COUNTING BASIS OF MATROIDS VIA ENTROPY MAXIMIZATION

ASHWIN MARAN AND LORENZO NAJT

1. Introduction. In this report for CS880 we study [3], which provides a deterministic approximation algorithm for counting the number of basis of a matroid given by a spanning oracle. To aid students of [3], we provide background on several key lemmas that are not fully explained in their article. In particular, we explain a key lemma (subsection 2.3) on the existence of certain probability distributions. In addition, we provide an implementation (Subsection 5.3) based on the Frank-Wolfe algorithm [4], which we use to investigate some conjectures stated in [3]. Playing with the implementation leads to several simple observations about the behavior of this algorithm.

All mistakes are due to the authors.

2. Definitions of Matroids and Log Concavity. In this section we present some background on the tools used in [3]. Some of these are well-known, such as Cauchy interlacing. For others, such as subsection 2.3, we hope that our exposition will tie together some threads that newcomers to this topic may have difficulty locating.

2.1. Matroids.

DEFINITION 2.1. *A matroid M is a pair $(E, \mathcal{I})$, where E is a finite set and $\mathcal{I} \subset 2^E$ with the following properties:*
1. $\emptyset \in \mathcal{I}$
2. *If $S \in \mathcal{I}$, then $T \in \mathcal{I}$ for all $T \subset S$*
3. *If $S, T \in \mathcal{I}$ such that $|S| > |T|$, then $\exists\ e \in S \backslash T$, such that $T \cup \{e\} \in \mathcal{I}$.*

Any subset of E in $\mathcal{I}$ is called an independent set, and any subset of E that is not in $\mathcal{I}$ is called a dependent set.

DEFINITION 2.2. *A basis of a matroid M is a maximal independent set $S \in \mathcal{I}$ such that $S \cup \{e\} \notin \mathcal{I}$ for all $e \in E \backslash S$. The set of all maximal independent sets in a matroid is called the bases of the matroid and is denoted by $\mathcal{B}_M$.*

An example of a matroid is the graphic matroid. Here, E is the set of edges on a graph G and $\mathcal{I}$ is the family of sets of all edges that form a forest in G. This $(E, \mathcal{I})$ trivially satisfies the first two properties. To prove that the third property also holds true, note that for any $S \in \mathcal{I}$ of size k, the number of connected components in G_S is exactly $n - k$. So, if $S, T \in \mathcal{I}$ such that $|S| > |T|$, then there exists at least one extra connected component in G_T, which means an edge from S can be added to T while keeping it a forest.
In the case of the graphic matroid, the bases is the set of all spanning trees on G.

2.2. Basis Generating Polynomial.

DEFINITION 2.3. *Given a matroid M, following basis generating polynomial is :*

$$g_M(z_1, ..., z_n) = \sum_{B \in \mathcal{B}_M} \prod_{i \in B} z_i$$

This polynomial encodes a lot of information about the matroid and is a natural object to study. For instance,

$$g_M(1, \cdots, 1) = |\mathcal{B}_M| \text{ and } \frac{\partial g_M}{\partial z_i}(1, \cdots, 1) = \#\text{ Bases containing } i$$

In particular, we will be able to study the combinatorics of $\mathcal{B}_M$ by examining analytical properties of g_M as a function.

2.3. External Fields.

In this section we explain the proof of Lemma 2.1. The proof we explain is based on [5], although we are responsible for the errors introduced when filling in the missing details.

DEFINITION 2.4 (External Fields). *Let $\lambda = (\lambda_1, \ldots, \lambda_n)$ be a sequence of positive numbers. Let μ be a probability distribution on $2^{[n]}$. The λ-external field applied to μ is a probability distribution $\lambda \star \mu$ on $2^{[n]}$ where every $S \subseteq [n]$ has mass proportional to $\mu(S) \prod_{i \in S} \lambda_i = \mu(S)\lambda^S$.*

Recall that if μ is a probability distribution on $2^{[n]}$ then P_μ is the convex hull of the indicator vectors of the support of μ: $P_\mu = ConvexHull(1_S : \mu(S) > 0)$.

LEMMA 2.1 (Theorem 2.10 of [3]). *Let $\mu : 2^{[n]} \to \mathbb{R}_{\geq 0}$ be a function. For any point p in the interior of P_μ, there are weights $\lambda \in (\mathbb{R}_{>0})^n$ such that the marginal probabilities of $\lambda \star \mu$ are p. That is, $\mathbb{P}_{S\sim\lambda\star\mu}[i \in S] = p_i$.*

Proof. Consider the function $f(\lambda_1, \ldots, \lambda_n) = \log \frac{g_\mu(\lambda_1,\ldots,\lambda_n)}{\prod \lambda_i^{p_i}}$ We will minimize f subject to $\lambda > 0$. We can make the substitution $\lambda_i = e^{y_i}$, where our domain is now $y_i \in \mathbb{R}$. After doing so, the objective function becomes $\log(\sum_m e^{y\cdot m+\log(\mu(m))}) - \sum p_i y_i$, where m runs over the indicator vectors all the subsets in $\mathrm{supp}(\mu)$. The $\log(\sum_m e^{y\cdot m+\log(\mu(m))})$ term is convex by Lemma 2.2. Since the remaining terms are linear, the objective function f is convex.

We will now show that the minimum value is attained, so that we can learn about the optimal value by the vanishing of the gradient. To prove this, we will calculate the limit in any particular direction.

