• Applications: "Entropy & counting" has many recent applications in theoretical computer science.

Entropy, Optimization and Counting

Mohit Singh
Microsoft Research, Redmond, USA
Email: mohits@microsoft.com

Nisheeth K. Vishnoi
Microsoft Research, Bangalore, India
Email: nisheeth.vishnoi@gmail.com

← STOC 2014

Abstract

In this paper we study the problem of computing max-entropy distributions over a discrete set of objects subject to observed marginals. Interest in such distributions arises due to their applicability in areas such as statistical physics, economics, biology, information theory, machine learning, combinatorics and, more recently, approximation algorithms. A key difficulty in computing max-entropy distributions has been to show that they have polynomially-sized descriptions. We show that such descriptions exist under general conditions. Subsequently, we show how algorithms for (approximately) counting the underlying discrete set can be translated into efficient algorithms to (approximately) compute max-entropy distributions. In the reverse direction, we show how access to algorithms that compute max-entropy distributions can be used to count, which establishes an equivalence between counting and computing max-entropy distributions.

Log-Concave Polynomials I: Entropy and a Deterministic Approximation Algorithm for Counting Bases of Matroids

STOC 2019 Best paper

Nima Anari[1], Shayan Oveis Gharan[2], and Cynthia Vinzant[3]

[1]Computer Science Department, Stanford University, anari@cs.stanford.edu
[2]Computer Science and Engineering, University of Washington, shayan@cs.washington.edu
[3]Department of Mathematics, North Carolina State University, clvinzan@ncsu.edu

November 6, 2018

The Entropy Rounding Method in Approximation Algorithms

Thomas Rothvoß*

M.I.T.
rothvoss@math.mit.edu

October 15, 2018

Big break-through after 44 years.

A (Slightly) Improved Approximation Algorithm for Metric TSP

Anna R. Karlin,* Nathan Klein,† and Shayan Oveis Gharan‡

University of Washington

September 1, 2020

shortest abstract ever!

Abstract

For some $\epsilon > 10^{-36}$ we give a $3/2 - \epsilon$ approximation algorithm for metric TSP.

§ Application: Bounding the binomial tail.

PHP: There are n movies, 2^n+1 people.
Each person i watch a subset of movies S_i ($|S_i| \geq 1$).
Then there are two people who have watched the same subset. [2^n+1 pigeons, 2^n holes].

Q. There are $2n$ movies, 2^n people. Each person has watched 90% of the movies [$|S_i| \geq \frac{9}{10} \cdot 2n$]
Then there are two people who have watched the same subset.

Proof: We would like to compute the number of possible S_is with $|S_i| \geq \frac{9}{10} \cdot 2n$.

i.e. $\displaystyle\sum_{i=\frac{9}{10}2n}^{2n} \binom{2n}{i} = \sum_{j=0}^{\frac{1}{10}2n} \binom{2n}{j}$ [$\because \binom{2n}{i} = \binom{2n}{2n-i}$].

Now we claim:

$$\boxed{\text{If } k \leq n/2, \text{ then } \sum_{i=0}^{k} \binom{n}{i} \leq 2^{n \cdot H(k/n)}}$$

Assuming the claim,

$$\sum_{j=0}^{\frac{1}{10}2n} \binom{2n}{j} \leq 2^{2n \cdot H\left(\frac{2n/10}{2n}\right)}$$

$H(0.1) \approx 0.47 < 0.5$
$\Rightarrow 2nH(0.1) < n.$

$$= 2^{2n \cdot H(0.1)} < 2^n.$$

Using PHP, we get the solution.

• Proof of claim:

$$\text{If } k \leq n/2, \text{ then } \sum_{i=0}^{k} \binom{n}{i} \leq 2^{n \cdot H(k/n)}.$$

$\Rightarrow$ $X_1 \dots X_n$ be a uniformly random string sampled from the set of n-bit strings with at most k 1's.

$$\therefore H(X_1, \dots, X_n) = \log\left(\sum_{i=0}^{k} \binom{n}{i}\right). \quad \dots (A)$$

Now X_i's can be think of Bernoulli RVs with

$Pr(X_i = 1) \leq k/n$. [$\because$ All X_i's are symmetric, Total k of them are 1]

Thus $H(X_i) = H(p)$ for $p \leq k/n$.

As, $k \leq n/2$, $p \leq k/n \leq 1/2$.

As $H(p)$ is an increasing fn for $p \in (0, \frac{1}{2}]$,

$$H(X_i) \leq H(k/n). \quad \dots (C)$$

Hence,

$\therefore H(X_1 X_2 \dots X_n)$

$\leq \sum_{i=1}^{n} H(X_i)$ [From subadditivity]

$\leq n \cdot H(X_i)$ [From symmetry]

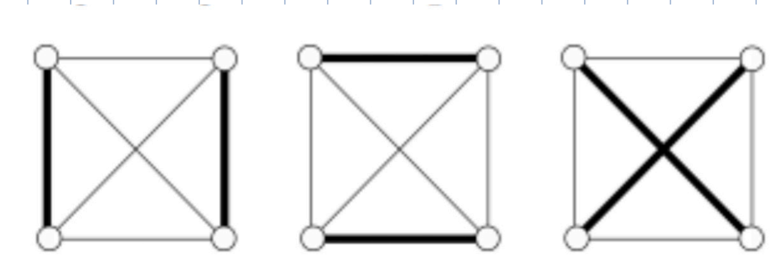
$\leq n \cdot H(k/n)$. [From (C)]

