Information theory in combinatorics

January 14, 2015

1 Basic definitions

Logarithms are in base 2.
Entropy: $H(X) = \sum_x \Pr[X = x] \log(1/\Pr[X = x])$.
For $0 \le p \le 1$ we shorthand $H(p) = p\log(1/p) + (1-p)\log(1/(1-p))$.
Conditional entropy: $H(X|Y) = \sum \Pr[Y = y] H(X|Y = y) = H(X, Y) - H(Y)$.
Chain rule: $H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_1, \ldots, X_{n-1})$.
Independence: If $X_1, \ldots, X_n$ are independent then $H(X_1, \ldots, X_n) = \sum H(X_i)$.
Basic inequalities:

- $H(X) \ge 0$.
- $H(X|Y) \le H(X)$ and $H(X|Y, Z) \le H(X|Y)$.
- If X is supported on a universe of size n then $H(X) \le \log n$, with equality if X is uniform.

2 Shearer's lemma

Shearer's lemma is a generalization of the basic inequality $H(X_1, \ldots, X_n) \le \sum H(X_i)$. For $S \subseteq [n]$ we shorthand $X_S = (X_i : i \in S)$.

Lemma 2.1 (Shearer)**.** *Let $X_1, \ldots, X_n$ be random variables. Let $S_1, \ldots, S_m \subseteq [n]$ be subsets such that each $i \in [n]$ belongs to at least k sets. Then*

$$k \cdot H(X_1, \ldots, X_n) \le \sum_{j=1}^{m} H(X_S).$$

Proof. By the chain rule

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_1, \ldots, x_{n-1}).$$

If $S_j = \{i_1, \ldots, i_{s_j}\}$ with $i_1 < \ldots < i_{s_j}$ then

$$H(X_{S_j}) = H(X_{i_1}) + H(X_{i_2}|X_{i_1}) + \ldots + H(X_{i_{s_j}}|X_{i_1}, \ldots, X_{i_{s_j-1}})$$
$$\leq H(X_{i_1}|X_1, \ldots, X_{i_1-1}) + H(X_{i_2}|X_1, \ldots, X_{i_2-1}) + \ldots$$

The lemma follows since each term $H(X_i|X_1, \ldots, X_{i-1})$ appears k times in the LHS and at least k times in the RHS. □

The following is an equivalent version, which is sometimes more convenient.

Lemma 2.2 (Shearer; distribution)**.** *Let $X_1, \ldots, X_n$ be random variables. Let $S \subseteq [n]$ be a random variable, such that $\Pr[X_i \in S] \geq \mu$ for all $i \in [n]$. Then*

$$\mu \cdot H(X_1, \ldots, X_n) \leq \mathbb{E}_S[H(X_S)].$$

3 Number of graph homomorphisms

Example 3.1. *Let $P \subset \mathbb{R}^3$ be a set of points whose projection on each of the XY, YZ, XZ planes have at most n points. How many points can P have? We can have $|P| = n^{3/2}$ if P is a grid of size $\sqrt{n} \times \sqrt{n} \times \sqrt{n}$. We will show that this is tight by applying Shearer's lemma. Let (X, Y, Z) be a uniform point in P. Then $H(X, Y, Z) = \log |P|$. On the other hand, by Shearer's lemma applied to the sets $\{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$,*

$$2H(X, Y, Z) \leq H(X, Y) + H(X, Z) + H(Y, Z) \leq 3 \log n.$$

Hence $\log |P| \leq H(X, Y, Z) \leq \frac{3}{2} \log n$.

This is an instance of a more general phenomena. Let G, T be undirected graphs. A homomorphism of T to G is $\sigma : V(T) \to V(G)$ such that $(u, v) \in E(T) \Rightarrow (\sigma(u), \sigma(v)) \in E(G)$. Let $\mathrm{Hom}(T, G)$ be the family of all homomorphisms from T to G. Our goal will be to bound $|\mathrm{Hom}(T, G)|$.