Without loss, we can assume that the $1 > p_i > 0$, since if $p_i = 0$ in the interior of P_μ, then $i \notin S$ for all $S \in \mathrm{supp}(\mu)$ and if $p_i = 1$ then $i \in S$ for all $S \in \mathrm{supp}(\mu)$, and in either case we could remove i from S. Suppose that $y_i(t) = t\tilde{y}_i$, with $\tilde{y} \cdot p > 0$. Let $\tilde{m} = \mathrm{argmax}_{m\in\mathrm{supp}(\mu)} m \cdot \tilde{y}$, we have that $\tilde{m} \cdot \tilde{y} \geq p \cdot \tilde{y} > 0$, since p is in the relative interior of the convex hull of $\mathrm{supp}(\mu)$. We define $q(t) = \log(\sum e^{t\tilde{y}\cdot m+\log(\mu(m))}) - t\tilde{y} \cdot p$, and aim to show that the limits as $t \to \infty$ and $t \to -\infty$ are both ∞. We use the following well known inequality (Lemma 2.3):

$$\max(x_1, \ldots, x_n) \leq \log(\sum_i \exp(x_i)) \leq \max(x_1, \ldots, x_n) + \log(n)$$

Thus, we have

$$t(\tilde{y} \cdot \tilde{m}) + \log(\mu(\tilde{m})) - t\tilde{y} \cdot p \leq q(t) \leq \log(|\mathrm{supp}(\mu)|) + t(\tilde{y} \cdot \tilde{m}) - t\tilde{y} \cdot p$$

In fact, we have that $\tilde{y}\cdot\tilde{m} > \tilde{y}\cdot p$, since otherwise $q(t)$ is bounded by $\log(\mathrm{supp}(\mu)|)$, and a bounded convex function on a line is constant, which contradicts the strict convexity of $f(y)$. The claim that $\lim_{t\to\infty}(q(t)) = \infty$ follows.

To obtain the other inequality, we observe that there is an m' with $m' \cdot \tilde{y} \leq \tilde{y} \cdot p$. Then we use that $\log(\sum e^{t\tilde{y}\cdot m+\log(\mu(m))}) - t\tilde{y}\cdot p \geq t\tilde{y}\cdot m' - t\tilde{y}\cdot p + \log(\mu(m'))$. Again if $m' \cdot \tilde{y} = \tilde{y} \cdot p$ for all m' with $m' \cdot \tilde{y} \leq p \cdot \tilde{y}$, then we have $\tilde{m} \cdot \tilde{y} = p \cdot \tilde{y}$, which gives the same contradiction as above. Thus, we can pick some with $m' \cdot \tilde{y} < p \cdot \tilde{y}$, and so as $t \to -\infty$, $q(t) \to \infty$.

Now, we see what we learn from the gradient vanishing optimality condition: $0 = \partial_i f = \frac{\partial_i g_\mu}{g_\mu} - \frac{p_i}{\lambda_i}$ That is, at the optimum λ^*, we have $\lambda_i^* \frac{\partial_i g_\mu}{g_\mu}(\lambda_1^*, \ldots, \lambda_n^*) = p_i$. Observing that the left hand term is the same as $\frac{\sum_{B:i\in B}(\lambda^*)^B\mu(B)}{\sum_B \lambda^{*B}\mu(B)} = \mathbb{P}_{\lambda^*\star\mu}(i \in B)$, we conclude the proof.

REMARK 2.5. *We note that if we can efficiently evaluate g_μ, then we can optimize the above program. This, however, is not necessary in the present context.*

LEMMA 2.2. *Let $g_i, i = 1, \ldots n$ be convex and twice differentiable functions on a convex domain D. Then $f = \log(\sum_i \exp(g_i))$ is convex on D. If $g_i \neq g_j$ everywhere on D for some i, j, then f is strictly convex.*

Proof. It suffices to check the one dimensional case, so convexity in D can be checked on all line segments. To prove the one dimensional case, we will compute $f'' > 0$.

By calculation, we have $f' = \frac{\sum_i g_i' \exp(g_i)}{\sum_i \exp(g_i)}$, and $f'' = \frac{(\sum_i [g_i'' \exp(g_i) + (g_i')^2 \exp(g_i)])[\sum_i \exp(g_i)] - [\sum_i g_i' \exp(g_i)]^2}{(\sum_i \exp(g_i))^2}$,

which , since the denominator is positive, has the same sign as the numerator :

$$(\sum_i [g_i'' \exp(g_i) + (g_i')^2 \exp(g_i)])[\sum_i \exp(g_i)] - [\sum_i g_i' \exp(g_i)]^2 \tag{2.1}$$

$$= \sum_{i,j} exp(g_i + g_j)[g_i'' + (g_i')^2 - g_i' g_j'] \tag{2.2}$$

$$= \sum_{i,j} \exp(g_i + g_j) g_i'' + \sum_{i,j} \exp(g_i + g_j)((g_i')^2 - g_i' g_j') \tag{2.3}$$

We have that $0 \leq \sum_{i,j} \exp(g_i + g_j) g_i''$ by the assumption that the g_i'' are convex. Hence, it remains to observe that:

$$\begin{aligned} &\sum_{i,j} \exp(g_i + g_j)((g_i')^2 - g_i' g_j') \\ &= \sum_{\{i,j\} \in \binom{n}{2}} exp(g_i + g_j)[(g_i')^2 - 2g_i' g_j' + (g_j')^2] \\ &= \sum_{\{i,j\} \in \binom{n}{2}} \exp(g_i + g_j)(g_i' - g_j')^2 \geq 0. \end{aligned}$$

□

We note that if $g_i' \neq g_j'$ everywhere, then the last term is > 0.