Thus, from (A),

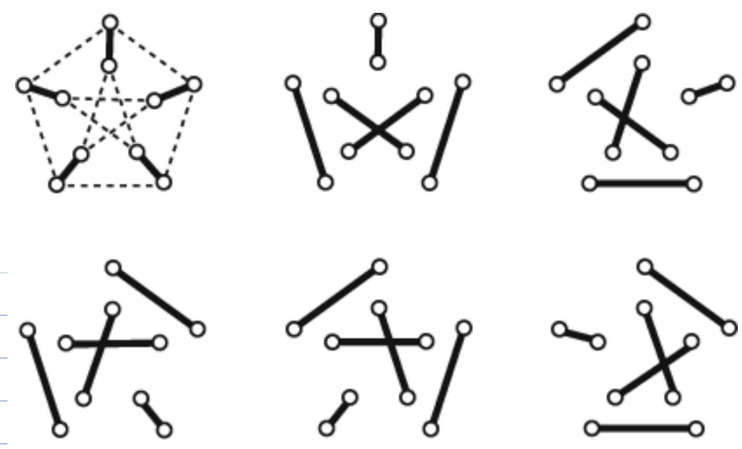
$$\log\left(\sum_{i=0}^{k} \binom{n}{i}\right) \leq n \cdot H(k/n).$$ ■

§ Application : Counting Perfect Matchings.

Perfect matching : set of edges where every vertex has exactly one edge incident on it.



K_4 has 3 perfect matchings.

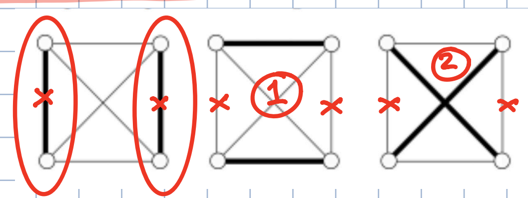


Petersen graph has six perfect matchings such that every edge is contained in precisely two of these perfect matchings.

- Let d_v = degree of vertex v.

- **Theorem [Brégman]** : Let $G := (A, B, E)$ be a bipartite graph with $|A| = |B| = n$. Then the number of perfect matchings in G is at most $\prod_{v \in A} (d_v!)^{1/d_v}$.

- Note : This is tight. Take $K_{n,n}$. Any bijections can be chosen. Such no. is $n! = \prod_{i=1}^{n} (n!)^{1/n}$.



$K_{2,2}$ has $2! = 2$ PMs.

Proof : [By Radhakrishnan '97]

- obvious bound : # perf. matchings $\leq \prod_{v \in A} d_v$.

We can justify this by entropy as well.

Let Σ be the set of perfect matchings (PM).

Let $\sigma \in \Sigma$ be a uniformly random PM.

Here, $\sigma : A \to B$.

Let $\sigma(v_i)$ be the neighbor of $v_i \in A$ in σ. So basically σ is a permutation of vertices in B.

$$\therefore \log|\Sigma| \overset{①}{=} H(\sigma) \overset{②}{=} H[\sigma(v_1)] + H[\sigma(v_2)|\sigma(v_1)] + \cdots + H[\sigma(v_n)|\sigma(v_1)\ldots\sigma(v_n)]$$

entropy & counting

chain rule

conditioning reduces entropy

$$\overset{③}{\leq} \sum_{i=1}^{n} H[\sigma(v_i)]$$

Now, use entropy & counting again:

$$\sum_{i=1}^{n} H[\sigma(v_i)] \overset{④}{\leq} \sum_{i=1}^{n} \log d_i \qquad \left[\because \text{support size of } \sigma(v_i) \text{ is } \underset{:=d_i}{d(v_i)}\right]$$

$$= \log\left[\prod_{v\in A} d_v\right].$$

Can we improve further? Ineq ③ seems lossy.

E.g. consider a term on LHS of ③:

$H[\sigma(v_i)|\sigma(v_1)\ldots\sigma(v_{i-1})]$ measures uncertainty in $\sigma(v_i)$ after $\sigma(v_1)\ldots\sigma(v_{i-1})$ has been revealed. We use $H[\sigma(v_i)]$ as upper bound without using the information from $\sigma(v_j)$'s for $j \in [i-1]$.

For example, $\sigma(v_i) \notin \{\sigma(v_1),\ldots,\sigma(v_{i-1})\}$.

Hence, number of possibilities of $\sigma(v_i)$ is not $d(v_i)$ but $|N(v_i) - \{\sigma(v_1),\ldots,\sigma(v_{i-1})\}| := R_\sigma(i)$.

However, we have no way of knowing (or controlling) how many neighbors of v_i have been used when $\sigma(v_i)$ is revealed.

Idea: Choose a random order to examine vertices of A, rather than a deterministic order.

To exploit this observation, we pick a random permutation $\pi : [n] \to A$ and examine σ in this order determined by π.

Then $|N(v_i) - \{\sigma(\pi(v_1)), \dots, \sigma(\pi(v_{i-1}))\}| =: R_{\sigma.\pi}(i)$ depends on how $N(v_i)$ are ordered by $\sigma.\pi$.

Since π is a random permutation $|N(v_i) \cap \{\sigma(\pi(v_1)), \dots, \sigma(\pi(v_{i-1}))\}|$ is equally likely to be any number in $\{1, \dots, d_i\}$.

Thus, $\Pr_{\pi}[R_{\sigma.\pi}(i) = j] = 1/d_i$ for $j \in [d_i]$. 4

• Now we show an useful inequality.

Lemma 1:

Let (X, Y) be a pair of random variables. Let support(X) can be partitioned into sets $A_1, \dots, A_r$ s.t. $\forall\, i \in [r]$ and $x \in A_i$, $|\text{support}(Y_x)| \le i$, then

$$H(Y|X) \le \sum_{i=1}^{r} \Pr[X \in A_i] \log i.$$

Note: support(X) is the set of values X takes with positive probability. $Y_x := Y | X = x$.

Proof: $H(Y|X) \overset{\text{def}}{=} \mathbb{E}_x[H[Y_x]]$

$$= \sum \Pr[X \in A_i]\, H[Y_x \mid x \in A_i]$$

$$\le \sum_{i=1}^{r} \Pr[X \in A_i] . \log i$$

support size $\le i$. entropy $\le \log i$

Fix $i \in [n]$ and a permutation π.