A *fractional independent set* of T is a mapping $\psi : V(T) \to [0, 1]$ such that for each edge $(u, v) \in E(T)$, $\psi(u) + \psi(v) \leq 1$. The fractional independent set number of T is the maximum size (eg $\sum \psi(v)$) of a fractional independent set, denoted $\alpha^*(T)$. It is given by a linear program, whose dual is the following. A *fractional cover* of T is a mapping $\phi : E(T) \to [0, 1]$ such that for each vertex $v \in V(T)$, $\sum_{(u,v) \in E(T)} \phi(u, v) \geq 1$. The fractional cover number of T is the minimum size (eg $\sum \phi(e)$) of a fractional cover of T. It is equal to $\alpha^*(T)$ by linear programming duality.

Theorem 3.2 (Alon [2], Freidgut-Kahn [6])**.** $|\mathrm{Hom}(T, G)| \leq (2|E(G)|)^{\alpha^*(T)}$.

This implies as a special case the previous example (up to constants). Let G be a tri-partite graph with parts X, Y, Z. For every point $(x, y, z) \in P$ add the edges $(x, y), (y, z), (x, z)$ to G. Then $|E(G)| \leq 3n$. Let $T = \triangle$, where $\alpha^*(\triangle) = 3/2$. Then

$$6|P| \leq |\mathrm{Hom}(\triangle, G)| \leq (6n)^{3/2}.$$

One can also show that the bound is essentially tight for fixed T, as there exist graphs G for which $|\mathrm{Hom}(T, G)| \geq (|E(G)|/|E(T)|)^{\alpha^*(T)}$. We will not show this here.

Proof. Let $\sigma : T \to G$ be a uniform homomorphism in $\mathrm{Hom}(T, G)$. If $v_1, \ldots, v_n$ are the vertices of T, then set $X_i = \sigma(v_i)$. We have $H(X_1, \ldots, X_n) = \log |\mathrm{Hom}(T, G)|$. Let ϕ be a fractional cover of T with $\sum \phi(e) = \alpha^*(T)$. Let $S \in E(T)$ be chosen with probability $\Pr[S = \{u, v\}] = \phi(u, v)/\alpha^*(T)$. Note that $S \subset [n]$, with $\Pr[i \in S] \geq 1/\alpha^*(T)$. Also, $H(X_S) \leq \log(2|E(G)|)$ since if $S = \{u, v\}$ then (X_u, X_v) is distributed over directed edges of G. By Shearer's lemma,

$$\log |\mathrm{Hom}(T, G)| = H(X_1, \ldots, X_n) \leq \alpha^*(T) \cdot \mathbb{E}_S[H(X_S)] \leq \alpha^*(T) \cdot \log(2|E(G)|).$$

□

4 Number of independent sets

Let G be a d-regular graph on n vertices. How many independent sets can G have? Let $\mathcal{I}(G)$ denote the family of all independent sets $I \subset V(G)$.

Theorem 4.1 (Kahn [8])**.** *If G is bi-partite then*

$$|\mathcal{I}(G)| \leq (2^{d+1} - 1)^{\frac{n}{2d}}.$$

This is tight: take G to be the union of $n/2d$ copies of $K_{d,d}$. The result was extended to general d-regular graphs by Zhao [11].

Proof. Assume $V(G) = [n]$, and let $A \cup B = [n]$ be a partition so that $E(G) \subset A \times B$, where we assume $|A| \geq |B|$. Let $I \subset [n]$ be a uniform independent set, and set $X_i = 1_{i \in I}$. Then $\log |\mathcal{I}(G)| = H(X_1, \ldots, X_n)$. We shorthand $X_A = \{X_i : i \in A\}, X_B = \{X_i : inB\}$. We have

$$H(X_1, \ldots, X_n) = H(X_A) + H(X_B|X_A).$$

For each $b \in B$ let $N(b) \subset A$ be the neighbors of b. Let $Q_b = [I \cap N(b) = \emptyset]$ be the event that non of the neighbors of b are in I, and let $q_v = \Pr[Q_v]$. We first bound the second term,

$$H(X_B|X_A) \leq \sum_{b \in B} H(X_b|X_A) \leq \sum_{b \in B} H(X_b|X_{N(b)}) \leq \sum_{b \in B} H(X_b|Q_b).$$

