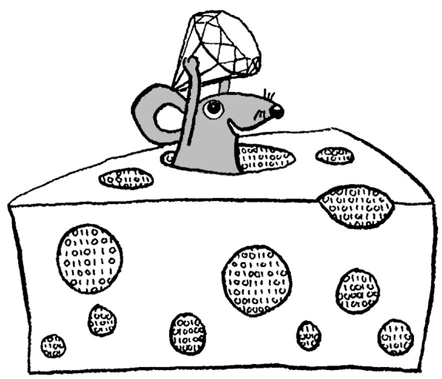
Hidden Gems

of

Sublinear Algorithms



A Latex Product of the Simons Institute

~ with contributions from numerous ~

Eminent, Illustrious, and Famed Scholars:

◇ **Lecturers** ◇	◇ **Scribes** ◇
Clément Canonne	Vipul Arora
Sepehr Assadi	Jane Lange
Nathan Harms	Nathan Harms
Guy Blanc	Guy Blanc
Santhoshini Velusamy	Santhoshini Velusamy
Abigail Gentle	Sandeep Silwal
Akash Kumar	Akash Kumar
Maryam Aliakbarpour	Ria Stevens
Ari Biswas	Ari Biswas
Krzysztof Onak	Kaiwen He
Vipul Arora	Vipul Arora
Samson Zhou	Shivam Nadimpalli
Shweta Jain	Sabyasachi Basu

Collected and Edited by

Nathan Harms

What is this?

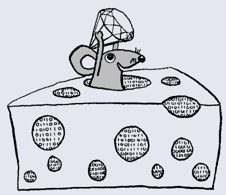
This is a collection of lecture notes from the "Hidden Gems" lecture series, part of the Sublinear Algorithms program at the Simons Institute, May–August 2024. In these lectures, speakers explained a theorem, lemma, or technique that they regard as a hidden gem – useful in many situations, yet outside the standard toolbox of sublinear algorithms. We hope the reader finds some of these gems valuable in their adventures through time and space, even if they do not read the entire input.

Acknowledgments

Thanks to Clément Canonne and Ronitt Rubinfeld, who proposed and encouraged this series. Thanks to Guy Blanc and Ari Biswas for help creating the LaTeX templates. Thanks to the Simons institute for hosting the Sublinear Algorithms program, and to the organizers of this program: Clément Canonne, Artur Czumaj, Piotr Indyk, Jelani Nelson, Noga Ron-Zewi, Ronitt Rubinfeld, and Asaf Shapira.

– *Nathan Harms*

Contents



Sublinear Algorithms Hidden Gem 1

Levin's Economical Work Investment Strategy

Lecturer: Clément Canonne. *Scribe:* Vipul Arora

May 28, 2024

1 Introduction

In this talk, we discussed the problem of finding a "high quality" element x from some universe U, when the cost of evaluating an element grows larger when the quality is smaller; of course we would like to minimize the total cost of finding the high quality element. For example, consider the task of finding a biased coordinate in a random binary vector. If the bias p of a coordinate is very close to 0.5, the task of distinguishing it from an unbiased coordinate becomes more difficult as the number of samples required becomes $\Theta(1/(2p-1)^2)$. So, the more biased a coordinate is, the less the work needed to find it.

The property of interest is quantified by a *quality* function $a : U \to [0,1]$, and the work required for verifying the element is quantified by its quality via the *cost* function $c : [0,1] \to (0,\infty)$. So, the cost of verifying an element $x \in U$ become $c(a(x))$, with $c(a(\cdot)) : U \to (0,\infty)$.

The quality function is unknown beforehand, so the task of verifying a claim of the kind: "$x \in U$ is such that $a(x) \geq \varepsilon$" can only be done by *investing* $c(a(x))$ units of work. The cost is assumed to be inversely proportional to the quality, $c \propto {}^1/_a$, typical examples being $c(t) = \Theta\left({}^1/_{t^{1+\delta}}\right)$, for some $\delta \geq 0$.

To avoid investing too much work into a *possibly low* quality element, a strategy of randomly sampling many elements and investing a limited amount of work in each of them is deployed. We assume that we know a lower bound on the expected quality of the sampled elements, i.e. for some $\varepsilon > 0$,

$$\mathbb{E}_{X\sim U}[a(X)] \geq \varepsilon. \tag{1}$$

Towards the goal of finding and verifying a *good* $x \in U$, we note this useful observation:

Observation 1. *If we have a guarantee of the form* $\Pr_X[a(X) \geq \alpha] \geq \beta$*, then* $O(c(\alpha)/\beta)$ *work suffices to find and verify some* $x \in U$ *such that* $a(x) \geq \alpha$*, with a constant probability of success.*

Proof. Let us randomly sample $t = O({}^1/_\beta)$ independent elements $X = \{x_1, \ldots, x_t\}$ from U. We have

$$\Pr_X[a(x_i) < \alpha, \forall i \in [t]] = \prod_{i=1}^{n} (1 - \Pr_X[a(x_i) \geq \alpha]) \leq (1-\beta)^{O({}^1/_\beta)} \leq {}^1/_e.$$

So, with constant success probability, we will find a *good* $x_i \in U : a(x_i) \geq \alpha$. To verify this, we do at most $c(a(x_i)) \leq c(\alpha)$ units of work on each sample, thus proving $O(c(\alpha)/\beta)$ units of work suffices. ∎

2 An Initial Strategy

Breaking the expectation in (1) over low and high quality elements, we get:

Observation 2 (Markov's Inequality). *If* $\mathbb{E}_X[a(X)] \geq \varepsilon$, *then* $\Pr_X[a(X) \geq \varepsilon/2] \geq \varepsilon/2$.

Combining Observation 1 with Observation 2, we get $O\left(\frac{2}{\varepsilon}c\left(\frac{2}{\varepsilon}\right)\right)$ work suffices, by randomly sampling $t = O(1/\varepsilon)$ elements and investing $c(2/\varepsilon)$ units of work on each. For $c(t) = O(1/t^{1+\delta})$, an upper bound of $O(1/\varepsilon^{2+\delta})$ is achieved, which isn't quite optimal if we consider the two extremes:

- A $\Theta(1)$ fraction of $x \in U$ are with quality $\Theta(\varepsilon)$. In this case, just by Observation 1, $O(c(\varepsilon)) = O(1/\varepsilon^{1+\delta})$ work suffices.
- A $\Theta(\varepsilon)$ fraction of $x \in U$ are with quality $\Theta(1)$. Again, by Observation 1, $O(1/\varepsilon)$ work suffices.

So, we would like to come up with a better strategy, via a counting argument:

Lemma 3. *Let* $L := \lceil \log_2 2/\varepsilon \rceil$. *If* $\mathbb{E}_X[a(X)] \geq \varepsilon$, *then there exists* $\ell \in [L] := \{1, 2, \ldots, L\}$ *such that*

$$\Pr_X[a(X) \geq 2^{-\ell}] \geq \frac{2^\ell \varepsilon}{4L} := \delta_L.$$

Proof. Define the following buckets: $B_0 := \{x \in U : a(x) \leq \varepsilon/2\}$, and $B_\ell := \{x \in U : a(x) \in (1/2^\ell, 1/2^{\ell-1}]\}$, for every $\ell \in [L]$, partitioning U. Breaking the expectation in (1) over these buckets, we get

$$\varepsilon \leq \mathbb{E}_X[a(X)] \leq \frac{\varepsilon}{2}\underbrace{\Pr_X[X \in B_0]}_{\leq 1} + \sum_{\ell=1}^{L} \frac{1}{2^{\ell-1}} \Pr_X[X \in B_\ell]. \tag{2}$$

So, by averaging, there must exists some $\ell^\star \in [L]$ such that $\Pr_X[X \in B_{\ell^\star}] \geq \frac{\varepsilon}{2L} 2^{\ell^\star - 1} = \delta_L$. And hence, $\Pr_X[a(X) \geq 2^{-\ell^\star}] = \sum_{\ell=\ell^\star}^{L} \Pr_X[X \in B_\ell] \geq \Pr_X[X \in B_{\ell^\star}] \geq \delta_L$. ∎

Using Lemma 3, we describe **Levin's Strategy**[Gol14]: For every $\ell \in [L]$, define $\varepsilon_\ell := 1/2^\ell$, and

- Sample $\eta_\ell := O(1/\delta_L) = O(L/2^\ell \varepsilon)$ random elements from U.
- Invest $c(\varepsilon_\ell) = O(1/\varepsilon_\ell^{1+\delta}) = O(2^{\ell(1+\delta)})$ work on each of the η_ℓ random elements sampled above.