Let $K = \pi^{-1}(i)$ [i.e. $\pi(K) = i$].

Now we study the expression:

$$H[\sigma] = H[\sigma(\pi(1))] + H[\sigma(\pi(2)) \mid \sigma(\pi(1))] + \dots$$
$$+ H[\sigma(\pi(n)) \mid \sigma(\pi(1)) \dots \sigma(\pi(n-1))].$$

By averaging over all π, we obtain

$$H[\sigma] = \mathbb{E}_{\pi}\Big[H[\sigma(\pi(1))] + H[\sigma(\pi(2)) \mid \sigma(\pi(1))] + \dots$$
$$+ H[\sigma(\pi(n)) \mid \sigma(\pi(1)) \dots \sigma(\pi(n-1))]\Big].$$

Let us collect contributions of different $\sigma(i)$ separately.

$i = \pi(K)$

$$H[\sigma] = \sum_{i=1}^{n} \mathbb{E}_{\pi}\big[H[\sigma(i) \mid \sigma(\pi(1)) \dots \sigma(\pi(K-1))]\big]$$

Breaking into diff. support sizes of $\sigma(i)$

$$\leq \sum_{i=1}^{n} \mathbb{E}_{\pi}\Big[\sum_{j=1}^{d_i} \Pr_{\sigma}[|R_{\sigma,\pi}(i)| = j] \cdot \log j\Big] \quad \text{[From Lemma 1]}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{d_i} \mathbb{E}_{\pi}\Big[\Pr_{\sigma}[|R_{\sigma,\pi}(i)| = j] \cdot \log j\Big]$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{d_i} \Pr_{\pi,\sigma}[|R_{\sigma,\pi}(i)| = j] \cdot \log j$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{d_i} \frac{1}{d_i} \log j \quad \text{[From (4)]}$$

$$= \sum_{i=1}^{n} \log (d_i!)^{1/d_i} = \log \Big(\prod_{i=1}^{n} d_i!\Big)^{1/d_i}$$

$\Rightarrow |\Sigma| = \left(\prod_{i=1}^{n} d_i!\right)^{1/d_i}$, which completes the proof ∎

§ Application: Shearer's lemma.

Puzzle:

> Suppose n distinct points in $\mathbb{R}^3$ have n_1 distinct projections on the XY-plane, n_2 .. on XZ-plane and n_3 ... on YZ plane.
>
> Then $n^2 \leq n_1 n_2 n_3$.

Proof: A trivial observation: $n \leq n_1 n_2 n_3$. $(\circ, \circ, \circ)$ $\downarrow n_1 \; \downarrow n_3 \; \downarrow n_2$

For the stronger bound, we use entropy.

Let $P = (x, y, z)$ be one of the n points picked at random with uniform distribution.

So by definition, $P_1 = (x, y)$, $P_2 = (x, z)$, $P_3 = (y, z)$ are its three projections.

Now,

$$+\begin{cases} H[P_1] \overset{\text{chain rule}}{=} H[x] + H[y|x] \\ H[P_2] = H[x] + H[z|x] \\ H[P_3] = H[y] + H[z|y] \end{cases}$$

seems tailor-made

$$H[P_1] + H[P_2] + H[P_3] = 2H[x] + H[y] + H[y|x] + H[z|x] + H[z|y]$$

$$\Rightarrow 2H[P] \overset{\text{chain rule}}{=} 2H[x] + 2H[y|x] + 2H(z|xy)$$

$$\leq 2H[x] + \underbrace{H[y] + H[y|x]} + \underbrace{H(z|x) + H(z|y)}$$

$$= H[P_1] + H[P_2] + H[P_3]. \qquad (*)$$

Now, $H[P] = \log n$ [$\because$ uniform distr.], and $H[P_i] \leq \log n_i$ for $i \in [3]$ as P_i can take at most n_i values.

Thus from $(*)$,

$$\Rightarrow 2\log n \leq \log n_1 + \log n_2 + \log n_3 \quad \left[\text{relating entropy \& support size}\right]$$

$$\Rightarrow n^2 \leq n_1 n_2 n_3.$$

• **Shearer's Lemma:** Let $X = X_1, \dots, X_n$ be a RV. If S is any distribution on subsets of $[n]$, s.t. $\forall\, i \in [n], \Pr[i \in S] \geq \mu$; then

$$\boxed{\mathbb{E}[H(X_S)] \geq \mu \cdot H(X).}$$

• Here X_S is projection of X onto the coordinates in S, i.e. $X_S := X_{i_1}, X_{i_2}, \dots, X_{i_k}$ when $S = \{i_1, \dots, i_k\}$. We define $X_{<i} := X_1, \dots, X_{i-1}$.

[Note: it generalizes subadditivity: $\sum_{i=1}^{n} H(X_i) \geq H(X)$. i.e. $\mathbb{E}[H(X_i)] \geq H(X)/n$. Informally, this says average coordinate carries at least average entropy]

Proof: Let $T = \{i_1, \dots, i_k\}$ with $i_1 < i_2 < \dots < i_k$. Then

$$H(X_T) \underset{\text{chain rule}}{=} H(X_{i_1}) + H(X_{i_2} | X_{i_1}) + \dots + H(X_{i_k} | X_{i_1} \dots X_{i_{k-1}}).$$

$$\underset{\substack{\text{conditioning} \\ \text{can't increase}}}{\geq} H(X_{i_1}) + H(X_{i_2} | X_{<i_2}) + \dots + H(X_{i_k} | X_{<i_k}).$$

$$\Rightarrow \mathbb{E}_S[H(X_S)] \geq \mathbb{E}_S\Big[\sum_{i \in S} H(X_i | X_{<i})\Big]$$

$$= \mathbb{E}_S\Big[\sum_{i \in [n]} \mathbb{1}_S(i) \cdot H(X_i | X_{<i})\Big]$$