Note that $H(X_b|Q_b) = q_b \cdot H(X_b|Q_b = 1) \leq q_b$, since $\overline{Q_b} \Rightarrow X_b = 0$ and $X_b \in \{0, 1\}$, hence

$$H(X_B|X_A) \leq \sum_{b \in B} q_b.$$

Next we bound $H(X_A)$. Note that the sets $N(b)$ cover each element of A exactly d times, hence by Shearer's lemma,

$$H(X_A) \leq \frac{1}{d} \sum_{b \in B} H(X_{N(b)}).$$

We can bound

$$H(X_{N(b)}) = H(X_{N(b)}|Q_b) + H(Q_b) \le (1-q_b)\log(2^d-1) + H(q_b).$$

Combining these estimates, we obtain

$$\begin{aligned} H(X_1,\dots,X_n) &\le \sum_{b\in B} q_b + \frac{1}{d}\sum_{b\in B}\big(H(q_b) + (1-q_b)\log(2^d-1)\big) \\ &= \frac{n}{2d}\log(2^d-1) + \frac{1}{d}\sum_{b\in B}\left(H(q_b) + q_b\log\frac{2^d}{2^d-1}\right) \end{aligned}$$

Differentiation gives that $H(x) + x\log\frac{2^d}{2^d-1}$ is maximized at $x_0 = \frac{2^d}{2^{d+1}-1}$, hence

$$H(X_1,\dots,X_n) \le \frac{n}{2d}\left(\log(2^d-1) + H(x_0) + x_0\log\frac{2^d}{2^d-1}\right) = \frac{n}{2d}\log(2^{d+1}-1).$$

□

5 Weighted version, and applications

The following is a combinatorial version of Shearer's lemma. A hypergraph $H = (V, E)$ is simply a family of subsets $E \subset 2^V$.

Lemma 5.1 (Shearer; hypergraphs). *Let H be a hypergraph. Let $S_1,\dots,S_m \subset V$ be subsets of vertices, such that each $v \in V$ belongs to at least k subsets. Define the projected hypergraph H_i with $V(H_i) = S_i$ and $E(H_i) = \{e \cap S_i : e \in E\}$. Then*

$$|E(H)|^k \le \prod |E(H_i)|.$$

Proof. Let $|V(H)| = n$, $X_1,\dots,X_n \in \{0,1\}$ be the indicator of a uniform edge $e \in E$. Then $H(X_1,\dots,X_n) = \log|E(H)|$ and $H(X_{V(H_i)}) \le \log|E(H_i)|$, since $X_{V(H_i)}$ is a random variable supported on $E(H_i)$. □

Freidgut proved a weighted version of Shearer's lemma. Let $w_i : E(H_i) \to \mathbb{R}_{\ge 0}$ be some nonnegative weight function. For $e \in E$ let $e_i = e \cap S_i \in E(H_i)$.

Theorem 5.2 (Weighted Shearer lemma, Freidgut [5]). *Under the same conditions,*

$$\left(\sum_{e\in E(H)}\prod_{i=1}^m w_i(e_i)\right)^k \le \prod_{i=1}^m \sum_{e_i\in E(H_i)} w_i(e_i)^k.$$

Corollary 5.3. *For any $n \times n$ matrices A, B, C,*

$$\mathrm{Tr}(ABC)^2 \le \mathrm{Tr}(AA^t)\cdot\mathrm{Tr}(BB^t)\cdot\mathrm{Tr}(CC^t).$$

Proof. We need to prove:

$$\left(\sum A_{i,j}B_{j,k}C_{k,i}\right)^2 \leq \sum A_{i,j}^2 \cdot \sum B_{j,k}^2 \cdot \sum C_{k,i}^2.$$

Clearly, we may assume all entries of A, B, C are nonnegative.

Let H be a complete tri-partite hypergraph with 3 parts I, J, K of size n each. Let H_1, H_2, H_3 be the projected graphs to $I \cup J, J \cup K, I \cup K$, respectively. Each vertex of H belongs to two of the projected graphs. Define weights (on 2-edges) by

$$w(i,j) = A_{i,j}, w(j,k) = B_{j,k}, w_{k,i} = C_{k,i}.$$

Then

$$\sum_{e \in E(H)} w_1(e_1)w_2(e_2)w_3(e_3) = \sum_{i,j,k} A_{i,j}B_{j,k}C_{k,i}$$

and (for example)