LEMMA 2.3. *Let $x_1, \ldots, x_n \in \mathbb{R}$. Then*

$$\max(x_1, \ldots, x_n) \leq \log(\sum_i \exp(x_i)) \leq \max(x_1, \ldots, x_n) + \log(n)$$

Proof. $\max(x_1, \ldots, x_n) = \log(\exp(\max(x_1, \ldots, x_n))) \leq \log(\sum \exp(x_i)) \leq \log(n \exp(\max(x_i))) = \log(n) + \max(x_1, \ldots, x_n)$. □

REMARK 2.6. *This perspective is related to the external field construction: https://golem.ph.utexas.edu/category/2016/06/how_the_simplex_is_a_vector_sp.html*

2.4. Cauchy's interlacing theorem. An important tool in this work is Cauchy's interlacing theorem, a key tool in theoretical computer science. For example, this tool also appeared in the recent resolution of the sensitivity conjecture [6]. There are several forms of Cauchy's interlacing theorem, which relate the spectrum of related matrices by showing that their eigenvalues interlace, in the following sense:

DEFINITION 2.7 (Interlacing). *Let $\beta = (\beta_1 \geq \ldots \geq \beta_n)$ and $\alpha = (\alpha_1 \geq \ldots \geq \alpha_n)$ be two sequences of real numbers. We say that β interlaces α if $\alpha_1 \geq \beta_1 \geq \alpha_2 \geq \ldots \geq \beta_{n-1} \geq \alpha_n \geq \beta_n$.*

THEOREM 2.4 (Cauchy Interlacing Theorem I). *For a symmetric matrix $A \in \mathbb{R}^{n \times n}$ and a vector $v \in \mathbb{R}^n$, the eigenvalues of A interlace the eigenvalues of $A + vv^T$.*

Proof. [1] For a matrix B, let $\chi(B)(x) = \det(xI - B)$ denote the characteristic polynomial. Since A is orthogonally diagonalizable, we let $v_1, \ldots, v_n$ be an orthonormal collection of eigenvectors, with $Av_i = \lambda v_i$. Thus, we have (using rank one update formula Lemma 2.5 on the first line):

$$\det(xI - A - vv^T) = \det(xI - A)(1 + v^T (xI - A)^{-1} v) \tag{2.4}$$

$$= \chi(A)(x)(1 + \sum_{i=1}^n \frac{\langle v, v_i \rangle^2}{x - \lambda_i}) \tag{2.5}$$

□

From this, it follows that $\frac{\chi(A + vv^T)(x)}{\chi(A)(x)} = 1 + \sum_{i=1}^n \frac{\langle v, v_i \rangle^2}{x - \lambda_i} = g(x)$. Observe that the poles of $g(x)$ occur at the eigenvalues of A. We first consider the case when A has no repeated eigenvalues.

[1] The proof we present comes from https://windowsontheory.org/2014/04/15/restricted-invertiblity-by-interlacing-polynomials/.

Around each λ_i, $g(x)$ behaves like $\frac{1}{x-\lambda_i}$, so that its limit on the right is $-\infty$ and its limit on the left is $+\infty$. Moreover, $\lim_{x\to\pm\infty} g(x) = 1$. From this description of the limits, it is clear that there is a zero of $g(x)$ between each of the λ_i, and also between $\max \lambda_i$ and ∞. This gives a total of n roots, so these are exactly the roots of $g(x)$. This shows that the eigenvalues of A interlace those of $A + vv^T$, and that the eigenvalues are distinct from those of A.

To handle the case that A has repeated eigenvalues, we observe that $g(x) = 1 + \sum_{i\in K}^{n} \frac{c_i\langle v, v_i\rangle^2}{x-\lambda_i}$, where $c_i \in \mathbb{N}$, and $K \subseteq [n]$. This shows that $\frac{\chi(A+vv^T)(x)}{\chi(A)(x)}$ is a degree $|K|$ rational function whose zeros and poles interlace as before, and the two characteristic polynomials have the remaining roots in common.

LEMMA 2.5 (Rank one update formula). *Let $A \in \mathbb{R}^{n\times n}$ be invertible, and let $v \in \mathbb{R}^n$. Then $det(A + uv^T) = det(A)(1 + v^T A^{-1}u)$.*

Proof. This is the proof we learned from wikipedia. First, it is enough to demonstrate this for $A = I$, since $det(A + uv^T) = \det(A(I + A^{-1}(uv^T)) = \det(A)\det(I + (A^{-1}u)v^T) = \det(A)(1 + v^T A^{-1}u)$. To prove the theorem for $A = I$, we have to calculate that $\det(I + uv^T) = (1 + v^T u)$. This in turn follows from applying multiplicativity of the determinant to:

$$\begin{pmatrix} I & 0 \\ v^T & 1 \end{pmatrix}\begin{pmatrix} I + uv^T & u \\ 0 & 1 \end{pmatrix}\begin{pmatrix} I & 0 \\ -v^T & 1 \end{pmatrix} = \begin{pmatrix} I & u \\ 0 & 1 + v^t u \end{pmatrix}$$