[$\mathbb{1}_S$ is indicator function for set S]

$$= \sum_{i \in [n]} \Big[\mathbb{E}_S[\mathbb{1}_S(i) \cdot H(X_i | X_{<i})]\Big]$$

$$= \sum_{i \in [n]} \Pr[i \in S] \cdot H(X_i | X_{<i})$$

$$\geq \mu \sum_{i \in [n]} H(X_i | X_{<i}) \underset{\text{chain rule.}}{=} \mu\, H(X).$$

- **Variant**: Let $X = (X_1, X_2, \dots, X_n)$ be a RV and $\mathcal{A} = \{A_i\}_{i \in I}$ be a collection of subsets of $[n]$ s.t. $\forall i \in [n]$, i appears in $\geq k$ sets; then

$$\sum_{i \in I} H[X_{A_i}] \geq k \cdot H[X].$$

Here, $X_A = (X_j : j \in A)$ for all $A \subseteq [n]$.

Proof: As in the previous proof, let $T = \{i_1, \dots, i_s\}$ with $i_1 < i_2 < \dots < i_s$. Then

$$H(X_T) \geq H(X_{i_1}) + H(X_{i_2} | X_{<i_2}) + \dots + H(X_{i_k} | X_{<i_s}).$$

$$\geq \sum_{j=1}^{s} H(X_{i_j} | X_{<i_j}).$$

Now if we sum over all $T \in \mathcal{A}$, then for each $i \in [n]$ the term $H(X_i | X_{<i})$ appears at least k times, as each i appears in at least k sets.

Hence, $$\sum_{T \in \mathcal{A}} H(X_T) \geq k \sum_{j=1}^{n} H(X_j | X_{<j})$$

$$= k H(X). \qquad \blacksquare$$

Original proof of Shearer's lemma was based on intricate induction argument. For the proof see Theorem 22.7 in Jukna Book.

- Note for the puzzle, we have $X = (x, y, z)$, $n = 3$, $\mathcal{A} = \{(1,2), (1,3), (2,3)\}$, i.e. $k = 2$.

corresponds to P_1, P_2, P_3

$$\Rightarrow H[P_1] + H[P_2] + H[P_3] \geq 2 H[P]$$

• Intersecting families of graphs.

Remember Erdos-Ko-Rado.

Suppose $\mathcal{F}$ is a family of subsets of $[n]$.

$\mathcal{F}$ is k-intersecting if $\forall A, B \in \mathcal{F}$, $|A \cap B| \geq k$.

Claim: If $\mathcal{F}$ is 1-intersecting, then $|\mathcal{F}| \leq 2^{n-1}$.

→ Follows from the fact that $\forall A \subseteq [n]$, $\mathcal{F}$ can either contain A or A^c, not both.

→ We can also get a large family of this size by taking all sets containing 1.

→ Similarly we can get a large k-intersecting family of size $2^n/2^k$ by taking all sets containing $[k]$. Can we do better? We dont need same k elements in all pairwise intersections.

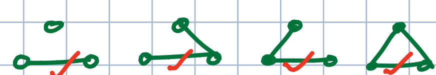
→ Let $\mathcal{F} = \{A \subseteq [n] : |A| \geq n/2 + k/2\}$. Then every two sets have $\geq k$ elements in common.

$$|\mathcal{F}| = \sum_{i=n/2+k/2}^{n} \binom{n}{i} \geq \left(\frac{2^n}{2}\right)\left(1 - O\left(\frac{k}{\sqrt{n}}\right)\right).$$

• Now we study similar properties for graphs.

• Let $\mathcal{G}$ be a family of graphs with vertex set $[n]$. $\mathcal{G}$ is intersecting, if $\forall T, K \in \mathcal{G}$, $T \cap K$ has an edge.

→ As previously we have a family of size $2^{\binom{n}{2}}/2$ s.t. all share a common edge.



• $\mathcal{G}$ is ∇-intersecting, if $\forall T, K \in \mathcal{G}$, $T \cap K$ contains a triangle.

As previous we do get a family of size $2^{\binom{n}{2}}/8$.

But can we have $2^{\binom{n}{2}}/2$ as above? → No!

Theorem: If $\mathcal{G}$ is ∇-intersecting, then
$$|\mathcal{G}| \leq 2^{\binom{n}{2}}/4.$$

Proof:

Step 1. Entropy & counting.

Let G be a uniformly random graph from $\mathcal{G}$

Hence, $H[G] = \log |\mathcal{G}|$ (I)

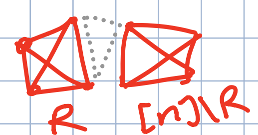
So, we can think of $G := (X_1, \dots, X_{\binom{n}{2}})$ where X_i is the RV corresponding to ith edge.

Step 2. Create a distribution S.
(To apply Shearer's lemma)

Let X_S be the RV from G restricted to edge sets in some graph G_S.
We want the support size of X_S $(:= \lambda_S)$ to be small.
As then, $H[X_S] \leq \log \lambda_S$ and to apply Shearer's lemma we need a good upper bound on X_S.

Step 3 Relate with 1-intersecting family.

For any $R \subseteq [n]$, let G_R be the graph consisting of two disconnected cliques, one on R & the other on $[n] \setminus R$.



Let E be number of edges in G_R.

Observation: As $\forall\, T, K \in \mathcal{G}$, $T \cap K$ contains a ∇, $T \cap K \cap G_R$ contains an edge as 2 vertices in the ∇ either belong to either R or $[n] \setminus R$.

Thus, the family of graphs $\{T \cap G_R : T \in \mathcal{G}\}$ is 1-intersecting, so has size $\le 2^E/2$. ... (*)

step 4. A good candidate 's'.