$$\sum_{e \in E(H_1)} w_1(e_1)^2 = \sum A_{i,j}^2.$$

□

6 Read-k functions

Let $x \in \{0,1\}^n$ be uniform bits. Let $f_1, \ldots, f_m : \{0,1\}^n \to \{0,1\}$ be boolean functions, where each f_i depends only on variables in some set $S_i \subset [n]$. Assume furthermore that $\Pr[f_i = 1] = p$. If the sets $S_1, \ldots, S_m$ are pairwise disjoint then $f_i(x)$ are independent, and in particular

$$\Pr[f_1(x) = \ldots = f_m(x) = 1] = p^m.$$

Shearer's lemma allows us to extend this to the case where there is limited intersections.

Definition 6.1 (read-k functions)**.** *The functions $f_1, \ldots, f_m$ are said to be read-k if each x_i participates in at most k functions. That is, $|\{j : i \in S_j\}| \leq k$ for all $i \in [n]$.*

Lemma 6.2. *If $f_1, \ldots, f_m$ are read-k with $\Pr[f_i = 1] = p$ then*

$$\Pr[f_1(x) = \ldots = f_m(x) = 1] \leq p^{m/k}.$$

Proof. Let $q = \Pr[f_1(x) = \ldots = f_m(x) = 1]$. We may assume wlog that each x_i is contained in *exactly* k sets. Let $A = \{x \in \{0,1\}^n : f_1(x) = \ldots = f_m(x) = 1\}$ and $A_i = \{x \in \{0,1\}^{S_i} : f_i(x) = 1\}$. We have $|A| = q2^n$ and $|A_i| = p2^{|S_i|}$. Let $(X_1, \ldots, X_n) \in A$ be uniformly distributed. By Shearer's lemma,

$$k \cdot H(X_1, \ldots, X_n) \leq \sum H(X_{A_i}).$$

The lemma follows since $H(X_1, \ldots, X_n) = \log |A| = \log q + n$ and $H(X_{A_i}) \le \log |A_i| = \log p + |S_i|$. Hence

$$k(\log q + n) \le m \cdot \log p + \sum |S_i| = m \cdot \log p + kn.$$

□

For example, if $G = G(n, 1/2)$ is a random graph on n vertices, and E_v is some event which depends only on the edges touching a vertex v, then

$$\Pr[\forall v \; E_v] \le \prod \Pr[E_v]^{1/2}.$$

The power $1/2$ is tight. For example, choose a maximal matching M on $\{1, \ldots, n\}$ (n even) and let E_v be the event "the unique edge in M which touches v appears in G".

We prove here an analog of the Chernoff bound for read-k functions. Recall that if $Y_1, \ldots, Y_m \in \{0, 1\}$ are independent, with $\Pr[Y_i = 1] = p$, then Chernoff bound tell us that

$$\Pr[Y_1 + \ldots + Y_m \ge (p + \varepsilon)m] \le \exp(-2\varepsilon^2 m).$$

Theorem 6.3 (Gavinsky-Lovett-Saks-Srinivasan [7])**.** *If $f_1, \ldots, f_m$ are read-k with $\Pr[f_i = 1] = p$ then*

$$\Pr[f_1(x) + \ldots + f_m(x) \ge (p + \varepsilon)m] \le \exp(-2\varepsilon^2 m/k).$$

The proof uses the Kullback-Leibler divergence between distributions.

Definition 6.4. *Let μ, μ' be two distributions on the same domain. The KL-divergence between them is defined as*

$$D_{\mathrm{KL}}(\mu \,||\, \mu') = \sum \mu(x) \log \frac{\mu(x)}{\mu'(x)}.$$

If X, X' are random variables distributed like μ, μ' then $D_{\mathrm{KL}}(X \,||\, X') = D_{\mathrm{KL}}(\mu \,||\, \mu')$.

Fact 6.5.

(i) $D_{\mathrm{KL}}(X \,||\, X') \ge 0$.

(ii) For any function ϕ, $D_{\mathrm{KL}}(\phi(X) \,||\, \phi(X')) \le D_{\mathrm{KL}}(X \,||\, X')$.

(iii) If X is supported on a set A, and U is uniform on A, then $D_{\mathrm{KL}}(X \,||\, U) = H[U] - H[X]$.