Since our interest is in controlling the number of positive eigenvalues, the following corollary will be useful to us:

COROLLARY 2.8 (Lemma 2.4 in [3]). *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix and let $P \in \mathbb{R}^{m\times n}$. If A has at most one positive eigenvalue, then PAP^T has at most one positive eigenvalue.*

Proof. Since A is real symmetric it is orthogonally diagonalizable, so we can write $A = O^T DO$, where D is diagonal and O is orthogonal. Since A has at most one positive eigenvalue, D has at most one positive entry, which we can assume to be in the $(1, 1)$ slot. We define $B = O^T(D - e_1 e_1^T)O = O^T DO - O^T e_1 e_1^T O = O^T DO - (O^T e_1)(O^T e_1)^T$, and observe that B is negative semidefinite, and that $A = B + vv^T$, for $v = O^T e_1$.

Thus, $PAP^T = PBP^T + Pvv^T P^T$. We have that PBP^T is negative definite, since $x^T PBP^T x = (P^T x)^T B(P^T x) \leq 0$. Since $Pvv^T P^T = (Pv)(Pv)^T$, the eigenvalues of PBP^T interlace the eigenvalues of $PAP^T = PBP^T + (Pv)(Pv)^T$ by Theorem 2.4, meaning that PAP^T has at most one positive eigenvalue. ∎

We also need the following form of Cauchy's interlacing theorem:

THEOREM 2.9 (Cauchy Interlacing II). *Let $A \in \mathbb{R}^{n\times n}$ be symmetric, and let B be a principal submatrix of A. Then the eigenvalues of B interlace the eigenvalues of A.*

Proof. A simple proof follows from Courant-Fischer: https://www.math.uh.edu/~bgb/Courses/Math6304/MatrixTheory-20121011.pdf ∎

Cauchy's interlacing theorem will be used to relate the Hessian of g to the Hessian of $\log(g)$. One of the tools necessary for this is the following lemma:

LEMMA 2.6. *Let $A \in \mathbb{R}^{n\times n}$ be a real symmetric matrix with nonnegative entries and at most one positive eigenvalue. Then for every $v \in \mathbb{R}^n_{\geq 0}$, the $n \times n$ matrix $(v^T Av)A - t(Av)(Av)^T$ is negative semidefinite for all $t \geq 1$.*

Proof. We can assume $v \in \mathbb{R}^n_{>0}$, since the set of negative semidifinite matrices is closed and the result will follow for $v \in \mathbb{R}^n_{\geq 0}$ by taking a sequence $v_n \to v$. We can also assume that $t = 1$, because $(Av)(Av)^T$ is PSD, so subtracting it will preserve negative semidefiniteness (PSD matrices form a cone[2]). WE have that $v^T Av > 0$, because of the non-negativity of the entries. Let $w \in \mathbb{R}^n$, and consider the $2 \times n$ matrix P with rows v^T and w^T. Then $PAP^T = \begin{pmatrix} v^T Av & v^T Aw \\ w^T Av & w^T Aw \end{pmatrix}$

[2]If A, B are PSD, then $x(A + B)x^T = xAx^T + xBx^T \geq 0$.

By Corollary 2.8, PAP^T has at most one positive eigenvalue. Since the minor $[v^TAv]$ is positive, Theorem 2.9 implies that PAP^T has a positive eigenvalue, so it has exactly one. Therefore, its determinant is non-positive, so we have $0 \geq det(PAP^T) = (v^TAv)(w^TAw) - (w^TAv)(v^TAw) = w^T((v^TAv)A - (Av)(v^TA))w$. Since this held for all w, the result follows. ∎

3. Complete log concavity in the rank 2 case.. In this section, we explain a proof of the complete log concavity for the rank 2-case sketched by Nima Anari in [2].

THEOREM 3.1 (Classification of rank-2 simple matroids). *Let $M = (V, I)$ be a simple matroid of rank 2. Let B denote the basis of M, and define a graph by $G(M) = (V, M(I))$. Then $G(M)$ is a complete multipartite graph.*

Proof. It suffices to show that G^c is a union of cliques, since then G is multipartite. Consider $\{a, b\}, \{b, c\} \in E(G^c)$, which are circuits in M. By the circuit axiom, there is a circuit in $\{a, b\} \cup \{b, c\} \setminus b = \{a, c\}$. Since M is simple, and therefore has no loops, this implies that $\{a, c\}$ is a circuit. ∎

Conversely, we have:

THEOREM 3.2. *If $G = (V, E)$ is a complete multipartite graph, then $E \cup V$ form the independent sets of a matroid on ground set V.*

Proof. We verify the basis exchange axiom. ∎

THEOREM 3.3. *If M is a simple rank 2 matroid, and A is the adjacency matrix of $G(M)$, then $g_M = \frac{1}{2}(z_1, \ldots, z_n)A(z_1, \ldots, z_n)^T$. $\nabla^2 g_M = A$, and A at most one positive eigenvalue.*

To apply this we need the following

LEMMA 3.4 (Euler's Identity). *If $g \in \mathbb{R}[z_1, \ldots, z_n]$ is homogeneous of degree d, then $\nabla^2 g \cdot z = (d-1)\nabla g$ and $z^T(\nabla^2 g)z = d(d-1)g$*