Let S be uniformly random graph G_R obtained by picking a random subset R of size $n/2$.

By symmetry, an edge is in G_R w.p. $E/\binom{n}{2}$.

Then as X_S is supported on an intersecting family (from *): $\mathbb{E}[H(X_S)] \le \log(2^E/2) = E-1$. (II)

Step 4. Apply Shearer's lemma.

Applying Shearer's lemma with $\mu = E/\binom{n}{2}$, we get

$$\mathbb{E}[H(X_S)] \geq \frac{E}{\binom{n}{2}} H[G]$$

we wanted μ to be large.
supp(X_S) to be small.

(II) ; (I)

$$\Rightarrow E-1 \geq \frac{E}{\binom{n}{2}} \log|\mathcal{G}|$$

Thus, $\log|\mathcal{G}| \le \binom{n}{2} - \binom{n}{2}/E$

$$= \binom{n}{2} - \binom{n}{2} \Big/ 2\binom{n/2}{2}$$

[∵ #edges in graph $= \binom{n/2}{2} + \binom{n/2}{2}$]

$$= \binom{n}{2} - \frac{n(n-1)}{2\left(\frac{n}{2}\right)\left(\frac{n}{2}-1\right)}$$

$$= \binom{n}{2} - \frac{n-1}{n/2-1} \le \binom{n}{2} - 2.$$

$$\Rightarrow |\mathcal{G}| \le 2^{\binom{n}{2}}/4.$$

■

• Application : lower bounds for bandits.

§ Properties of KL-divergence :

Remember, for two probability distributions p, q on a sample space S, relative entropy or KL-divergence : $D(p\|q) = \sum_{x\in S} p(x) \ln \frac{p(x)}{q(x)} = \mathbb{E}_p\left[\ln \frac{p(x)}{q(x)}\right]$.

① Gibbs' inequality : $D(p\|q) \geq 0$.
Equality iff $p = q$.

[proved using Jensen's / log-sum inequality].

② Chain rule for product distributions :

Let the sample space be $S := S_1 \times S_2 \times \dots \times S_n$.
Let p, q be two distributions on S such that $p = p_1 \times \dots \times p_n$ and $q = q_1 \times \dots \times q_n$, where p_j, q_j are distributions on S_j, for each $j \in [n]$.

Then $D(p\|q) = \sum_{j=1}^{n} D(p_j \| q_j)$

Proof: Let $x = (x_1, \dots, x_n) \in S$ s.t. $x_j \in S_j \ \forall j \in [n]$.
$h_i(x_i) = \ln(p_i(x_i)/q_i(x_i))$.

Then $D(p\|q) = \sum_{x\in S} p(x) \ln(p(x)/q(x))$

$= \sum_{i=1}^{n} \sum_{x\in S} p(x) h_i(x_i)$ $\quad$ [Since, $\ln(p(x)/q(x)) = \sum_{i=1}^{n} h_i(x_i)$]

$= \sum_{i=1}^{n} \sum_{x_i^* \in S_i} \sum_{\substack{x\in S \\ x_i = x_i^*}} h_i(x_i^*) p(x)$

$= \sum_{i=1}^{n} \sum_{x_i^* \in S_i} h_i(x_i^*) \sum_{\substack{x\in S \\ x_i = x_i^*}} p(x)$

$= \sum_{i=1}^{n} \sum_{x_i \in S_i} p_i(x_i) h_i(x_i)$ $\quad$ [since, $\sum_{x\in S,\, x_i = x_i^*} p(x) = p_i(x_i^*)$]

$= \sum_{i=1}^{n} D(p_i \| q_i)$

③ **Pinsker's inequality**: (relates individual events & KL-divergence)

For any event $A \subset S$, we have

$$2\left(p(A) - q(A)\right)^2 \leq D(p \| q).$$

• **Proof**:

From log-sum inequality, for each event $B \subset S$,

$$\sum_{x \in B} p(x) \ln\left[\frac{p(x)}{q(x)}\right] \geq \left(\sum_{x \in B} p(x)\right) \ln\left[\frac{\sum_{x \in B} p(x)}{\sum_{x \in B} q(x)}\right]$$

$$= p(B) \ln\left(p(B)/q(B)\right).$$

Hence,

$$\begin{aligned} \sum_{x \in A} p(x) \ln\left(\frac{p(x)}{q(x)}\right) &\geq p(A) \ln\left(\frac{p(A)}{q(A)}\right). \\ \sum_{x \notin A} p(x) \ln\left(\frac{p(x)}{q(x)}\right) &\geq p(\bar{A}) \ln\left(\frac{p(\bar{A})}{q(\bar{A})}\right). \end{aligned} \tag{$*$}$$

Let $a = p(A)$, $b = q(A)$, w.l.o.g. assume $a < b$.

Then, $D(p \| q)$

$$= \sum_{x \in A} p(x) \ln \frac{p(x)}{q(x)} + \sum_{x \in \bar{A}} p(x) \ln \frac{p(x)}{q(x)}$$

$$\geq p(A) \ln\left(\frac{p(A)}{q(A)}\right) + p(\bar{A}) \ln\left(\frac{p(\bar{A})}{q(\bar{A})}\right)$$

$$= a \ln \frac{a}{b} + (1-a) \ln\left(\frac{1-a}{1-b}\right)$$

$$= \int_a^b -\frac{a}{x}\, dx + \int_a^b \frac{(1-a)}{(1-x)}\, dx$$

$$= \int_a^b \frac{(1-a)x - (1-x)a}{x(1-x)}\, dx = \int_a^b \frac{(x-a)}{x(1-x)}\, dx$$

$$\geq \int_a^b 4(x-a)\, dx \quad \left[\text{since, } x(1-x) \leq 1/4 \text{ for } x \in (0,1)\right]$$

$$= 2(b-a)^2 = 2\left[q(A) - p(A)\right]^2. \quad \blacksquare$$

ℓ_1-metric seems a natural extension. metric does not add up in product space, KL divergence helps here.