(iv) Let U be uniform over a set A. Let $A' \subset A$ with $|A'| = p|A|$. Let X be any random variable of A with $\Pr[X \in A'] = q$. Then

$$D_{\mathrm{KL}}(X \,||\, U) \ge D_{\mathrm{KL}}(q \,||\, p),$$

where $D_{\mathrm{KL}}(q \,||\, p) = q \log \frac{q}{p} + (1 - q) \log \frac{1-q}{1-p}$.

Lemma 6.6 (Shearer lemma for KL divergence)**.** *Let $X_1, \dots, X_n$ be random variables. Let $U_1, \dots, U_n$ be independent random variables, where U_i is uniform over a set containing the support of X_i. Let $S_1, \dots, S_m \subset [n]$ be such that each $i \in [n]$ belongs to at most k sets. Then*

$$k \cdot D_{\mathrm{KL}}(X_1, \dots, X_n \,||\, U_1, \dots, U_n) \geq \sum D_{\mathrm{KL}}(X_{S_i} \,||\, U_{S_i}).$$

Proof. We may assume wlog that each $i \in [n]$ belongs to exactly k sets. Hence by Shearer's lemma, $k \cdot H(X_1, \dots, X_n) \leq \sum H(X_{S_i})$. Now apply fact (iii).

$$\begin{aligned} k \cdot D_{\mathrm{KL}}(X_1, \dots, X_n \,||\, U_1, \dots, U_n) &= kH(U_1, \dots, U_n) - kH(X_1, \dots, X_n) \\ &= k \sum H(U_i) - kH(X_1, \dots, X_n) \end{aligned}$$

and

$$\sum D_{\mathrm{KL}}(X_{S_i} \,||\, U_{S_i}) = \sum H(U_{S_i}) - H(X_{S_i}) = k \sum H(U_i) - \sum H(X_{S_i}).$$

□

Proof of Theorem 6.3. Let

$$A = \{x \in \{0,1\}^n : f_1(x) + \ldots + f_m(x) \geq (p + \varepsilon)m\}.$$

Let $X \in A$ be uniformly distributed, and let $U \in \{0,1\}^n$ be uniform. We have

$$\log \Pr[f_1(x) + \ldots + f_m(x) \geq (p + \varepsilon)m] = \log \frac{|A|}{2^n} = H[X] - H[U] = -D_{\mathrm{KL}}(X \,||\, U).$$

Let X_{S_i}, U_{S_i} be the restrictions of X, U to S_i, respectively. Then by Shearer's lemma for KL divergence,

$$k \cdot D_{\mathrm{KL}}(X \,||\, U) \geq \sum D_{\mathrm{KL}}(X_{S_i} \,||\, U_{S_i}).$$

Let $A_i = \{0,1\}^{S_i}$ and let $A'_i = \{x \in A_i : f_i(x) = 1\}$. Then $|A'_i| = p|A_i|$, and U_{S_i} is uniform on A_i. Let $q_i = \Pr[X_i \in A_i]$. Hence by fact (iv),

$$D_{\mathrm{KL}}(X_{S_i} \,||\, U_{S_i}) \geq D_{\mathrm{KL}}(q_i \,||\, p).$$

By convexity of the KL divergence function, we have

$$D_{\mathrm{KL}}(X \,||\, U) \geq \frac{1}{k} \sum_{i=1}^{m} D_{\mathrm{KL}}(q_i \,||\, p) \geq \frac{m}{k} D_{\mathrm{KL}}(q \,||\, p),$$

where $q = (q_1 + \ldots + q_m)/m$. By assumption, any X satisfies $f_i(X) = 1$ for at least $(p+\varepsilon)m$ indices $i \in [m]$, hence

$$q_1 + \ldots + q_m = \sum \Pr[X_i \in A_i] = \sum \mathbb{E}[1_{X_i \in A_i}] = \sum \mathbb{E}[f_i(X)] = \mathbb{E}\left[\sum f_i(X)\right] \geq (p + \varepsilon)m.$$

Hence $q \geq p + \varepsilon$, and we conclude that

$$\log \Pr[f_1(x) + \ldots + f_m(x) \geq (p + \varepsilon)m] \leq -D_{\mathrm{KL}}(X \,||\, U) \leq -(m/k) \cdot D_{\mathrm{KL}}(p + \varepsilon \,||\, p).$$