Proof. For a homogeneous function g of degree d, we know that $g(tz_1, ..., tz_n) = t^d g(z_1, ..., z_n)$. On taking the derivative with respect to t on both sides,

$$\sum_{i\in[n]} \frac{\partial}{\partial(tz_i)} g(tz_1, ..., tz_n) \cdot z_i = dt^{d-1} g(z_1, ..., z_n)$$

Since this is true for all $t \in \mathbb{R}$, we can substitute $t = 1$ in the above equation to find that $\nabla g \cdot z = d \cdot g(z)$. Consider the i^{th} entry of the vector $\nabla^2 g \cdot z$. $(\nabla^2 g \cdot z)_i = \sum_{j\in[n]} z_j \partial_j \partial_i g(z)$. Since $\partial_i g$ is a $(d-1)$ homogeneous function, $\sum_{j\in[n]} z_j \partial_j \partial_i g(z) = (d-1)\partial g_i(z)$. So,

$$\nabla^2 g \cdot z = (d-1)\nabla g$$

Multiplying both sides by z^T gives us

$$z^T(\nabla^2 g)z = (d-1)(\nabla g \cdot z) = (d-1)(d \cdot g)$$

THEOREM 3.5. *If M is a simple matroid of rank 2, then $g_M(z)$ is log concave.*

Proof. Let M be a simple matroid of rank 2 with the generating polynomial $g \in \mathbb{R}[z_1, ..., z_n]$. Consider

$$\begin{aligned} g^2 \cdot \nabla^2 \log(g) &= [g \cdot \partial_i \partial_j g - \partial_i g \cdot \partial_j g]_{1\leq i,j\leq n} \\ &= g \cdot \nabla^2 g - (\nabla g)(\nabla g)^T \end{aligned}$$

For any arbitrary $\lambda \in \mathbb{R}^n_{>0}$, from Theorem 3.3, we know that $\nabla^2 g|_{z=\lambda}$ has at most one positive eigenvalue. Together with Lemma 2.6, this means that the following is negative semi-definite.

$$(\lambda^T(\nabla^2 g|_{z=\lambda})\lambda)\nabla^2 g|_{z=\lambda} - t\left((\nabla^2 g|_{z=\lambda}\lambda)(\nabla^2 g|_{z=\lambda}\lambda)^T\right)$$

Using Euler's identity and the fact that g is homogenous of degree r, we can simplify the above to be

$$r(r-1) \cdot g(\lambda) \cdot \nabla^2 g|_{z=\lambda} - t(r-1)^2(\nabla g|_{z=\lambda})(\nabla g|_{z=\lambda})^T$$

Therefore, the matrix

$$r(r-1)\left(g\cdot\nabla^2 g - t\cdot\frac{r-1}{r}(\nabla g)(\nabla g)^T\right)\Big|_{z=\lambda}$$

is negative semi-definite. If we now allow $t=\frac{r}{r-1}$, we find that

$$g(\lambda)^2\cdot\nabla^2\log(g)|_{z=\lambda}$$

is semi-definite. Since our choice of λ was arbitrary, $g_M(z)$ is log concave, and follows from results in Matroid Hodge theory. ∎

We will now state the following theorem without proof. This proof follows the same outline as above, but showing that $\nabla^2 g$ has at most one positive eigenvalue for any g is much harder.

THEOREM 3.6 (Anari, Oveis-Gharan, Vinzant). *If M is a matroid, then $g_M(z)$ is log concave.*

In fact, matroids exhibit a much stronger property called complete log concavity. We will say that a polynomial $g\in\mathbb{R}[z_1,...,z_n]$ is completely log concave if for every $k\geq 0$ and non-negative matrix $V\in\mathbb{R}^{n\times k}_{\geq 0}$, $D_V g(z)$ is non-negative and log concave, where

$$D_V g(z)=\left(\prod_{j=1}^{k}\sum_{i=1}^{n}V_{ij}\partial_i\right)g(z)$$

THEOREM 3.7 (Anari, Oveis Gharan, Vinzant). *If M is a matroid, then $g_M(z)$ is completely log concave.*

4. Description of algorithm. In this section we explain the entropy maximization algorithm from [3], and explain the guarantees on its approximation ratio.

4.1. Entropy of Log Concave Distributions. The algorithm relies on several results about marginal entropies that we develop in this section.

DEFINITION 4.1 (Entropy). *The entropy of a distribution μ on $\{0,1\}^n$ is $\mathcal{H}(\mu)=-\sum_{i\in\{0,1\}^n}\mu(i)\log(\mu(i))$.*

If we assume that $\mu_1,...,\mu_n$ are the marginal probabilities of μ, then the following holds true:

LEMMA 4.2.

$$\mathcal{H}(\mu)\leq\sum_{i=1}^{n}\mathcal{H}(\mu_i),\ \text{where } \mathcal{H}(\mu_i)=-\mu_i\log(\mu_i)-(1-\mu_i)\log(1-\mu_i)$$

This is true because the product distribution of $\mathrm{Ber}(\mu_1),...,\mathrm{Ber}(\mu_n)$ will have a higher entropy than any other distribution with the same marginal probabilities.
Just like with matroids, probability distributions on $\{0,1\}^n$ can be characterized by a generating function. For a distribution μ on $\{0,1\}^n$, the generating function is given as:

$$g_\mu(z_1,...,z_n)=\sum_{S\subset\{0,1\}^n}\mu(S)\prod_{i\in S}z_i$$

A distribution is called log-concave if the generating polynomial is a log concave function.