Combining Observation 1, and Lemma 3, we will find some $x \in U$ such that $a(x) \geq \varepsilon_\ell$, with a constant probability of success. The total work entailed will be ($\sim$ indicates ignoring constant factors):

$$\begin{aligned}
\sum_{\ell=1}^{L} \eta_\ell c(\varepsilon_\ell) &\lesssim \frac{L}{\varepsilon} \sum_{\ell=1}^{L} \frac{1}{2^\ell} \cdot 2^{\ell(1+\delta)} = \frac{L}{\varepsilon} \sum_{\ell=1}^{L} 2^{\ell\delta} = \frac{L}{\varepsilon} \cdot \frac{2^{L\delta} - 1}{2^\delta - 1} \cdot 2^\delta && \text{(assuming } \delta \neq 0\text{)} \\
&\lesssim \frac{L}{\varepsilon} 2^{L\delta} \lesssim \frac{\log_2 2/\varepsilon}{\varepsilon} \left(\frac{2}{\varepsilon}\right)^\delta \lesssim \frac{\log 1/\varepsilon}{\varepsilon^{1+\delta}}, && \text{(Using } L = \lceil \log_2 2/\varepsilon \rceil\text{)}
\end{aligned}$$

improving the $O(1/\varepsilon^{2+\delta})$ upper bound shown earlier, via Observation 2. However, if $\delta = 0$, i.e., $c(t) = O(1/t)$, then the total work entailed via this strategy would be worse by a $\log 1/\varepsilon$ factor:

$$\sum_{\ell=1}^{L} \eta_\ell c(\varepsilon_\ell) \lesssim \frac{L}{\varepsilon} \sum_{\ell=1}^{L} \frac{1}{2^\ell} \cdot 2^\ell = \frac{L^2}{\varepsilon} \lesssim \frac{\log 1/\varepsilon}{\varepsilon^2}.$$

3 A Better Strategy

To get an even better strategy, we refine the counting argument presented in Lemma 3:

Lemma 4. *Let* $L := \lceil \log_2 2/\varepsilon \rceil$. *If* $\mathbb{E}_X[a(X)] \geq \varepsilon$, *then there exists* $\ell \in [L]$ *such that*

$$\Pr_X[a(X) \geq 2^{-\ell}] \geq \frac{2^\ell \varepsilon}{8(L+1-\ell)^2} := \delta_L.$$

Proof. Using the same bucketing as in the proof of Lemma 3, we assume for contradiction that: $\forall \ell \in L$,

$$\Pr_X[a(X) \geq 2^{-\ell}] < \frac{2^\ell \varepsilon}{8(L+1-\ell)^2}.$$

Again breaking the expectation in (1) over these buckets, we get

$$\begin{aligned}
\varepsilon \leq \mathbb{E}_X[a(X)] &\leq \frac{\varepsilon}{2} \underbrace{\Pr_X[X \in B_0]}_{\leq 1} + \sum_{\ell=1}^{L} \frac{1}{2^{\ell-1}} \underbrace{\Pr_X[X \in B_\ell]}_{\leq \Pr_X[a(X) \geq 2^{-\ell}]} \leq \frac{\varepsilon}{2} + \sum_{\ell=1}^{L} \frac{1}{2^{\ell-1}} \Pr_X[a(X) \geq 2^{-\ell}] \\
&< \frac{\varepsilon}{2} + \frac{\varepsilon}{8} \sum_{\ell=1}^{L} \frac{1}{2^{\ell-1}} \cdot \frac{2^\ell}{(L+1-\ell)^2} = \frac{\varepsilon}{2} + \frac{\varepsilon}{4} \sum_{\ell=1}^{L} \frac{1}{\ell^2} && (\text{Reindexing } L+1-\ell \to \ell) \\
&< \frac{\varepsilon}{2} + \frac{\varepsilon}{4} \sum_{\ell=1}^{\infty} \frac{1}{\ell^2} \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{4} \cdot \frac{\pi^2}{6} && \left(\because \textstyle\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}\right) \\
&= \frac{\varepsilon}{2} + \varepsilon \frac{\pi^2}{24} < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon, && (\because \pi^2 < 12)
\end{aligned}$$

which is a contradiction. So, the conclusion must hold. ∎

Using Lemma 4, we improve Levin's strategy by updating δ_L, and the total work thus becomes:

$$\begin{aligned}
\sum_{\ell=1}^{L} \eta_\ell c(\varepsilon_\ell) &\lesssim \sum_{\ell=1}^{L} \frac{(L+1-\ell)^2}{2^\ell \varepsilon} \cdot 2^{\ell(1+\delta)} \\
&= \frac{1}{\varepsilon} \sum_{\ell=1}^{L} (L+1-\ell)^2 2^{\ell\delta} = \frac{2^{(L+1)\delta}}{\varepsilon} \sum_{\ell=1}^{L} 2^{-\ell\delta} \ell^2 && (\text{Reindexing } L+1-\ell \to \ell) \\
&\lesssim \frac{1}{\varepsilon^{1+\delta}} \sum_{\ell=1}^{L} \frac{\ell^2}{2^{\ell\delta}} \coloneqq (\star) && (\text{Using } L = \lceil \log_2 2/\varepsilon \rceil) \\
&< \frac{1}{\varepsilon^{1+\delta}} \underbrace{\sum_{\ell=1}^{\infty} \frac{\ell^2}{2^{\ell\delta}}}_{=O(1)} \lesssim \frac{1}{\varepsilon^{1+\delta}}, && (\text{Assuming } \delta \neq 0)
\end{aligned}$$

shaving off a $\log 1/\varepsilon$ factor from the bound achieved in Section 2, via Lemma 3. If $\delta = 0$, the total work entailed becomes: $(\star) = \frac{1}{\varepsilon} \sum_{\ell=1}^{L} \ell^2 \lesssim \frac{L^3}{\varepsilon} \lesssim \frac{1}{\varepsilon} \log^3 1/\varepsilon$, trading off a $1/\varepsilon$ factor for a $\log^2 1/\varepsilon$ factor.

4 An Optimal Strategy for the Linear Case

For $\delta = 0$, we design an improved strategy by tweaking the argument in Lemma 3. From (2), we have:

$$\sum_{\ell=1}^{L} \frac{1}{2^\ell} \Pr_X[X \in B_\ell] \geq \frac{\varepsilon}{4}. \tag{3}$$

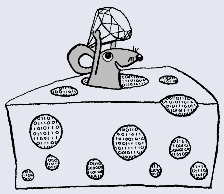
For each $\ell \in [L]$, we sample $\eta_\ell \coloneqq {}^{8}/_{2^\ell \varepsilon}$ random elements from U, and invest $c(2^{-\ell}) = O(2^\ell)$ units of work in each of them. The probability of not hitting even a single element in any of the ℓ buckets is

$$\begin{aligned}
\prod_{\ell=1}^{L} (\Pr[X \notin B_\ell])^{\eta_\ell} &= \prod_{\ell=1}^{L} (1 - \Pr[X \in B_\ell])^{\eta_\ell} \le \prod_{\ell=1}^{L} (\exp(-\Pr[X \in B_\ell]))^{\eta_\ell} && \text{(Using } 1 - x \le e^{-x}\text{)} \\
&= \exp\left(-\sum_{\ell=1}^{L} \eta_\ell \Pr[X \in B_\ell]\right) = \exp\left(-\frac{8}{\varepsilon}\sum_{\ell=1}^{L} \frac{1}{2^\ell} \Pr[X \in B_\ell]\right) && \text{(Using } \eta_\ell = {}^{8}/_{2^\ell\varepsilon}\text{)} \\
&\le e^{-2}. && \text{(Using (3))}
\end{aligned}$$

So, we hit at least one bucket with probability at least $1 - {}^{1}/_{e^2}$. while the total cost entailed becomes:

$$\sum_{\ell=1}^{L} \eta_\ell c(2^{-\ell}) \lesssim \frac{1}{\varepsilon} \sum_{\ell=1}^{L} \frac{1}{2^\ell} \cdot 2^\ell \lesssim \frac{L}{\varepsilon} \lesssim \frac{\log {}^{1}/_{\varepsilon}}{\varepsilon}.$$

In fact [BRY14, Lemma D.1] proves that this cost is optimal for linear cost functions, by showing the existence of a distribution $\mathcal{D}$ such that $\mathbb{E}_{X\sim\mathcal{D}}[a(X)] = \varepsilon$, yet for which every strategy costs $\Omega({}^{1}/_{\varepsilon} \log {}^{1}/_{\varepsilon})$.



Sublinear Algorithms Hidden Gem 2

Concentration Inequalities

Lecturer: Sepehr Assadi. *Scribe:* Jane Lange

May 29, 2024

1 Introduction

Sometimes, we want to prove concentration for a random variable, but the standard form of the Chernoff bound is loose or inapplicable. Here, we discuss other concentration inequalities that are better suited for certain cases than the standard Chernoff bound. All the examples are in the setting of the following experiment:

Definition 1 (Balls and bins experiment)**.** For problem size n, we have n balls and n bins. Each ball is thrown independently and is equally likely to land in each bin.