The bound

$$\Pr[f_1(x) + \ldots + f_m(x) \geq (p + \varepsilon)m] \leq \exp(-2\varepsilon^2 m/k)$$

follows from $2^{-D_{\mathrm{KL}}(p+\varepsilon \,||\, p)} \leq \exp(-2\varepsilon^2)$. □

7 Moore bound in irregular graphs

Let G be a d-regular graph on n vertices with girth g. We assume here throughout that $g = 2r+1$ is odd, although the results can be extended to even girth. Moore's bound gives a lower bound on n:

$$n \geq 1 + d \sum_{i=0}^{r-1} (d-1)^i.$$

The proof is simple: fix a vertex $v \in V(G)$. Let $n_i(v)$ be the number of vertices of distance i from v, for $i = 0, \ldots, r$. The number of non backtracking paths of length $i \geq 1$ from v is $n_i(v) = d(d-1)^{i-1}$, and they all must lead to distinct vertices by the girth assumption. Hence, $n \geq n_0(v) + \ldots + n_r(v)$.

Alon, Hoory and Linial extended this bound to the case where the average degree is d.

Theorem 7.1 (Alon-Hoory-Linial [3])**.** *Let G be a graph on n vertices with average degree d and girth $g = 2r+1$. Then*

$$n \geq 1 + d \sum_{i=0}^{r-1} (d-1)^i.$$

We present an information theoretic proof due to Ajesh Babu and Radhakrishnan [1]. In the proof, we may assume that the minimum degree is 2, as removing vertices of degree 1 can only increase the average degree, and does not change the girth.

Proof. Let $d_v = \deg(v)$. Let π be a distribution on vertices given by $\pi(v) = \frac{d_v}{2|E|}$. We will prove: $\mathbb{E}_{v \sim \pi}[n_i(v)] \geq d(d-1)^{i-1}$, and the theorem follows. To prove that, let $v \sim \pi$ and sample a uniform non backtracking path of length i from v, which we denote $v = v_0, v_1, \ldots, v_i$. That is, v_1 is a uniform neighbor of v, and for $j \geq 1$, v_{j+1} is a uniform neighbor of v_j other than v_{j-1}. We make two observations: each vertex v_j is distributed according to π; and each edge (v_j, v_{j+1}) is a uniform directed edge in G. Now,

$$\begin{aligned}
\log \mathbb{E}[n_i(v)] &\geq \mathbb{E}[\log n_i(v)] \\
&\geq H[v_1, \ldots, v_i | v] \\
&= H[v_1|v] + H[v_2|v, v_1] + \ldots + H[v_i | v, v_1, \ldots, v_{i-1}] \\
&= \mathbb{E}\left[\log d_v + \sum_{j=1}^{i-1} \log(d_{v_j} - 1)\right] \\
&= \mathbb{E}\left[\log\left\{d_v (d_v - 1)^{i-1}\right\}\right] \\
&= \frac{1}{dn} \sum_v d_v \log\left\{d_v (d_v-1)^{i-1}\right\} \\
&\geq \frac{1}{d} \cdot d \log\left\{d(d-1)^{i-1}\right\} = \log\left\{d(d-1)^{i-1}\right\},
\end{aligned}$$

where the last inequality follows from the convexity of the function $x \log(x(x-1)^{i-1})$ for $x \geq 2$. □

8 Brégman theorem: bounding the permanent

Let A be an $n \times n$ matrix with $0,1$ entries. The permanent of A is $\sum_{\pi \in S_n} A_{i,\pi(i)}$. Minc conjectured, and Brégman proved, the following theorem.

Theorem 8.1 (Brégman's theorem [4])**.** *Let $d_1, \ldots, d_n$ be the row sums of A. Then*

$$\mathrm{per}(A) \leq \prod (d_i!)^{1/d_i}.$$

It is tight, eg if $d_1 = \ldots = d_n = d$ and A consists of n/d blocks of size $d \times d$ of all ones. We present an entropy based proof due to Radhakrishnan [9].