THEOREM 4.3. *For a log concave probability distribution μ on $\{0,1\}^n$ with marginal probabilities $\mu_1,...,\mu_n$,*

$$\mathcal{H}(\mu)\geq-\sum_{i=1}^{n}\mu_i\log(\mu_i)$$

Proof. Jensen's inequality says that when f is concave and X is a $(\mathbb{R}^n_{\geq 0})$ valued random variable,

$$f(\mathbb{E}[X])\geq\mathbb{E}[f(X)]$$

Here, we will let $f(z_1, ..., z_n) = \log\left(g_\mu\left(\frac{z_1}{\mu_1}, ..., \frac{z_n}{\mu_n}\right)\right)$ and $X = \mathbb{1}_S$ where S is chosen randomly according to the distribution μ.
Then $\mathbb{E}[X] = (\mu_1, ..., \mu_n)$ which implies that $f(\mathbb{E}[X]) = 0$.
On the other hand,

$$f(X) = \log\left(\sum_{T\subseteq S}\mu(T)\prod_{i\in T}\frac{1}{\mu_i}\right) \geq \log\left(\mu(S)\prod_{i\in S}\frac{1}{\mu_i}\right) = \log(\mu(S)) - \sum_{i\in S}\log(\mu_i)$$

$$\begin{aligned}\therefore \mathbb{E}[f(X)] &= \sum_S \mu(S)\log(\mu(S)) - \sum_S \mu(S)\sum_{i\in S} log(\mu_i)\\ &= -\mathcal{H}(\mu) - \sum_{i\in[n]}\log(\mu_i)\sum_{S:S\ni i}\mu(S)\\ &= -\mathcal{H}(\mu) - \sum_{i\in[n]}\mu_i\log(\mu_i)\end{aligned}$$

□

Therefore,

$$0 \geq -\mathcal{H}(\mu) - \sum_{i\in[n]}\mu_i\log(\mu_i) \implies \mathcal{H}(\mu) \geq -\sum_{i=1}^n \mu_i\log(\mu_i)$$

Corollary 4.4. *If the log concave distribution μ is such that $|S| = r\ \forall\ S \in supp(\mu)$ (i.e., r-homogeneous), then*

$$\sum_{i=1}^n \mathcal{H}(\mu_i) - r \leq \mathcal{H}(\mu) \leq \sum_{i=1}^n \mathcal{H}(\mu_i)$$

Proof. From Theorem 4.3 and Lemma 4.2,

$$\sum_{i=1}^n \mu_i\log\left(\frac{1}{\mu_i}\right) \leq \mathcal{H}(\mu) \leq \sum_{i=1}^n \mathcal{H}(\mu)$$

$$\therefore \sum_{i=1}^n \mathcal{H}(\mu_i) + \sum_{i=1}^n (1-\mu_i)\log(1-\mu_i) \leq \mathcal{H}(\mu) \leq \sum_{i=1}^n \mathcal{H}(\mu)$$

Note that for all $p \in (0,1), (1-p)\cdot\log\left(\frac{1}{1-p}\right) \leq p$. So,

$$-\sum_{i=1}^n (1-\mu_i)\log(1-\mu_i) \leq \sum_{i=1}^n \mu_i = \mathbb{E}[|S|] = r$$

$$\therefore \sum_{i=1}^n \mathcal{H}(\mu_i) - r \leq \mathcal{H}(\mu) \leq \sum_{i=1}^n \mathcal{H}(\mu_i)$$

Corollary 4.5. *If the distribution μ and μ^* are both log concave, then*

$$\frac{1}{2}\sum_{i=1}^n \mathcal{H}(\mu_i) \leq \mathcal{H}(\mu) \leq \sum_{i=1}^n \mathcal{H}(\mu_i)$$

Proof. From Theorem 4.3,

$$\mathcal{H}(\mu) \geq -\sum_{i=1}^n \mu_i\log(\mu_i) \text{ and } \mathcal{H}(\mu^*) \geq -\sum_{i=1}^n (1-\mu_i)\log(1-\mu_i)$$

Since $\mathcal{H}(\mu) = \mathcal{H}(\mu^*)$, this means that

$$\mathcal{H}(\mu) \geq \frac{1}{2}\left(-\sum_{i=1}^n \mu_i\log(\mu_i) + (1-\mu_i)\log(1-\mu_i)\right) = \frac{1}{2}\sum_{i=1}^n \mathcal{H}(\mu_i)$$

Now, for any matroid M with bases $\mathcal{B}_M$, if we let μ be the uniform distribution on the bases, then by definition,

$$\mathcal{H}(\mu) = \log(|\mathcal{B}_M|)$$

Moreover,

$$g_\mu(z) = \frac{1}{|\mathcal{B}_M|} \sum_{B \in \mathcal{B}_M} \prod_{i \in B} z_i = \frac{1}{|\mathcal{B}_M|} g_M(z)$$

Since M is a matroid, Theorem 3.6 implies that $g_M(z)$ is log concave. Since $g_\mu(z) = \frac{1}{|\mathbb{B}_M|} g_M(z)$, it is also log-concave. $g_M(z)$ is log concave, which implies that μ is a log concave distribution. Since all bases of a matroid have the same size, μ is $r - homogeneous$, where r is the rank of the matroid. Additionally, μ^* is the uniform distribution on the dual matroid M^*. So, μ^* is also log concave.
So, from Corollary 4.4 and Corollary 4.5,

$$\max\left(\frac{1}{2}\sum_{i=1}^{n}\mathcal{H}(\mu_i), \sum_{i=1}^{n}\mathcal{H}(\mu_i) - r\right) \leq \log(|\mathcal{B}_M|) \leq \sum_{i=1}^{n}\mathcal{H}(\mu_i)$$

where μ is the uniform distribution on the bases of the matroid M.