2 Chernoff Bounds and Maximum Load

First we will try to bound the maximum load with inverse-polynomial failure probability by applying the standard Chernoff bound. For each i, let X_i be the indicator that the i^{th} ball lands in the first bin. Then the load of the first bin is $X := \sum_{i=1}^{n} X_i$. The standard Chernoff bound says:

Theorem 2 (Standard Chernoff bound)**.** *Let $X_1, \ldots, X_n$ be independent random variables over $[0,1]$ and let $X = \sum X_i$.*

$$\Pr[X > (1+\delta)\mathbb{E}[X]] \leq \exp\left(-\frac{\delta^2 \mathbb{E}[X]}{2+\delta}\right).$$

In our case we have $\mathbb{E}[X] = n\mathbb{E}[X_i] = 1$, so to get an inverse-polynomial tail bound, we need to set $\exp(-\delta^2/(2+\delta)) = n^{-c}$ for some $c > 1$ so that failure probability is still inverse-poly after union bounding over the bins. The union bound doesn't really matter – δ is still $\Theta(\log(n))$. So we have shown that for some constant k, we have

$$\Pr[\text{maximum load } \geq k\log(n)] \leq 1/n.$$

2.1 Tightening the bound

However, by explicitly counting instead of using the Chernoff bound, we see that this bound is loose. X is a binomial random variable, so we have

$$\Pr[X \geq k] \leq \tbinom{n}{k} \cdot n^{-k}.$$

Using the bound $\binom{n}{k} \leq (en/k)^k$, we have $\Pr[X \geq k] \leq (e/k)^k$. So we can set $(e/k)^k \leq n^{-c}$, which gives $k = O(\frac{\ln n}{\ln\ln n})$. So the standard Chernoff bound gave a deviation which is loose by a factor of $1/\ln\ln n$. There is a different form that gives the right expression:

Theorem 3 (Chernoff bound for small expectations). *Let $X_1, \ldots, X_n$ be independent random variables over* $[0,1]$ *and let* $X = \sum X_i$. *Then*

$$\Pr[X > (1+\delta)\mathbb{E}[X]] \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mathbb{E}[X]}.$$

Since $\mathbb{E}[X]$ is just 1 in our case, this bound has basically the same form as what we get from the explicit counting: $(e/\delta)^{\delta}$. Then by using this bound we can set δ to be $\Theta(\ln n/\ln\ln n)$ as desired.

3 Talagrand's Inequality and Counting Empty Bins

Let Y_i be the indicator of the i^{th} bin being empty, and let $Y = \sum Y_i$. We have $\mathbb{E}[Y_i] = (1-1/n)^n \to 1/e$ and $\mathbb{E}[Y] = n/e$. We cannot simply use the Chernoff bound on Y because the Y_i's are not independent. Instead, we can use Talagrand's inequality.

Definition 4 (c-Lipschitz). Let Ω be some set. We say that $f : \Omega^n \to \mathbb{R}$ is c-Lipschitz if for all $i \in [n]$ and all $x_1, \ldots, x_n, x_i' \in \Omega$,

$$f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) \leq c \cdot f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n).$$

Theorem 5 (Talagrand's inequality/bounded differences inequality). *Let $X_1, \ldots, X_n$ be independent random variables over a set Ω and let $f : \Omega^n \to \mathbb{R}$ be a c-Lipschitz function. Then*

$$\Pr[|f - \mathbb{E}[f]| \geq t] \leq \exp(-\frac{t^2}{2c^2 n}).$$

In our case, the independent random variables are over the set $[n]$, where $X_i = j$ if ball i goes to bin j. Then Y, the variable we care about bounding, is a 1-Lipschitz function of the X_i's, because moving one ball from one bin to another can only change the number of empty bins by at most 1. So we have

$$\begin{aligned}\Pr[|Y - \mathbb{E}[Y]| \geq \varepsilon\mathbb{E}[Y]] &\leq \exp(-\frac{\varepsilon^2\mathbb{E}[Y]^2}{2n}) \\ &\leq \exp(-\frac{\varepsilon^2\mathbb{E}[Y]}{2e^2})\end{aligned}$$

To make this tail probability inverse-polynomial, we set ε to $\Theta(\ln n/\sqrt{n})$, thus making the deviation $\Theta(\sqrt{n}\ln n)$. Then $\varepsilon^2\mathbb{E}[Y] = \Theta(\ln^2(n))$, so $\exp(-\varepsilon^2\mathbb{E}[Y]) = n^{-\Omega(1)}$. Note that this bound is effective because $\mathbb{E}[Y]$ is large compared to the number of independent variables — it is linear in n.

3.1 Certifiable functions

What if the random variable you want to bound doesn't have a large expectation relative to the number of independent variables? Then the Talagrand bound will say nothing unless ε is made very large. However, we can still say something if the function is *r-certifiable.*

Definition 6 (Certifiable). The function $f : \Omega^n \to \mathbb{R}$ is r-certifiable if for all k, to prove that $f(X_1, \ldots, X_n) \geq k$, it suffices to reveal at most rk of the inputs to f.

Theorem 7 (Concentration of certifiable functions)**.** *Let $X_1, \ldots, X_n$ be independent random variables over a set Ω and let $f : \Omega^n \to \mathbb{R}$ be a c-Lipschitz, r-certifiable function. Then*

$$\Pr[|f - \mathbb{E}[f]| \geq t + O(1)\sqrt{c^2 r \mathbb{E}[f]}] \leq \exp(-\frac{t^2}{O(c^2 r) \cdot \mathbb{E}[f]}).$$

For an example, consider the number of bins with load at least $\ell := \frac{0.9 \ln n}{\ln \ln n}$. We can define Z_i to be the indicator of the i^{th} bin having load at least ℓ, and thus our variable of interest is $Z = \sum Z_i$. We have $\mathbb{E}[Z_i] = \Theta(n^{-0.9})$ and thus $\mathbb{E}[Z] = \Theta(n^{0.1})$. The right choice of r is exactly ℓ: if k bins are ℓ-loaded, it suffices to reveal the ℓk balls that went to those bins. So we have

$$\Pr[|Z - \mathbb{E}[Z]| > t + O(1)\sqrt{\ell n^{0.1}}] \leq \exp(\frac{t^2}{\ell n^{0.1}})$$

To make this at most inverse-poly, t will need to be $\Theta(\sqrt{\ell n^{0.1} \ln n})$; this is dominated by the $n^{0.05}$ term.

3.2 Exercise: making the function certifiable

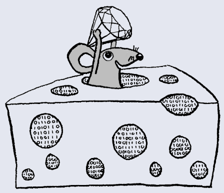
What if we want to prove concentration for the number of bins with load between $\ell = \frac{0.9 \ln n}{\ln \ln n}$ and $\ell' = \frac{0.8 \ln n}{\ln \ln n}$? Note that this is not r-certifiable for any small r. The challenge is to express this quantity as a certifiable function. One way to do this is to express it as the difference of two certifiable functions and prove concentration for each: the number of bins with load at least ℓ, which we will denote Z_ℓ, and the number with load at least ℓ', denoted $Z_{\ell'}$. We aim to show concentration for $Z_\ell - Z_{\ell'}$. We showed above that $|Z_\ell - \Theta(n^{0.1})| \leq O(n^{0.05})$ with probability $\leq n^{-2}$. By the same argument we can make $|Z_{\ell'} - \Theta(n^{0.2})| \leq O(n^{0.1})$ with probability $\leq n^{-2}$. By union bounding over these two events, we have with probability $2n^{-2}$ that $Z_{\ell'} - Z_\ell = \Theta(n^{0.2})$.

3.3 Another exercise (actually left as an exercise)

What if we want to prove concentration for the number of balls that land in bins with an odd number of balls? Note that this is not only not certifiable, it's also not even Lipschitz — there could be, for instance, a bin with $\Theta(n)$ balls in it, and removing a ball from this bin would change the parity for all $\Theta(n)$ remaining balls.

4 Further Reading

- Dubhashi and Panconesi: Concentration of Measure for the Analysis of Randomized Algorithms [DP12]
- Alon and Spencer: The Probabilistic Method [AS00]
- Molloy-Reed: Graph Coloring and the Probabilistic Method [MR02]
- David Wajc: Negative Association — Definition, Properties, and Applications [Waj17]



Sublinear Algorithms Hidden Gem 3

Lower Bounds via Ramsey's Theorems

Lecturer: Nathan Harms. *Scribe:* Nathan Harms

May 31, 2024

1 Motivation

The challenge in proving lower bounds is that algorithms can do whatever they want. Sometimes you think "It seems like any good algorithm must surely do something like *this*, and I can prove my lower bound if that is true". But then we need to prove that the algorithm "must do something like this". In this note we will see a general technique using Ramsey's theorems for this purpose – to force algorithms to behave in specific ways. We will see two examples:

1. A distribution testing problem called *testing data binning*, due to Canonne & Wimmer [CW20].
2. Testing properties of sequences, due to Fischer [Fis04], which (I believe) is the original application of this technique in sublinear algorithms.

2 Ramsey's Theorems

The next three theorems are from Ramsey's paper *On a Problem of Formal Logic* [Ram30]. From modern discussions of Ramsey theory, one may get the impression that Ramsey was quite interested in graphs and colorings thereof, but the purpose of the paper was to study the *Entscheidungsproblem* – to try to find a way to algorithmically decide the truth or falsity of a given statement (graphs are not mentioned in his paper). This was about 6 years before Church and Turing proved that undecidable problems exist. So the original application of Ramsey's theorem was to theoretical computer science.

Prior to introducing this new field of combinatorics, Ramsey first made important contributions to philosophy [Mis20], and then separately to economics, making what John Maynard Keynes called "one of the most remarkable contributions to mathematical economics ever made"[1] [Got20]. And all this before age 26. We will accomplish nearly as much in the present note.