Proof. Let $P = \{\pi \in S_n : A_{i,\pi(i)} = 1 \; \forall i \in [n]\}$. Then $|P| = \mathrm{per}(A)$. Let $\pi \in P$ be uniformly chosen, and consider the random variable $(\pi(1), \ldots, \pi(n))$. We have

$$\begin{aligned}\log |P| &= H(\pi(1), \ldots, \pi(n)) \\ &= H(\pi(1)) + H(\pi(2)|\pi(1)) + \ldots + H(\pi(n)|\pi(1), \ldots, \pi(n-1)).\end{aligned}$$

Consider the i-th term in the sum. Let $D_i = \{j : A_{i,j} = 1\}$ with $|D_i| = d_i$, and consider some fixing of $\pi(1) = x_1, \ldots, \pi(i-1) = x_{i-1}$. Then $\pi(i)$ can take any value in $D_i \setminus \{x_1, \ldots, x_{i-1}\}$, and hence $H(\pi(i)|\pi(1) = x_1, \ldots, \pi(i-1) = x_{i-1}) \leq \log |D_i \setminus \{x_1, \ldots, x_{i-1}\}|$. It is not clear how to evaluate this directly. The trick is to enumerate the rows in a random order.

For $\sigma \in S_n$ and consider the random variable $\pi(\sigma(1)), \ldots, \pi(\sigma(n))$. We have

$$H(\pi) = H(\pi(\sigma(1))) + H(\pi(\sigma(2))|\pi(\sigma(1))) + \ldots + H(\pi(\sigma(n))|\pi(\sigma(1)), \ldots, \pi(\sigma(n-1)))$$

Averaging over uniformly chosen $\sigma \in S_n$, we get

$$H(\pi) = \mathbb{E}_\sigma \sum_{i=1}^{n} H(\pi(\sigma(i))|\pi(\sigma(1)), \ldots, \pi(\sigma(i-1))).$$

(note: we think of σ as a fixed permutation, and not a random variable. Equivalently, we can condition also on σ in the entropy calculations). Letting $k_{\sigma,i} = \sigma^{-1}(i)$, we can reorder the terms as

$$\begin{aligned}H(\pi) &= \sum_{i=1}^{n} \mathbb{E}_\sigma H(\pi(i)|\pi(\sigma(1)), \ldots, \pi(\sigma(k_{\sigma,i} - 1))) \\ &\leq \sum_{i=1}^{n} \mathbb{E}_{\pi,\sigma} \log |D_i \setminus \{\pi(\sigma(1)), \ldots, \pi(\sigma(k_{\sigma,i} - 1))\}| \\ &= \sum_{i=1}^{n} \mathbb{E}_{\pi,\sigma} \log |\pi^{-1}(D_i) \setminus \{\sigma(1), \ldots, \sigma(k_{\sigma,i} - 1)\}|.\end{aligned}$$

Fix π, and consider the i-th term. For all $\pi \in P$ we have $\pi(i) \in D_i$, and hence $i \in \pi^{-1}(D_i)$. Consider the ordering of $\pi^{-1}(D_i)$ induced by σ. The set $\pi^{-1}(D_i) \cap \{\sigma(1), \ldots, \sigma(k_{\sigma,i} - 1)\}$

is the set of all elements of $\pi^{-1}(D_i)$ which appear before i; moreover, as σ is uniform, the ordering of $\pi^{-1}(D_i)$ by σ is uniform, and hence

$$\Pr_{\sigma}[|\pi^{-1}(D_i) \setminus \{\sigma(1), \ldots, \sigma(k_{\sigma,i}-1)\}| = j] = \frac{1}{d_i} \quad \forall j = 1, \ldots, d_i.$$

We thus conclude

$$H(\pi) \leq \sum_{i=1}^{n} \sum_{j=1}^{d_i} \frac{\log j}{d_i} = \log \prod_{i=1}^{n} (d_i!)^{1/d_i}.$$

□

9 Spencer theorem

Let A be an $n \times n$ matrix with $0, 1$ entries. If $x \in \{-1, 1\}^n$ is chosen uniformly, then whp $|(Ax)_i| \leq O(\sqrt{n})$; however the largest entry can be of the order of $\sqrt{n \log n}$. While this is true for most x, Spencer proved that there exist x for which $|(Ax)_i| \leq O(\sqrt{n})$ for all $i \in [n]$.