4.2. The Algorithm. Unfortunately, estimating $\sum_i \mathcal{H}(\mu_i)$ is not any easier than estimating $\mathcal{H}(\mu)$. Instead we will estimate $\sum_i \mathcal{H}(p_i^*)$, where $p^* = \arg\max_{p \in \mathcal{P}_M} \sum_i \mathcal{H}(p_i)$, where $\mathcal{P}_M$ is the matroid polytope. Since $\mu \in \mathcal{P}_M$, it follows that:

$$\sum_{i \in [n]} \mathcal{H}(p_i^*) \geq \sum_{i \in [n]} \mathcal{H}(\mu) \geq \log(|\mathcal{B}_M|)$$

Now, from Lemma 2.1, we can see that there exists a probability distribution $\tilde{\mu} = \lambda * \mu$ for some λ such that $\tilde{\mu}$ has the marginal probabilities of p^*. By construction, $\tilde{\mu}$ preserves the log concavity of μ. So, once again applying Corollary 4.4 and Corollary 4.5,

$$\max\left(\frac{1}{2}\sum_{i \in [n]}\mathcal{H}(p_i^*), \sum_{i \in [n]}\mathcal{H}(p_i^*) - r\right) = \max\left(\frac{1}{2}\sum_{i \in [n]}\mathcal{H}(\tilde{\mu}_i), \sum_{i \in [n]}\mathcal{H}(\tilde{\mu}_i) - r\right) \leq \log(|\mathcal{B}_M|)$$

$$\therefore \max\left(\frac{1}{2}\sum_{i=1}^{n}\mathcal{H}(p_i^*), \sum_{i=1}^{n}\mathcal{H}(p_i^*) - r\right) \leq \log(|\mathcal{B}_M|) \leq \sum_{i=1}^{n}\mathcal{H}(p_i^*)$$

This means that the optimization of $\arg\max_{p \in \mathcal{P}_M} \sum_i \mathcal{H}(p_i)$ provides the promised approximation algorithm. Since $\mathcal{P}_M$ has a separation oracle, this concave maximization problem can be solved in polynomial time. In the next section we describe our implementation of it.

5. Implementation. For our implementation, we study the behavior of the entropy maximiation algorithm on forests and matchings. As suggested by Nima Anari [1], instead of basing our concave maximization implementation on the separation oracle, we will base it on the linear optimization oracle. That is, we will exploit the fact that we can quickly determine a max weight forest or max weight matching, in order to optimize over the corresponding polytope. This connection is made possible by the Frank-Wolfe algorithm [4], which we will explain in the following section.

5.1. Frank-Wolfe Algorithm. Let $P \subseteq \mathbb{R}^n$ be a polytope, and let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable convex function. We wish to solve

$$\min_{x \in P} f(x) \tag{5.1}$$

We assume that we have an oracle that can solve the following family of problems: For any linear function l,

$$\min_{x \in P} l(x) \tag{5.2}$$

We note that if P is given by a system of inequalities $Ax \leq b$, or by a separation oracle, then we can solve this via linear programming. However, we can sometimes solve (5.2) via other algorithms - we will refer to such algorithms as linear optimization oracles. We will explain two such oracles in the next section. For now, we explain how to use such a separation oracle to optimize, via the Frank-Wolfe algorithm. Frank-Wolfe is also sometimes called the conditional gradient method.

The idea of the Frank-Wolfe algorithm is simple and natural: we will use the gradient of f, ∇f to linearize the problem, and use the linear separation oracle to determine a vertex of the polytope that optimizes the linearized problem. That is, we consider the linear function $L_y(x) = f(y) + \langle \nabla f(y), x - y \rangle$, which is the first order approximation to f at the point y. We now optimize L_y over P, to find an optimal boundary point of P to move towards, and pick a point between the current point and the target point. There are several schemes for choosing that intermediate point. Each of these schemes will calculate some $\lambda_t \in (0, 1)$ at each stage, and move to $(1 - \lambda_t)$current point $+ \lambda_t$target point.

That is:

DEFINITION 5.1 (Frank-Wolfe Algorithm).
1. *Initialize with* $x_0 \in P$.
2. *Compute* $s_t \in argmin_{s \in P} \langle s, \nabla f(x_t) \rangle$
3. *Set* $x_{t+1} = (1 - \lambda_t)x_t + \lambda_t s_t$.

Since the gradient of the entropy is infinite on $\{x_i = 0, x_j = 1 : i \in I, j \in J\}$, to apply Frank-Wolfe to entropy maximization we will need to avoid points on the coordinate subspaces. It suffices to initialize with a point in the interior of P. For spanning trees, this is possible as long as the graph contains no loop or bridge edge. For matchings, this is possible as long as every edge is in some matching, and no edge is in every matching.