Notation. For any set T and any number k, we will write $\binom{T}{k}$ for the set of subset $S \subset T$ with cardinality $|S| = k$, and we will write $\binom{T}{\leq k}$ for the set of subsets $S \subset T$ with cardinality $|S| \leq k$.

The first theorem of Ramsey's paper is the *infinite* Ramsey theorem[2].

Theorem 1 (Infinite Ramsey Theorem)**.** *For every $k, c \in \mathbb{N}$, and every coloring $\mathsf{col} : \binom{\mathbb{N}}{\leq k} \to \{0,1\}^c$, there exists an infinite subset $T \subset \mathbb{N}$ such that $\forall S, S' \in \binom{T}{k}$, we have*

$$\mathsf{col}(S) = \mathsf{col}(S') .$$

[1] From my reading of [Got20], Ramsey's contribution (based on the idea that one should not discount the interests of people in the future) may be the origin of current "longtermist" ideas.

[2] Ramsey's original proof used the axiom of choice, but this is not necessary, see e.g. [ER50].

The next theorem is perhaps the most familiar one. Modern restatements of this theorem use the language of graphs and edge colorings, and Ramsey's theorem is usually interpreted as a theorem about graphs (we will use the graph interpretation to sketch a proof of this theorem.)

Theorem 2 ("Standard" Ramsey Theorem)**.** *For every $n, c \in \mathbb{N}$, there exists $N \coloneqq N(n,c)$ such that for every coloring* $\mathsf{col}\colon \binom{[N]}{2} \to \{0,1\}^c$*, there is a subset $T \subset [N]$ of cardinality $|T| = n$ where $\forall S, S' \in \binom{T}{2}$,* $\mathsf{col}(S) = \mathsf{col}(S')$.

Proof sketch. For $c = 1$, the theorem is equivalent to the more well-known statement: For every $n \in \mathbb{N}$, there exists $N \coloneqq N(n)$ such that, for every red-blue edge coloring of the complete graph K_N, there is a complete n-vertex subgraph such that either every edge is red, or every edge is blue.

Consider the following recursive greedy algorithm that constructs a sequence of vertices. Given a complete graph K_t with edges colored red and blue, choose an arbitrary vertex v_1. Let R be all the red neighbors of v, and B all the blue neighbors. Either $|R| \geq (t-1)/2$ or $|B| \geq (t-1)/2$. Take the larger set R or B and recurse on the subgraph on those vertices to obtain a sequence $v_2, v_3, \ldots$.

Since each step reduces the number of vertices by (roughly) half, the algorithm applied to K_N will produce a sequence $v_1, v_2, \ldots, v_\ell$ of length $\ell = \Omega(\log N)$. These vertices have the property that for each $i \in [\ell - 1]$, the edges from vertex v_i to later vertices v_j, $j > i$ are either all red, or all blue.

Let V_R be the set of vertices v_i such that the edges to v_j, $j > i$ are all red, and let V_B be the set of vertices v_i such that the edges to v_j, $j > i$ are all blue. The vertices of V_R form a complete graph with all red edges, and the vertices of V_B form a complete graph with all blue edges. One of these sets has cardinality at least $\ell/2 = \Omega(\log N)$. Choosing $N = 2^{Cn}$ for sufficiently large constant C guarantees that the desired subgraph of cardinality n exists.

The theorem for larger values of c can be proved by induction, with the base case being the two-color case. ■

The hypergraph Ramsey theorem is often thought of as an "extension" of the standard Ramsey theorem (at least, this is how it is presented on Wikipedia), but in fact Ramsey only used the "standard" theorem as a step on the way to the hypergraph theorem, which was the one he actually needed.

Theorem 3 (Hypergraph Ramsey Theorem)**.** *For every $n, k, c \in \mathbb{N}$, there exists $N \coloneqq N(n,k,c)$ such that for every coloring* $\mathsf{col}\colon \binom{[N]}{k} \to \{0,1\}^c$*, there is a subset $T \subset [N]$ of cardinality $|T| = n$ where $\forall S, S' \in \binom{T}{k}$,* $\mathsf{col}(S) = \mathsf{col}(S')$.

In this note we will use only the infinite Ramsey theorem, but the hypergraph Ramsey theorem is often more appropriate. We will actually need a small extension of these theorems:

Exercise 4. Prove the following extension of the hypergraph Ramsey theorem. *For every $n, k, c \in \mathbb{N}$, there exists $N \coloneqq N(n,k,c)$ such that for every coloring* $\mathsf{col}\colon \binom{[N]}{\leq k} \to \{0,1\}^c$*, there is a subset $T \subset [N]$ of cardinality $|T| = n$ where $\forall S, S' \in \binom{T}{\leq k}$ with $|S| = |S'|$ it holds that* $\mathsf{col}(S) = \mathsf{col}(S')$.

(The analogous extension for the infinite Ramsey theorem can be proved in the same way.)

Exercise 5. Show that the condition $|S| = |S'|$ in Exercise 4 is necessary, i.e. that for large enough c, there is a coloring $\mathsf{col}\colon \binom{[N]}{\leq k} \to \{0,1\}^c$ such that, for every $T \subset [N]$ of cardinality $|T| = n$, one can find two sets $S, S' \in \binom{T}{\leq k}$ with $\mathsf{col}(S) \neq \mathsf{col}(S')$.

3 Application I: Distribution Testing

Our first example application comes from *distribution testing*. There are several papers using Ramsey theory to get lower bounds in distribution testing, e.g. [DKN15, DKL24, CW20]. We will use an example from work of Canonne & Wimmer [CW20] about *testing data binning*, because this example probably requires the fewest definitions.

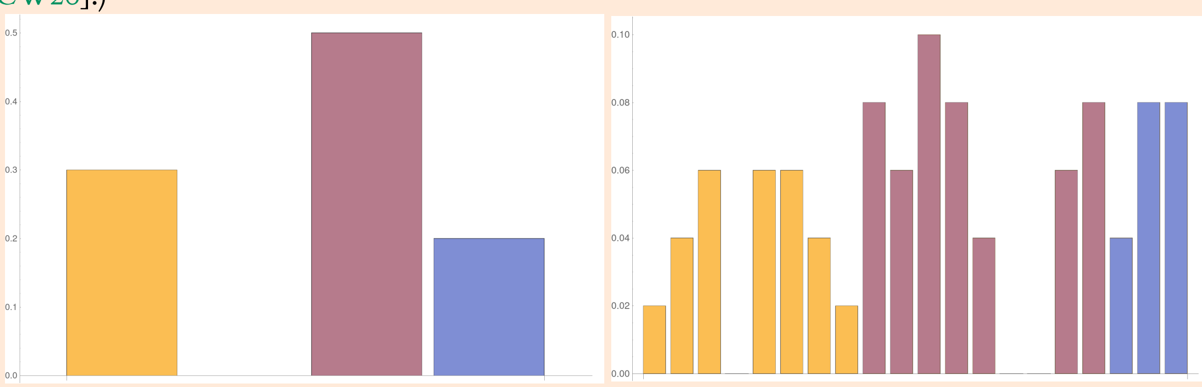
The problem is this: we have some distribution q over $[k]$ in mind, and we have sample access to an unknown distribution p over $\mathbb{N}$. We want to know if p is a "refinement" of q. By "refinement" we mean that we can take p and partition it into intervals, so that lumping together the probabilities in each interval gives us q. For example, if q is uniform over $[k]$, we want to know whether we can evenly divide p into k equally-weighted intervals.

Definition 6 (Refinement)**.** For probability distributions q over $[k]$ and p over $\mathbb{N}$, we say that p is a *refinement* of q if there exists a partition of $\mathbb{N}$ into k intervals $I_1, I_2, \ldots, I_k$, in such a way that

$$\forall i \in [k] : \qquad p[I_i] = q_i \,,$$

where $p[I] := \sum_{j \in I} p_j$ is the total probability mass contained in interval I.

Example 7. On the left, a distribution q over $[k]$. On the right, a refinement p of q. (Figures from [CW20].)



We want to test whether an unknown distribution is a refinement of q, or far (in total variation distance) from being a refinement of q. For fixed distribution q over $[k]$, we say that an algorithm A is refinement tester[3] for q with sample complexity m if, for all distributions p,

1. If p is a refinement of q, $\Pr_S[A(S) = 1] \geq 3/4$; and
2. If p is ϵ-far from any refinement of q, $\Pr_S[A(S) = 0] \geq 3/4$,

where S is a multiset of m independent random samples from p.

Question 8. *What is the minimum $m := m(k, \epsilon)$ such that for every distribution q over $[k]$, there is a refinement tester for q with sample complexity m?*

A useful observation about Question 8 is that our desired output from the algorithm is unaffected by certain transformations of the unknown distribution p. For example, suppose we take the distribution p and move all

[3] I am changing the terminology from [CW20] so that instead of "testing identity up to binning" we are "testing refinement". If apologies are necessary, they are also provided.

the values t spaces to the right: define a new distribution p' as

$$p'_i := \begin{cases} 0 & \text{if } i \le t \\ p_{i-t} & \text{if } i > t\,. \end{cases}$$

If p is a refinement of q then so is p'. Also, if p is ϵ-far from being a refinement of q then so is p'. So the desired output of the algorithm does not change when we translate the input.