Theorem 9.1 (Spencer [10])**.** *For any $n \times n$ matrix A with $0, 1$ entries, there exists $x \in \{-1, 1\}^n$ such that $\|Ax\|_\infty \leq O(\sqrt{n})$.*

The main idea is to find a *partial coloring*: a partial solution $x \in \{-1, 0, 1\}^n$ such that $\|Ax\|_\infty \leq O(\sqrt{n})$, and such that a constant fraction of the coordinates of x are in $\{-1, 1\}$. Then, we recurse upon the uncolored (set to zero) variables. The error terms form a geometric sequence (almost), and hence sum to $O(\sqrt{n})$. Here we will just describe this partial coloring lemma.

Lemma 9.2 (partial coloring lemma)**.** *For any $n \times n$ matrix A with $0, 1$ entries, there exists $x \in \{-1, 0, 1\}^n$ such that*

1. $\|Ax\|_\infty \leq O(\sqrt{n})$.
2. *At least $n/4$ (say) of the coordinates of x are in $\{-1, 1\}$.*

Proof. Let $C \geq 1$ be a constant to be determined later. We will find $x', x'' \in \{-1, 1\}^n$ such that $\|Ax' - Ax''\|_\infty \leq C\sqrt{n}$, and such that x', x'' disagree on $n/4$ of the coordinates. Then setting $x = (x' - x'')/2$ gives the required solution. To this end, let $X \in \{-1, 1\}^n$ be uniformly chosen, and consider the random variables $Y_i(X) = \lfloor (AX)_i / C\sqrt{n} \rfloor$ for $i \in [n]$. Standard estimates show that $\Pr[Y_i \geq t] \leq \exp(-\Omega(C^2 t^2))$, and in particular if we choose C a large enough constant, we get $H(Y_i) \leq 1/4$. Hence

$$H(Y_1, \ldots, Y_n) \leq \sum_{i=1}^{n} H(Y_i) \leq n/4.$$

In particular, there must be some values $y_1, \ldots, y_n$ such that $\Pr[Y_1 = y_1, \ldots, Y_n = y_n] \geq 2^{-n/4}$. Let $S = \{x \in \{-1, 1\}^n : Y_i(x) = y_i \ \forall i \in [n]\}$. Then $|S| \geq 2^{3n/4}$, and for any $x', x'' \in S$ we have $\|Ax' - Ax''\|_\infty \leq C\sqrt{n}$. To conclude the lemma, observe that any subset of $\{0, 1\}^n$ of size $2^{3n/4}$ must contain two points which disagree on at least $n/4$ coordinates. □

References

[1] S. Ajesh Babu and J. Radhakrishnan. An entropy based proof of the moore bound for irregular graphs. *Arxiv-eprints*, pages 1–6, 2010.

[2] N. Alon. On the number of subgraphs of prescribed type of graphs with a given number of edges. *Israel Journal of Mathematics*, 38(1-2):116–130, 1981.

[3] N. Alon, S. Hoory, and N. Linial. The moore bound for irregular graphs. *Graphs and Combinatorics*, 18(1):53–57, 2002.

[4] L. Bregman. Some properties of nonnegative matrices and their permanents. In *Soviet Math. Dokl*, volume 14, pages 945–949, 1973.

[5] E. Friedgut. Hypergraphs, entropy, and inequalities. *American Mathematical Monthly*, pages 749–760, 2004.

[6] E. Friedgut and J. Kahn. On the number of copies of one hypergraph in another. *Israel Journal of Mathematics*, 105(1):251–256, 1998.

[7] D. Gavinsky, S. Lovett, M. Saks, and S. Srinivasan. A tail bound for read-k families of functions. *Random Structures & Algorithms*, 2014.

[8] J. Kahn. An entropy approach to the hard-core model on bipartite graphs. *Combinatorics, Probability and Computing*, 10(03):219–237, 2001.

[9] J. Radhakrishnan. An entropy proof of bregman's theorem. *journal of combinatorial theory, Series A*, 77(1):161–164, 1997.

[10] J. Spencer. Six standard deviations suffice. *Transactions of the American Mathematical Society*, 289(2):679–706, 1985.

[11] Y. Zhao. The number of independent sets in a regular graph. *Combinatorics, Probability and Computing*, 19(02):315–320, 2010.