• Pinsker's inequality also relates relative entropy and total variation distance (TV).

Pinsker's inequality imply:

$$\delta_{TV}(p,q) \le \left(\tfrac{1}{2} D(p\|q)\right)^{1/2}.$$

§ Total variation distance between two probability distribution functions p & q is:

$$\delta_{TV}(p,q) = \sup_{A \subseteq S} |p(A) - q(A)|.$$

[Claim: $\delta_{TV}(p,q) = \frac{1}{2} \sum_{\omega \in S} |p(\omega) - q(\omega)| =: \frac{1}{2} \|p-q\|_1$.

Proof: Let $B = \{\omega \in S : p(\omega) \geq q(\omega)\}$.

Then $\|p-q\|_1 = \sum_{\omega \in S} |p(\omega) - q(\omega)|$

$$= \sum_{\omega \in B} \overbrace{(p(\omega) - q(\omega))}^{\geq 0} + \sum_{\omega \in B^c} \overbrace{(q(\omega) - p(\omega))}^{> 0}$$

$$= p(B) - q(B) + q(B^c) - p(B^c)$$

$$= p(B) - q(B) + (1 - q(B)) - (1 - p(B))$$

$$= 2(p(B) - q(B)). \quad \cdots (*)$$

Now, $\delta_{TV}(p,q) = \sup_{A \subseteq S} |p(A) - q(A)|$.

$$= p(B) - q(B)$$

$$= \frac{1}{2} \sum_{\omega \in S} |p(\omega) - q(\omega)|.$$

$$= \frac{1}{2} \|p-q\|_1 \quad \text{[From } (*)\text{]} \quad \blacksquare$$

There are many other notions of distances between probability distributions such as Hellinger distance, Wasserstein distance Kolmogorov-Smirnov distance etc.

④ Relative entropy of Bernoulli RVs.

Let $B(p)$ be Bernoulli RV with mean p.

Then for all $\varepsilon \in (0, \frac{1}{2})$,

$$D\left(B\left(\frac{1+\varepsilon}{2}\right) \| B\left(\frac{1}{2}\right)\right) \leq 2\varepsilon^2, \text{ and}$$

$$D\left(B\left(\frac{1}{2}\right) \| B\left(\frac{1+\varepsilon}{2}\right)\right) \leq \varepsilon^2.$$

$$D\left(B\left(\frac{1}{2}\right) \| B\left(\frac{1-\varepsilon}{2}\right)\right) \leq \varepsilon^2.$$

• Proof:

We earlier showed

$$D(B(p) \| B(q)) = p \ln \frac{p}{q} + (1-p) \ln \left(\frac{1-p}{1-q}\right)$$

Hence, $D\left(B\left(\frac{1+\varepsilon}{2}\right) \| B\left(\frac{1}{2}\right)\right)$

$$= \left(\frac{1+\varepsilon}{2}\right) \ln(1+\varepsilon) + \left(\frac{1-\varepsilon}{2}\right) \ln(1-\varepsilon)$$

$$= \underbrace{\frac{1}{2} \ln(1-\varepsilon^2)}_{\text{negative}} + \frac{\varepsilon}{2} \ln\left(\frac{1+\varepsilon}{1-\varepsilon}\right)$$

$$< 0 + \frac{\varepsilon}{2} \cdot \frac{2\varepsilon}{1-\varepsilon} \quad \left[\because \ln\left(\frac{1+\varepsilon}{1-\varepsilon}\right) = \ln\left(1 + \frac{2\varepsilon}{1-\varepsilon}\right) < \frac{2\varepsilon}{1-\varepsilon}\right]$$

$$= \frac{\varepsilon^2}{1-\varepsilon} \leq 2\varepsilon^2.$$

Similarly, $D\left(B\left(\frac{1}{2}\right) \| B\left(\frac{1+\varepsilon}{2}\right)\right)$

$$= \frac{1}{2} \ln\left(\frac{1}{1+\varepsilon}\right) + \frac{1}{2} \ln\left(\frac{1}{1-\varepsilon}\right)$$

$$= -\frac{1}{2} \ln(1-\varepsilon^2) \leq -\frac{1}{2}(-2\varepsilon^2) \quad \left[\because \ln(1-x) \geq -2x \text{ for } x \leq \frac{1}{2}\right]$$

$$\leq \varepsilon^2.$$

Also, $D\left(B\left(\frac{1}{2}\right) \| B\left(\frac{1-\varepsilon}{2}\right)\right)$

$$= \frac{1}{2} \ln\left(\frac{1}{1-\varepsilon}\right) + \frac{1}{2} \ln\left(\frac{1}{1+\varepsilon}\right) \leq \varepsilon^2.$$

§ Example: Flipping a coin.

Given: A biased random coin:
a distribution on $\{0,1\}$ with unknown mean $\mu \in (0,1)$.
We know μ is either μ_1 or μ_2, where $\mu_1 > \mu_2$.

Goal: Flip the coin T times. Identify whether $\mu = \mu_1$ or μ_2 with high probability.

Formally, if $S := \{0,1\}^T$ be the sample space for outcomes of T coin flips, then we need a decision rule $\text{Rule}: S \to \{1,2\}$, s.t.

$$\Pr[\text{Rule}(\text{Observations}) = 1 \mid \mu = \mu_1] \geq (1-\delta)$$
$$\Pr[\text{Rule}(\text{Observations}) = 2 \mid \mu = \mu_2] \geq (1-\delta),$$

where $0 < \delta < 1/4$.

Q. How large should T be for such a decision rule to exist?

Claim: $T \sim O(\mu_1 - \mu_2)^{-2}$ is sufficient.