More generally, the algorithm's output should be invariant under any order-preserving transformation of the input. A map $\phi : \mathbb{N} \to \mathbb{N}$ is order-preserving if $\phi(i) < \phi(i+1)$ for all $i \in \mathbb{N}$. For any distribution p over $\mathbb{N}$, write ϕp as the distribution defined by

$$(\phi p)_i := \begin{cases} 0 & \text{if } i \text{ is not in the image of } \phi \\ p_{\phi^{-1}(i)} & \text{otherwise.} \end{cases}$$

Intuitively, we therefore also expect that the algorithm's decision should be unaffected by order-preserving transformations of the *samples* it receives. Suppose the algorithm receives a random sample $S = \{s_1, s_2, \dots, s_m\}$ (which is a multiset) and outputs the answer $A(S)$ depending on S. If ϕ is any order-preserving transformation, applying ϕ to the *distribution* p does not change the desired output, so it seems like applying ϕ to the *sample* should not change the actualy output. That is, we want to say that for all samples S and all order-preserving transformations ϕ,

$$A(S) = A(\phi S) \text{ where } \phi S := \{\phi(s) \mid s \in S\}\,.$$

Unfortunately, algorithms are not guaranteed to behave this way. But we can force them to behave this way using the infinite Ramsey theorem!

Lemma 9. *Fix any distribution q over $[k]$ and suppose there is a refinement tester for q with sample complexity m. Then there is a refinement tester A for q such that for every sample S and every order-preserving map ϕ, $A(S) = A(\phi S)$.*

Proof. Let B be a refinement tester for q with sample complexity m. We identify any sample S (which is a multiset) with a pair (X, μ) where $X \subset \mathbb{N}$ is a set (not a multiset) and $\mu \in [m]^{|S|}$ is the sequence of multiplicities of elements in S, i.e. if we write $X = \{x_1, x_2, \dots, x_t\}$ in sorted order so that $x_1 < x_2 < \cdots < x_t$, then the i^{th} element x_i of X appears in S with multiplicity μ_i.

We can therefore identify any algorithm B with the function $B(X, \mu)$ which takes as input a set $X \subset \mathbb{N}$ and a sequence μ of multiplicities, which uniquely determines a multiset on $\mathbb{N}$ with cardinality $\sum_i \mu_i$.

Our goal is to apply the infinite Ramsey theorem (Theorem 1) to obtain a new algorithm A whose decision $A(X, \mu)$ depends *only* on μ. To accomplish this, we think of the algorithm B itself as a "coloring" of subsets of $\mathbb{N}$. Think of this coloring as a complete description of "what B would do" if it was given a sample multiset S supported on the set $X \subset \mathbb{N}$.

Definition 10 (Coloring)**.** For every subset $X \subset \mathbb{N}$ with cardinality $|X| \le m$, we define a color $\mathsf{col}(X)$ as a binary string $\mathsf{col}(X) \in \{0,1\}^c$ (where c is to be determined below) as follows. Let $t := |X|$. Then for every $\mu \in [m]^t$ with $\sum_i \mu_i = m$, we append to $\mathsf{col}(X)$ a description of the pair

$$(\mu, B(X, \mu))\,.$$

In other words, "If B receives X with multiplicities μ, it outputs $B(X, \mu)$". This requires at most $1 + t\lceil \log m \rceil$ bits for each μ, and therefore at most $m^t(1 + t\lceil \log m \rceil) \le m^m(1 + m\lceil \log m \rceil)$ bits in total; so we can set $c := m^m(1 + m\lceil \log m \rceil)$.

Now we apply the infinite Ramsey theorem (the version in Exercise 4) to obtain an infinite set $T \subset \mathbb{N}$ such that for all $X, X' \in \binom{T}{\leq m}$, if $|X| = |X'|$ then $\mathsf{col}(X) = \mathsf{col}(X')$.

≫ Consider the operation of B on samples drawn from T. For any two $X, X' \in \binom{T}{\leq m}$ with $|X| = |X'|$, if B receives X with multiplicities μ then it outputs $B(X, \mu)$, and if B receives X' with multiplicities μ then it outputs $B(X', \mu)$. Since $\mathsf{col}(X) = \mathsf{col}(X')$, we must have $B(X, \mu) = B(X', \mu)$. So B is invariant under order-preserving transformations of the sample *within* T.

Now we define a new algorithm A as follows. Since T is infinite, there is an order-preserving injective map $\phi : \mathbb{N} \to T$. Given sample S supported on any subset $X \subset \mathbb{N}$ with multiplicities μ, we define

$$A(X, \mu) \coloneqq B(\phi X, \mu) \text{ where } \phi X \coloneqq \{\phi(x) \mid x \in X\} .$$

We must first check that this is a refinement tester for q. Suppose p is a refinement of q. Then ϕp is also a refinement of q, and $S', \phi S$ are identically distributed when S' is m samples from ϕp and S is m samples drawn from p. So

$$\Pr_S [A(S) = 1] = \Pr_S [B(\phi S)] = \Pr_{S'} [B(S')] \geq 3/4 .$$

By a similar argument, $\Pr_S [A(S) = 0] \geq 3/4$ when p is ϵ-far from any refinement of q.

Now we check that the output of A is invariant under order-preserving transformations of the sample. Consider any order-preserving transformation ψ. For any multiset $S = (X, \mu)$, we have $\psi S = (\psi X, \mu)$. So

$$\begin{aligned} A(\psi S) = A(\psi X, \mu) &= B(\phi(\psi X), \mu) \\ &= B(\phi X, \mu) && \text{(Since } |\phi(\psi X)| = |\phi X| \text{ and both are subsets of } T) \\ &= A(X, \mu) = A(\phi S) . \end{aligned}$$

This concludes the proof. ∎

Canonne & Wimmer use (a version of) Lemma 9 to prove their lower bound:

Theorem 11 ([CW20]). *Let $\epsilon > 0$ be a sufficiently small constant. For every $k \in \mathbb{N}$ there exists a distribution q over $[k]$ such that any refinement tester for q requires $m = \Omega(k^{1-o(1)})$.*

4 Application II: Testing Order-Based Properties of Sequences

Our second application is to testing properties of sequences. This application is due to Fischer [Fis04], which appears to be the first application of Ramsey theory to prove lower bounds in sublinear algorithms, although Fischer cites [BCDM97], who prove lower bounds for parallel random access machines (PRAMs).

Definition 12 (Order-Based Properties). We say two integer sequences $x, y \in \mathbb{N}^n$ are *order-equivalent* if the relative order of their elements is the same, i.e.

$$\forall i, j \in [n] \ : \qquad x_i < x_j \iff y_i < y_j, \text{ and } x_i > x_j \iff y_i > y_j .$$

A property $\mathcal{P} \subset \mathbb{N}^*$ of integer sequences is an *order-based* property if, for every two order-equivalent sequences $x, y \in \mathbb{N}^n$,

$$x \in \mathcal{P} \iff y \in \mathcal{P} .$$

In other words, membership in $\mathcal{P}$ is determined only by the relative order of the elements in the sequence.

Example 13 (Monotonicity)**.** The most common example of an order-based property is *monotonicity*. A sequence $x \in \mathbb{N}^n$ is *monotone* if and only if $x_i < x_j \implies i < j$.

We want to prove lower bounds for testing order-based properties. In the previous example, the algorithm only received random samples from the input, and its decision was deterministic given those samples. In this example, the algorithms are allowed to make queries to the input sequence $x \in \mathbb{N}^n$, they are randomized, and they may even be adaptive.

As in the first application, our desired output from the algorithm is invariant under a certain class of transformations of the input. For any order-preserving map $\phi : \mathbb{N} \to \mathbb{N}$ and any sequence $x \in \mathbb{N}^n$, define ϕx as the sequence

$$\phi x := (\phi(x_1), \phi(x_2), \dots, \phi(x_n)) .$$

Note that ϕx has the same relative order between elements as does x, so in particular $\phi x \in \mathcal{P} \iff x \in \mathcal{P}$ when $\mathcal{P}$ is an order-based property. Equally important, if x is *ϵ-far* from $\mathcal{P}$ in Hamming distance, then so is ϕx; i.e. for $x, y \in \mathbb{N}^n$ we define

$$\mathrm{dist}(x, y) := |\{i \in [n] \mid x_i \neq y_i\}| ,$$

and we say x is ϵ-far from $\mathcal{P}$ if

$$\forall y \in \mathcal{P} , \qquad \mathrm{dist}(x, y) \geq \epsilon n .$$

If x is ϵ-far from $\mathcal{P}$ then so is ϕx. This means that the desired output of a tester for $\mathcal{P}$ is invariant under order-preserving maps. Intuitively then, we expect the "operation" of the tester to be invariant under order-preserving maps – i.e. the algorithm's query and decision strategy should depend only on the relative order of elements of x, and **not** on the specific values x_i.

