, 2000 (cit. on p. 2).

[YFW05] Jonathan S Yedidia, William Freeman, and Yair Weiss. “Constructing free-energy approximations and generalized belief propagation algorithms”. In: *IEEE Transactions on Information Theory* 51.7 (2005), pp. 2282–2312. DOI: 10.1109/TIT.2005.850085 (cit. on p. 2).

A Examples and Unfinished Proofs

Example 1. Consider, for instance, a degree-2 pseudo-distribution over $\{\pm 1\}^3$ given by letting the single-coordinate marginals be $\tilde{\boldsymbol{p}}_1 = \tilde{\boldsymbol{p}}_2 = \tilde{\boldsymbol{p}}_3 = \mathsf{Unif}\{\pm 1\}$, while the pair marginals $\tilde{\boldsymbol{p}}_{ij}$ have the form

- $\tilde{\boldsymbol{p}}(i \leftarrow +1, j \leftarrow +1) = \tilde{\boldsymbol{p}}(i \leftarrow -1, j \leftarrow -1) = \frac{t_{ij}}{2}$
- $\tilde{\boldsymbol{p}}(i \leftarrow +1, j \leftarrow -1) = \tilde{\boldsymbol{p}}(i \leftarrow -1, j \leftarrow +1) = \frac{1-t_{ij}}{2}$

for three parameters $0 \leq t_{12}, t_{23}, t_{13} \leq 1$.

If $t_{12} > 1/2$, then intuitively $\tilde{\boldsymbol{p}}_{12}$ encourages coordinates 1 and 2 to receive the same ±1 assignment. Conversely, if $t_{12} < 1/2$, then coordinates 1 and 2 are negatively correlated. This already suggests a way to construct a pseudo-distribution which does not arise as the marginals of a globally consistent distribution: Set $t_{12}, t_{23} > 1/2$ (e.g. 1) and $t_{23} < 1/2$ (e.g. 0). The former says that all coordinates $1, 2, 3$ are all positively correlated, while the latter says that coordinates 2 and 3 are negatively correlated, a contradiction. Global consistency would enforce additional inequalities on the parameters t_{12}, t_{23}, t_{13}.