Proof: Say $T = K(\mu_1 - \mu_2)^{-2}$, and $\hat{\mu}$ be the empirical mean. Say $\theta := \mu_1 - \mu_2$.

Chernoff: $X_1, \ldots, X_\beta$ be indep. RV with support in $[0,1]$, then $\forall t > 0$, $\Pr(|\Sigma X_i - \mathbb{E}(\Sigma X_i)| > \alpha) < 2 \cdot \exp(-2\alpha^2/\beta)$

Thus if the coin has mean μ_1, $\Pr(\hat{\mu} \leq \mu_1 - \theta/2)$
$\leq \Pr(|\hat{\mu}T - \mu_1 T| > \theta T/2) \leq 2 \cdot \exp\left(-2 \cdot \frac{\theta^2 T^2}{4} \cdot \frac{1}{T}\right)$

$$= 2\exp\left(-\frac{1}{2} \cdot \theta^2 \cdot \frac{K}{\theta^2}\right) = 2\exp(-K/2).$$

Similarly, if coin has mean μ_2 then

$$\Pr(\hat{\mu} \geq \mu_2 + \theta/2) \leq 2\exp[-(\theta/2)^2 \cdot (K/\theta^2)] = 2\exp[-K/4]$$

Thus if $\hat{\mu} \geq \frac{\mu_1+\mu_2}{2}$, we return mean to be μ_1 and else return μ_2.

Claim: $T \sim \Omega(\mu_1-\mu_2)^{-2}$ is necessary.

For simplicity, we assume $\mu_1 = \frac{1+\varepsilon}{2}$, $\mu_2 = \frac{1}{2}$ and show $T > 1/4\varepsilon^2$.

<u>Proof</u>: For a valid decision rule, let $A_0 \subseteq S$ be the event that the rule returns "1".

Then,

$$\Pr[A_0 \mid \mu = \mu_1] - \Pr[A_0 \mid \mu = \mu_2] \geq 1 - 2\delta \quad \cdots \; (*)$$

Let $P_i(A) = \Pr[A \mid \mu = \mu_i]$, for event $A \subseteq S$, $i \in \{1,2\}$.

Let $P_{i,t}$ be the distribution of the t'th toss if $\mu = \mu_i$. Then $P_i = P_{i,1} \times \cdots \times P_{i,T}$.

So,
$$
\begin{aligned}
2\left[P_1(A) - P_2(A)\right]^2 &\leq D(P_1 \,\|\, P_2) && [\text{Pinsker}] \; (3) \\
&\leq \sum_{t=1}^{T} D[P_{1,t} \,\|\, P_{2,t}] && \left[\begin{matrix}(2)\\ \text{chain} \\ \text{rule}\end{matrix}\right] \\
&\leq T \cdot 2\varepsilon^2 && [(4)\ \text{Bernoulli RV}]
\end{aligned}
$$

$\Rightarrow |P_1(A) - P_2(A)| \leq \varepsilon\sqrt{T}$.

So for $A = A_0$ and $T \leq 1/4\varepsilon^2$, we obtain

$$|P_1(A_0) - P_2(A_0)| \leq 1/2 < 1 - 2\delta.$$

This contradicts $(*)$.

Note: Lower bound proof applies to all decision rules at once.

- Generalization to more than two coins.

We have n coins, at most one is biased (mean $\frac{1+\varepsilon}{2}$). The algorithm can choose a single coin x_t out of n coins, to flip at time $t \in [T]$.

At the end of time T, algorithm needs to guess the biased coin, if any. Let the guess be y_T.

To show the lower bound, we construct the following $(n+1)$ distributions on coin-flip outcomes.

p_0 : all coins are fair.

p_j : j'th coin has mean $\frac{1+\varepsilon}{2}$, other coins are fair.
$(j \in [n])$

Note that in all these distributions, the different coin flips are mutually independent events.

For $j \in [0, n]$, we denote the probability and expectation of an event under distribution p_j by $\mathbb{P}r_j$ and $\mathbb{E}_j$, respectively.

- Theorem: Let ALG be any coin-flipping algorithm. If $T \leq \frac{n}{100\,\varepsilon^2}$ then there exists at least $n/3$ distinct values of $j > 0$ s.t. $\mathbb{P}r_j(y_T \neq j) \geq 1/2$.

Proof: Let Q_j denote RV that counts the number of times ALG flips coin j.

Then $\sum_{j=1}^{n} \mathbb{E}_0(Q_j) = \mathbb{E}_0\left(\sum_{j=1}^{n} Q_j\right) = T$.

So at most $n/3$ coins can have $Q_j > 3T/n$ [Averaging argument].

and at most $n/3$ coins can have $\mathbb{P}r_0(y_T = j) > 3/n$.

[This also follows from averaging argument. Say we have x coins with $\mathbb{P}r_0(y_T = j) > 3/n$.

Then, $1 = \sum_j \mathbb{P}r_0(y_T = j) > x \cdot 3/n \Rightarrow x < \frac{n}{3}$]

Consider the sets:

$J_1 = \{j : \mathbb{E}_0(Q_j) \leq 3T/n\}$, $J_2 = \{j : \mathbb{P}r_0(y_T = j) \leq 3/n\}$.

Then $|J_1| \geqslant 2n/3$, $|J_2| \geqslant 2n/3$.

Let $J = |J_1 \cap J_2|$. Then $|J| \geqslant n/3$.

Let $j \in J$ and define the event $\mathcal{E} := \{y_T = j\}$.