Formally, for each $x \in \mathbb{N}^n$, a (non-adaptive or adaptive) randomized q-query algorithm A defines a probability distribution $\mathsf{queries}_{A,x}$ over sequences $(i_1, \dots, i_q, z) \in [n]^q \times \{0, 1\}$ of query positions i_j and output decisions z. We will show that testers for any order-invariant property $\mathcal{P}$ can be transformed into testers where this distribution of queries is invariant under order-preserving transformations of the input.

Lemma 14. *Let $\mathcal{P} \subset \mathbb{N}^*$ be any order-invariant property, and let A be any tester for $\mathcal{P}$ which makes $q := q(n, \epsilon)$ queries on inputs $x \in \mathbb{N}^n$ of length n, and succeeds with probability $3/4$. Then for every $\delta > 0$, there exists a q-query tester B for $\mathcal{P}$ which succeeds with probability $3/4 - \delta$, such that, for every two order-equivalent sequences $x, y \in \mathbb{N}^n$,*

$$\mathsf{queries}_{B,x} = \mathsf{queries}_{B,y} .$$

Proof sketch. We will again use the infinite Ramsey theorem. It would suffice to find an infinite set $T \subset \mathbb{N}$ such that $\mathsf{queries}_{A,x} = \mathsf{queries}_{A,y}$ for every order-equivalent pair x, y *whose entries all belong to* T. From here, we can define a new tester B by taking any order-preserving map $\phi : \mathbb{N} \to T$ and then, on input $x \in \mathbb{N}^n$, we simulate A on input ϕx. The argument will then proceed similarly to the distribution testing example.

To complete this argument, we would like to design a coloring $\mathsf{col} : \binom{\mathbb{N}}{\leq n} \to \{0, 1\}^c$ where, for any set $S \subset \mathbb{N}$ of cardinality $|S| \leq n$, $\mathsf{col}(S)$ encodes every possible distribution $\mathsf{queries}_{A,x}$ for the strings $x \in S^n$ (i.e. the sequences x whose entries belong to S). Note that, to apply the infinite Ramsey theorem, it is only necessary that the number of bits c in this encoding is *finite* (e.g. it is OK if c depends on the finite input size n). However, the set of all possible probability distributions over query-output sequences $(i_1, i_2, \dots, i_q, z) \in [n]^q \times \{0, 1\}$ is not finite.

Exercise 15. Complete the proof by finding an infinite set $T \subset \mathbb{N}$ such that $\mathsf{queries}^*_{A,x} = \mathsf{queries}^*_{A,y}$ for every order-equivalent pair $x, y \in \mathbb{N}^n$ with entries in T, where $\mathsf{queries}^*_{A,\cdot}$ can be described by finitely-many bits and is close enough to $\mathsf{queries}_{A,\cdot}$ for the proof sketch above to work[a].

[a] Hint: consult a New Man.

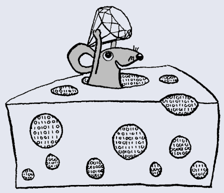
■

5 Further Reading

In this note we have only used the infinite Ramsey theorem, but some works in distribution testing have used other theorems from Ramsey theory [DKN15, DKL24].

Somewhat related to sublinear algorithms, we have used Ramsey theory in a similar way in our recent work in communication complexity [FHHH24, FGHH25]. Specifically, we used the hypergraph Ramsey theorem to show that certain communication protocols must "behave" in a way that is invariant under certain transformations of the input.

Gasarch [Gas] has a list of some applications of Ramsey theory in computer science.



Sublinear Algorithms Hidden Gem 4

Area Theorems for the CDF Curve

Lecturer: Santhoshini Velusamy. *Scribe:* Santhoshini Velusamy

June 5, 2024

1 Introduction

In Algorithmic Game Theory, there are many situations where we want to analyze the outcome of certain games even when it is extremely hard to compute their equilibrium. For instance, in auction theory, one of the most commonly used auction formats is the "first-price" auction where bidders submit bids for an item and the highest bidder wins the item and pays a price equivalent to the bid that they submit to the auctioneer. Even approximately computing the equilibrium of this auction is "PPAD-complete" [FRGH+23]. Yet this auction is widely used in practice owing to its simplicity and *credibility*.[4] In this note, we will discuss some area theorems in calculus that help analyze some parameters like the expected revenue generated by these auctions, which does not involve any direct computation of the equilibrium. My hope is that these tools may have broader applications within TCS.

2 Warm-up: A Folklore Area Theorem

We will start with a folklore theorem that is widely applied in calculus.

Theorem 1. *Let F denote the Cumulative Distribution function (CDF) of some distribution $\mathcal{D}$ supported on positive real numbers. Let f be the Probability Density function (PDF) of $\mathcal{D}$. Say F and f are integrable, then the area of the shaded region in Figure 1 is* $\mathop{\mathbb{E}}_{X\sim\mathcal{D}}[X]$.

[4] Formally, an auction is said to be credible if it is in the best interest of the auctioneer to be truthful with the bidders. In a first-price auction, an auctioneer always gains the most when they sell the item to the highest bidder and they cannot charge more than the bid of that bidder. In contrast, another popular auction format called the "second-price" auction is not credible. Here the auction rules state that the highest bidder wins and pays the second-highest bid. In sealed-bid auctions, the auctioneer always has an incentive to be dishonest about what the second highest bid was.

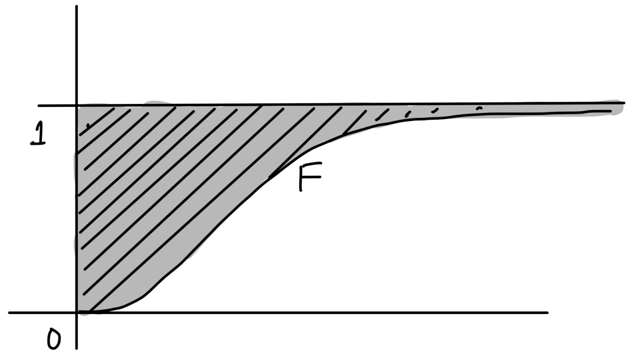


Figure 1: Area above the CDF curve is equal to the expectation of the random variable

Proof.

$$\begin{aligned}\text{Area of shaded region} &= \int_{t=0}^{\infty} (1 - F(t))\, dt \\ &= \int_{t=0}^{\infty} \Pr[X > t]\, dt \\ &= \int_{t=0}^{\infty} \left(\int_{x=t}^{\infty} f(x)\, dx\right) dt \\ &= \int_{x=0}^{\infty} \int_{t=0}^{x} f(x)\, dx\, dt \\ &= \int_{x=0}^{\infty} \left(\int_{t=0}^{x} dt\right) f(x)\, dx \\ &= \int_{x=0}^{\infty} x\, f(x)\, dx \\ &= \mathop{\mathbb{E}}_{X\sim\mathcal{D}} [X]\,.\end{aligned}$$

■

3 An Area Theorem for Welfare Optimality of First-Price Auctions

The next theorem is due to Hartline, Hoy, and Taggart [HHT14]. They use this theorem to prove the *revenue* and *welfare* optimality of first-price auctions. Following the proof of the theorem, we will discuss one of its applications.

Theorem 2 ([HHT14])**.** *Let F denote the CDF of some distribution $\mathcal{D}$ supported on positive real numbers. For $t > 0$, let $u(t) = \max_{b\le t}(t-b)F(b)$ (area of the green region in Figure 2). Say F^{-1} exists and both F and F^{-1} are integrable. We have*

$$\mathop{\mathbb{E}}_{X\sim\mathcal{D}} [X] + u(t) \ge \frac{e-1}{e} t\,,$$

for all $t > 0$.

Proof. We will first prove the following claim.

Claim 1. $\mathop{\mathbb{E}}_{X\sim\mathcal{D}} [X] \ge \int_{u(t)/t}^{1} t - \frac{u(t)}{x}\, dx$

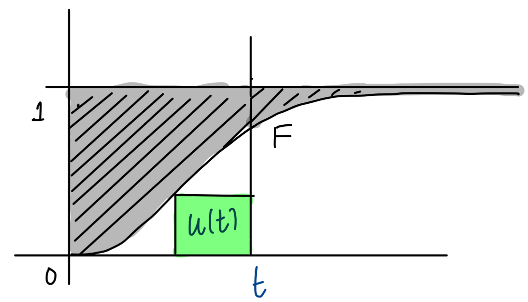


Figure 2: Area theorem from [HHT14]

Proof of Claim 1.

$$\mathop{\mathbb{E}}_{X\sim\mathcal{D}}[X] = \int_0^\infty x\, dF(x) = \int_0^1 F^{-1}(z)\, dz \geq \int_{\frac{u(t)}{t}}^1 F^{-1}(z)\, dz\,.$$

Observe that $F^{-1}(z) \geq t - \frac{u(t)}{z}$ for all z, since $u(t) \geq (t - F^{-1}(z))z$ by definition. Thus,

$$\mathop{\mathbb{E}}_{X\sim\mathcal{D}}[X] \geq \int_{\frac{u(t)}{t}}^1 \left(t - \frac{u(t)}{z}\right) dz\,.$$

∎

Applying Claim 1, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{X\sim\mathcal{D}}[X] + u(t) &\geq t + u(t)\ln\left(\frac{u(t)}{t}\right)\\ &\geq \min_y \left(t + y\ln\left(\frac{y}{t}\right)\right) = t(1-1/e)\,.\end{aligned}$$

∎

We will now discuss an application of the above theorem. In particular, we will use the above theorem to prove that first-price auctions are approximately welfare optimal.