In the remainder of this section, we explain the convergence of the Frank-Wolfe algorithm under the update rule $\lambda_t = \frac{2}{2+t}$. We follow http://people.csail.mit.edu/stefje/fall15/notes_lecture14.pdf

5.1.1. Convergence.

DEFINITION 5.2 (Curvature constant). *The curvature constant of a convex function g is the smallest C such that, for all $w, w' in P$ and $\eta \in [0, 1]$, $g(w + \eta(w' - w)) \leq g(w) + \eta \langle \nabla g(w), w' - w \rangle + \frac{C}{2}\eta^2$.*

THEOREM 5.3. *Consider a differentiable, convex function f on a convex domain P. Let $w^* \in argmin_{x \in P} f(x)$. Let w_t be the point returned on the tth step of the Frank-Wolfe algorithm,with $\lambda_t = \frac{2}{2+t}$. Then $g(w_t) - g(w^*) \leq \frac{C}{t+2}$.*

Proof. We have :

$$g(w_{t+1}) = g((1 - \lambda_t)w_t + \lambda_t s_t) \tag{5.3}$$

$$= g(w_t + \lambda_t(s_t - w_t)) \tag{5.4}$$

$$\leq g(w_t) + \lambda_t \langle \nabla g(w_t), s_t - w_t \rangle + \frac{C}{2}\lambda_t^2 \tag{5.5}$$

$$\leq g(w_t) + \lambda_t \langle \nabla g(w_t), s_t - w_t \rangle + \frac{C}{2}\lambda_t^2 \tag{5.6}$$

$$\leq g(w_t) + \lambda_t \langle \nabla g(w_t), w^* - w_t \rangle + \frac{C}{2}\lambda_t^2 \tag{5.7}$$

$$= (1 - \lambda_t)g(w_t) + \lambda_t g(w_t) + \lambda_t \langle \nabla g(w_t), w^* - w_t \rangle + \frac{C}{2}\lambda_t^2 \tag{5.8}$$

$$\leq (1 - \lambda_t)g(w_t) + \lambda_t g(w^*) + \frac{C}{2}\lambda_t^2 \tag{5.9}$$

(4)to (5) follows because s_t minimizes the linearization. (6) to (7) follows because of convexity, that is : $g(w_t) + \langle \nabla g(w_t), w^* - w_t \rangle = L_{w_t}(w^*) \leq g(w^*)$, i.e. the graph of the first order approximation is always below the true function.

From this, obtain $\lambda_t g(w_t) - \lambda_t g(w^*) \leq \frac{C}{2}\lambda_t^2$, so $g(w_t) - g(w^*) \leq \frac{C}{2}\lambda_t = \frac{C}{2+t}$. ∎

For the negative entropy function on $[0, 1]$, we have that $C = \infty$, as the second derivative is $\frac{1}{p(1-p)}$. Therefore, when we apply the Frank-Wolfe algorithm to the entropy maximization problem on a polytope, we

do not obtain bounds on the convergence. However, this seems to work reasonably well in our implementation, where we started with a random interior point obtained by averaging random basis obtained by applying the greedy algorithm to random weights.

5.2. Linear Optimization Oracles for Matchings and Forests. In this section we explain the two linear optimization oracles for matchings and for forests. The implementations that we use are built into the python library networkx. Our code is here: https://github.com/LorenzoNajt/EntropyCounting

5.2.1. Matchings. For matchings, we use an implementation of Edmonds Blossom algorithm.

5.2.2. Forests. For forests, we use the greedy algorithm. Indeed, the greedy algorithm works for any Matroid given by an independence oracle.

5.3. Empirical Observations. We record our observations about the code here, and link to the repository.

5.3.1. Some Data:. Spanning trees on a $k \times k$ grid graph:

k	[Entropy, True]
3	[2076, 192]
4	[7861602, 100352]
5	[491358645942, 557568000]

Perfect matchings on a $2k \times k$ grid graph: [3]

k	[log(Entropy) - log(True)]/V
12	0.805526983963
13	0.809016320395
14	0.80973110506

5.3.2. Symmetry. For any graph where the symmetry group is edge transitive , the optimum value and the true marginals will agree: this is because in that case the vector $(n-1)/E$ is in the polytope, so the entropy maximization upper bound from that point is met.

We noticed that the optimal values founds by entropy maximization tended to have a lot more symmetry than the underlying graph. We don't know what to make of this. For example, when we run the entropy maximization on the 5×5 grid graph, we obtain get that every edge weight is about .600.

REFERENCES

[1] N. Anari, *Email correspondence.*

[2] N. Anari, *Geometry of polynomials boot camp: Completely log-concave polynomials and distributions, part i - b*, https://simons.berkeley.edu/talks/tba-25.

[3] N. Anari, S. O. Gharan, and C. Vinzant, *Log-concave polynomials, entropy, and a deterministic approximation algorithm for counting bases of matroids*, in 2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS), IEEE, 2018, pp. 35–46.

[4] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval research logistics quarterly, 3 (1956), pp. 95–110.

[5] S. O. Gharan, *Geometry of polynomials boot camp, real stable polynomials, strongly rayleigh distributions, and applications, part ii - a*, https://simons.berkeley.edu/talks/tba-16.

[6] H. Huang, *Induced subgraphs of hypercubes and a proof of the sensitivity conjecture*, Annals of Mathematics, 190 (2019), pp. 949–955.

[3] Bipartite matching is a matroid intersection. Their theory extends to this case.