Then,

$$\mathbb{P}r_j(\mathcal{E}) \leq \mathbb{P}r_0(\mathcal{E}) + |\mathbb{P}r_j(\mathcal{E}) - \mathbb{P}r_0(\mathcal{E})|$$

$$\leq \mathbb{P}r_0(\mathcal{E}) + \frac{1}{2}\|P_0 - P_j\|_1 \quad \left[\text{Definition of } \delta_{TV}(p,q)\right]$$

$$\leq \frac{3}{n} + \sqrt{\frac{1}{2} D(P_0 \| P_j)} \quad \left[\text{From Pinsker's inequality}\right]$$

Now, using chain rule: $D(P_0 \| P_j)$

$$= \sum_{t=1}^{T} D(P_0(x_t) \| P_j(x_t))$$

Unlike two coins, now we have many coins to choose from & the choice may depend on the outcomes of previous tosses.

Then using conditional rel. entropy, $\sum_{t=1}^{T} D(P_0(x_t) \| P_j(x_t))$

$$= \sum_{t=1}^{T} \sum_{x_1, \dots, x_{t-1}} \mathbb{P}r_0[x_1, \dots, x_{t-1}] \, D(P_0(x_t) \| P_j(x_t) \mid x_1, \dots, x_{t-1})$$

Let $x_1, \dots, x_T$ be the outputs of the coin tosses seen by ALG. Let $P_0(x_t)$ and $P_j(x_t)$ denote the distribution of t'th coin toss seen by ALG, given the outputs of the first $t-1$ tosses.

Now $P_j(x_t)$ is a single coin toss, which is a fair coin for $x_t \neq j$ and biased coin if $x_t = j$. $P_0(x_t)$ is always corresponds to fair coin toss.

Then $D(P_0 \| P_j)$

$$= \sum_{t=1}^{T} \sum_{x_1,\dots,x_{t-1}} \Pr_0[x_1,\dots,x_t] \cdot \mathbb{1}_{\{x_t = j\}} \cdot D\left(B\left(\tfrac{1}{2}\right) \| B\left(\tfrac{1+\varepsilon}{2}\right)\right)$$

$$\leq \mathbb{E}_0[Q_j] \cdot \varepsilon^2 \leq \frac{3T}{n} \cdot \varepsilon^2.$$

Hence, $\Pr_j(\mathcal{E}) \leq \frac{3}{n} + \sqrt{\frac{1}{2} \cdot \frac{3T}{n} \cdot \varepsilon^2}$

As $T \leq \frac{n}{100\varepsilon^2}$, for large enough n: $\Pr_j(\mathcal{E}) \leq \frac{1}{2}$.

This proves the theorem. ■

§ Multi-armed Bandits: (MAB)

An important problem in online decision-making.

Given: K arms, T rounds.

In each round $t \in [T]$

1. ALGO picks arm a_t.
2. ALGO observes reward $r_t \in [0,1]$ for the chosen arm.

- We consider stochastic MAB, where reward for each arm a_t is IID, say Bernoulli RV with mean μ_t.

Let $\mu^* := \max_{a_t} \mu_t$,

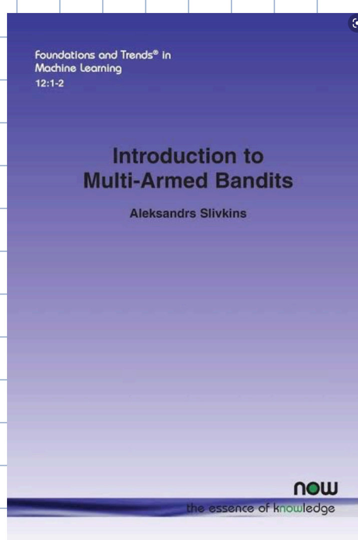
i.e.

best mean reward

Goal:

minimize regret

$$R(T) := \mu^* \cdot T - \sum_{t=1}^{T} \mu_t.$$



• **Theorem**: For stochastic multi-armed bandit problem. for fixed time horizon T and the number of arms K, for any algorithm, there exists a problem instance s.t. $\mathbb{E}[R(T)] \geq \Omega(\sqrt{KT})$. (large enough K)

Proof: We define the distribution by chosing a random $i^* \in [K]$ and defining the $r_t(i)$ as

$$r_t(i) = \begin{cases} 1 & \text{w.p. } 1/2 \\ 0 & \text{w.p. } 1/2 \end{cases} \text{ if } i \neq i^*, \quad r_t(i) = \begin{cases} 1 & \text{w.p. } \frac{1+\varepsilon}{2} \\ 0 & \text{w.p. } \frac{1-\varepsilon}{2} \end{cases} \text{ if } i = i^*$$

We choose $1/\varepsilon = \sqrt{100T/K}$.

We can think of an algorithm that chooses action $x_t \in [K]$ at time t as a coin-guessing algorithm which chooses coin x_t at time t.

For $t \leq \frac{n}{100\varepsilon^2}$, $\exists\, J_t \subseteq [K]$ with $|J_t| \geq K/3$ s.t. $\forall j \in J_t$, $\Pr_j(x_t = j) \leq 1/2$.

Hence, $\mathbb{E}[r_t(x_t)]$

$$\leq \overset{J_t}{\frac{1}{3}} \cdot \left(\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \left(\frac{1+\varepsilon}{2}\right)\right) + \overset{\bar{J}_t}{\frac{2}{3}} \left(\frac{1+\varepsilon}{2}\right) \leq \frac{1}{2} + \frac{5\varepsilon}{12}.$$

Here the expectation is also over the choice of i^*.

On the other hand,

$$\mathbb{E}\left[\min_{i \in [K]} \sum_{t=1}^{T} r_t(i)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} r_t(i^*)\right] \leq \left(\frac{1+\varepsilon}{2}\right) T.$$

Thus we have $\mathbb{E}[R_T]$

$$\geq \left(\frac{1+\varepsilon}{2}\right) T - \left(\frac{1}{2} + \frac{5\varepsilon}{12}\right) T \geq \frac{\varepsilon T}{12} \geq \frac{1}{6}\sqrt{\frac{KT}{100}}. \qquad \blacksquare$$

→ So there is one instance with regret $\geq \frac{1}{6}\sqrt{\frac{KT}{100}}$.