An auction is said to achieve optimal welfare if it always assigns the item to the bidder who values it the most. Formally, we describe the value of the item to the i-th bidder by a number t_i and in the first-price auction we can assume without loss of generality that the bid b_i submitted by the bidder is at most t_i. Let $\mathcal{D}$ denote the joint value distribution of the bidders. The optimal welfare is given by

$$\mathsf{Wel}_{\mathsf{OPT}} = \mathop{\mathbb{E}}_{(t_1,\ldots,t_i,\ldots,t_n)\sim\mathcal{D}}\left[\max_i t_i\right]\,.$$

On the other hand, the welfare achieved by the first-price auction is given by

$$\mathsf{Wel}_{\mathsf{FP}} = \mathop{\mathbb{E}}_{(t_1,\ldots,t_i,\ldots,t_n)\sim\mathcal{D}}[t_{i^*}]\,,$$

where $i^* = \arg\max_i b_i$. We will show that for any *equilibrium* bid strategy[5], $\mathsf{Wel}_{\mathsf{FP}} \geq \frac{e-1}{e}\cdot\mathsf{Wel}_{\mathsf{OPT}}$. This result was originally proved by Syrgkanis and Tardos [ST13b]. Here, we will give the proof due to Hartline, Hoy, and

[5]Informally, the bidding strategies of the players is said to be in an equilibrium if every player is playing their best strategy conditioned on the strategies of the other players.

Taggart [HHT14] that uses Theorem 2. Let F_{-i} denote the CDF of the random variable $\max_{j\neq i} b_j$.[6] Applying Theorem 2 on F_{-i} we get

$$\mathbb{E}\left[\max_{j\neq i} b_j\right] + \max_{b\leq t}(t-b)\Pr[\max_{j\neq i} b_j \leq b] \geq \frac{e-1}{e}t\,,$$

for all t. It follows that

$$\mathbb{E}\left[\max_{j\neq i} b_j\right] + \max_{b\leq t_i}(t_i-b)\Pr[\max_{j\neq i} b_j \leq b] \geq \frac{e-1}{e}t_i\,.$$

Since the bidding strategies are in an equilibrium, it must be the case that bidder i places the bid that maximizes their expected "payoff", i.e., $b_i = \arg\max_{b\leq t_i}(t_i-b)\Pr\left[\max_{j\neq i} b_j \leq b\right]$. Therefore,

$$\mathbb{E}\left[\max_{j\neq i} b_j\right] + (t_i-b_i)\Pr[\max_{j\neq i} b_j \leq b_i] \geq \frac{e-1}{e}t_i\,.$$

Multiplying both sides of the inequality by $\mathbb{1}[t_i \geq \max_{j\neq i} t_j]$ and summing over all i, we get

$$\sum_i \left(\mathbb{E}\left[\max_{j\neq i} b_j\right] + (t_i-b_i)\Pr[\max_{j\neq i} b_j \leq b_i]\right)\mathbb{1}[t_i \geq \max_{j\neq i} t_j] \geq \sum_i \frac{e-1}{e}t_i\,\mathbb{1}[t_i \geq \max_{j\neq i} t_j]\,.$$

Finally, we take expectation over $\mathcal{D}$ on both sides to get

$$\mathop{\mathbb{E}}_{(t_1,\ldots,t_n)\sim\mathcal{D}}\left[\sum_i \left(\mathbb{E}\left[\max_{j\neq i} b_j\right] + (t_i-b_i)\Pr[\max_{j\neq i} b_j \leq b_i]\right)\mathbb{1}[t_i \geq \max_{j\neq i} t_j]\right] \geq \mathop{\mathbb{E}}_{(t_1,\ldots,t_n)\sim\mathcal{D}}\left[\sum_i \frac{e-1}{e}t_i\,\mathbb{1}[t_i \geq \max_{j\neq i} t_j]\right]. \tag{4}$$

We have

$$\begin{aligned}
\text{RHS of Equation (4)} &= \mathop{\mathbb{E}}_{(t_1,\ldots,t_n)\sim\mathcal{D}}\left[\sum_i \frac{e-1}{e}t_i\,\mathbb{1}[t_i \geq \max_{j\neq i} t_j]\right]\\
&= \frac{e-1}{e}\mathop{\mathbb{E}}_{(t_1,\ldots,t_n)\sim\mathcal{D}}\left[\sum_i t_i\,\mathbb{1}[t_i \geq \max_{j\neq i} t_j]\right]\\
&= \frac{e-1}{e}\mathop{\mathbb{E}}_{(t_1,\ldots,t_n)\sim\mathcal{D}}\left[\max_i t_i\right]\\
&= \frac{e-1}{e}\,\mathsf{Wel}_{\mathsf{OPT}}\,,
\end{aligned}$$

[6]Here we are considering $t_1,\ldots,t_n$ to be random variables jointly distributed according to $\mathcal{D}$ and b_i to be a deterministic function of t_i.

and

$$
\begin{aligned}
\text{LHS of } \textit{Equation } (4) &= \mathop{\mathbb{E}}_{(t_1,\dots,t_n)\sim\mathcal{D}}\left[\sum_i \left(\mathbb{E}\left[\max_{j\neq i} b_j\right] + (t_i - b_i)\Pr[\max_{j\neq i} b_j \le b_i]\right)\mathbb{1}[t_i \ge \max_{j\neq i} t_j]\right] \\
&= \mathbb{E}\left[\sum_i \mathbb{E}\left[\max_{j\neq i} b_j\right]\mathbb{1}[t_i \ge \max_{j\neq i} t_j]\right] + \mathbb{E}\left[\sum_i (t_i - b_i)\Pr[\max_{j\neq i} b_j \le b_i]\mathbb{1}[t_i \ge \max_{j\neq i} t_j]\right] \\
&\le \mathbb{E}\left[\sum_i \mathbb{E}\left[\max_{j} b_j\right]\mathbb{1}[t_i \ge \max_{j\neq i} t_j]\right] + \mathbb{E}\left[\sum_i (t_i - b_i)\Pr[\max_{j\neq i} b_j \le b_i]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\max_{j} b_j\right]\underbrace{\sum_i \mathbb{1}[t_i \ge \max_{j\neq i} t_j]}_{=1}\right] + \mathbb{E}\left[\sum_i (t_i - b_i)\ \mathbb{E}\left[\mathbb{1}\left[\max_{j\neq i} b_j \le b_i\right]\right]\right] \\
&= \mathbb{E}\left[\max_{j} b_j\right] + \mathbb{E}\left[\sum_i (t_i - b_i)\ \mathbb{1}\left[\max_{j\neq i} b_j \le b_i\right]\right] \\
&= \mathbb{E}\left[\sum_i b_i\ \mathbb{1}\left[\max_{j\neq i} b_j \le b_i\right]\right] + \mathbb{E}\left[\sum_i (t_i - b_i)\ \mathbb{1}\left[\max_{j\neq i} b_j \le b_i\right]\right] \\
&= \mathbb{E}\left[\sum_i t_i\ \mathbb{1}\left[\max_{j\neq i} b_j \le b_i\right]\right] \\
&= \mathbb{E}\left[\underbrace{t_{i^*}}_{i^*=\arg\max_i b_i}\right] = \mathsf{Wel}_{\mathsf{FP}}\,.
\end{aligned}
$$

4 An Area Theorem for Multi-Item First-Price Auctions

Finally, we conclude with another area theorem that appeared in one of our joint works with Daskalakis, Fishelson, Lucier, and Syrgkanis [DFL+22] that was useful in analyzing the revenue optimality of *multi-item* first-price auctions. However, we will not discuss the application in this lecture as it involves more intricate analysis than the one we discussed above.

Theorem 3 ([DFL+22])**.** *Let F denote the CDF of some distribution $\mathcal{D}$ supported on positive real numbers. For $t > 0$, let $u(t) = \max_{b\le t}(t-b)F(b)$ (area of the green region in Figure 3) and $a(t) = \max_{b\le t} b(1-F(b))$ (area of the blue region in Figure 3). We have*

$$\sqrt{u(t)} + \sqrt{a(t)} \ge \sqrt{t}\,,$$

for all $t > 0$.

Proof. Let us define $\overline{F}(x,t) := \frac{u(t)}{t-x}$ and $\underline{F}(x,t) := 1 - \frac{a(t)}{x}$. Observe that by definition, for all $x \in [0,t)$, we have $F(x) \le \overline{F}(x,t)$ and $F(x) \ge \underline{F}(x,t)$. Hence, for all $t > 0$ and $x \in [0,t)$, $\overline{F}(x,t) - \underline{F}(x,t) \ge 0$, and thus

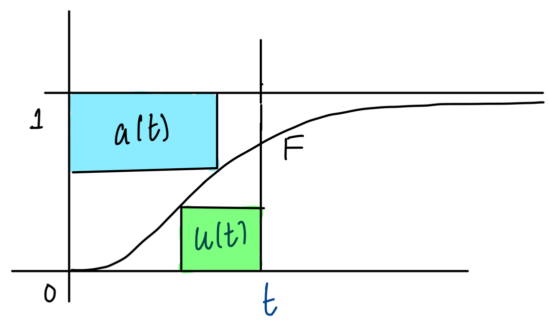


Figure 3: Area theorem from [DFL+22]

$\min_{x\in[0,t)} \overline{F}(x,t) - \underline{F}(x,t) \geq 0$. We have

$$\begin{aligned}\min_{x\in[0,t)} \overline{F}(x,t) - \underline{F}(x,t) &= \min_{x\in[0,t)]} \left(\frac{u(t)}{t-x} + \frac{a(t)}{x}\right) - 1 \\ &= \frac{(\sqrt{u(t)} + \sqrt{a(t)})^2}{t} - 1\,,\end{aligned}$$

where the minima is attained at $x = \frac{t\sqrt{a(t)}}{\sqrt{u(t)}+\sqrt{a(t)}}$. We conclude that $\sqrt{u(t)} + \sqrt{a(t)} \geq \sqrt{t}$, for all $t > 0$. ■

Sublinear Algorithms Hidden Gem 5

The Convex Order and Sampling Without Replacement

Lecturer: Guy Blanc. *Scribe:* Guy Blanc

June 11, 2024

1 Introduction

Concentration inequalities are among the most ubiquitous tools used in theoretical computer science. Perhaps the simplest setting in which they apply is when we want to understand how $\boldsymbol{X} \coloneqq \boldsymbol{x}_1 + \cdots + \boldsymbol{x}_n$ concentrates when $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ are independent. For example, the standard proof of Chernoff bounds bound the moment generating function of $\boldsymbol{X}$ using

$$[e^{\lambda(\boldsymbol{x}_1+\cdots+\boldsymbol{x}_n)}] = \prod_{i=1}^{n} \mathbb{E}[e^{\lambda \boldsymbol{x}_i}], \qquad \text{(by independence)}$$

and then applies Markov's inequality to show $\boldsymbol{X}$ is unlikely to exceed its mean. Chebyshev's inequality can be similarly applied, this time using

$$\mathrm{Var}[\boldsymbol{x}_1 + \cdots + \boldsymbol{x}_n] = \mathrm{Var}[\boldsymbol{x}_1] + \cdots + \mathrm{Var}[\boldsymbol{x}_n]\,. \qquad \text{(also by independence.)}$$

Both concentration inequalities heavily use independence. These notes are about a technique – via the *convex order* – that can often be applied in the case where $\boldsymbol{Y}$ is not a sum of *independent* random variables. In a bit more detail, these notes will give a representative example of what can be proven using the convex order. As we prove that example we will also make some general remarks about the convex order.

Remark 1 (The road not taken: Negative association)**.** Another powerful tool for showing that $\boldsymbol{Y} \coloneqq \boldsymbol{y}_1 + \cdots + \boldsymbol{y}_n$ concentrates even if $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ are not independent is via negative association. We refer the reader to this excellent manuscript, [Waj17], by David Wajc covering negative association.

2 The Example Concentration Inequality, and Why the Convex Order Helps

We will prove the following.

Theorem 2 (Concentration of sums sampled without replacement [Hoe63])**.** *For any finite and bounded population $S \in [0,1]^m$ and $n \leq m$, let $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ be a sample of n points chosen uniformly* without replacement *from S. Then, for $\boldsymbol{Y} \coloneqq \boldsymbol{y}_1 + \cdots + \boldsymbol{y}_n$,*

$$\Pr[\boldsymbol{Y} \geq [\boldsymbol{Y}] + n\varepsilon] \leq e^{-2\varepsilon^2 n}. \tag{5}$$

Notice that $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ are *not independent* since they are chosen without replacement, and yet, the concentration of Theorem 2 is as strong as if they were independent. This is no coincidence! The proof of Theorem 2 builds on the following easier concentration inequality.

Fact 3 (Hoeffding's inequality [Hoe63]). *Let* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ *be any independent random variables each bounded on* $[0, 1]$*, and* $\boldsymbol{X} \coloneqq \boldsymbol{x}_1 + \cdots + \boldsymbol{x}_n$ *their sum.*

$$\Pr[\boldsymbol{X} \geq [\boldsymbol{X}] + n\varepsilon] \leq e^{-2\varepsilon^2 n}. \tag{6}$$

We will need a bit more than Fact 3 as a black-box. The standard proof of Fact 3 follows the same blueprint as many other concentration inequalities. It first upper bounds $[e^{\lambda \boldsymbol{X}}]$ for an appropriate choice of λ and then applies Markov's inequality to this quantity. Therefore, we have the following.

Fact 4 (Generalization of Hoeffding's inequality). *Let* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ *be any independent random variables each bounded on* $[0, 1]$*, and* $\boldsymbol{X} \coloneqq \boldsymbol{x}_1 + \cdots + \boldsymbol{x}_n$*. For any* $\boldsymbol{Y}$ *satisfying* $[e^{\lambda \boldsymbol{Y}}] \leq \mathbb{E}[e^{\lambda \boldsymbol{X}}]$ *for every* $\lambda \in \mathbb{R}$*,*

$$\Pr[\boldsymbol{Y} \geq [\boldsymbol{X}] + n\varepsilon] \leq e^{-2\varepsilon^2 n}.$$

As a consequence, if we furthermore have that $[\boldsymbol{Y}] = \mathbb{E}[\boldsymbol{X}]$*, then*

$$\Pr[\boldsymbol{Y} \geq [\boldsymbol{Y}] + n\varepsilon] \leq e^{-2\varepsilon^2 n}.$$

With this version of Hoeffding's inequality, Theorem 2 is an easy consequence of the following.

Lemma 5 ([Hoe63]). *For any finite population* $S \in \mathbb{R}^m$ *and* $n \leq m$*, consider two sets of random variables.*

1. *A sample,* $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$*, drawn uniformly* without replacement *from* S*, and* $\boldsymbol{Y} \coloneqq \boldsymbol{y}_1 + \cdots \boldsymbol{y}_n$ *their sum.*
2. *A sample* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \overset{\text{iid}}{\sim} \mathrm{Unif}(S)$ *drawn uniformly* with replacement *from* S*, and* $\boldsymbol{X} \coloneqq \boldsymbol{x}_1 + \cdots + \boldsymbol{x}_n$ *their sum.*

Then for any convex $\Phi : \mathbb{R} \to \mathbb{R}$*,*

$$[\Phi(\boldsymbol{Y})] \leq \mathbb{E}[\Phi(\boldsymbol{X})].$$

Proof of Theorem 2 assuming Lemma 5. We observe that $x \mapsto e^{\lambda x}$ is convex, so

$$[e^{\lambda \boldsymbol{Y}}] \leq \mathbb{E}[e^{\lambda \boldsymbol{X}}].$$

Furthermore, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independent, and if each element of S is bounded on $[0, 1]$, then each $\boldsymbol{x}_i$ is also bounded on $[0, 1]$. Therefore, using the first part of Fact 4.

$$\Pr[\boldsymbol{Y} \geq [\boldsymbol{X}] + n\varepsilon] \leq e^{-2\varepsilon^2 n}.$$

To finish, we will argue that $[\boldsymbol{Y}] = \mathbb{E}[\boldsymbol{X}]$. Since $x \mapsto x$ is convex, we have that $\mathbb{E}[\boldsymbol{Y}] \leq \mathbb{E}[\boldsymbol{X}]$. Similarly, since $x \mapsto -x$ is convex, $\mathbb{E}[\boldsymbol{Y}] \geq \mathbb{E}[\boldsymbol{X}]$. Therefore, $\mathbb{E}[\boldsymbol{Y}] = \mathbb{E}[\boldsymbol{X}]$ and the desired result holds. ■

Hoeffding's inequality is just one example of a concentration we can transfer from $\boldsymbol{X}$ to $\boldsymbol{Y}$. Using Lemma 5, we can transfer any concentration inequality that only relies on a bound of $[\Phi(\boldsymbol{X})]$. For example, we could have similarly transferred Chebyshev's inequality using the convex function $\Phi(x) = (x - \mu)^2$, or a multiplicative form of Chernoff's inequality using the MGF again, $\Phi(x) = e^{\lambda x}$.

This ability to transfer concentration inequalities from $\boldsymbol{X}$ to $\boldsymbol{Y}$ holds whenever $\boldsymbol{Y}$ is smaller than $\boldsymbol{X}$ in the *convex order*.

Definition 6 (The convex order, see [SS07] for a textbook treament). For any real-valued $\boldsymbol{X}, \boldsymbol{Y}$, we say that $\boldsymbol{Y}$ is smaller than $\boldsymbol{X}$ in the convex order if, for all convex $\Phi : \mathbb{R} \to \mathbb{R}$,

$$[\Phi(\boldsymbol{Y})] \leq \mathbb{E}[\Phi(\boldsymbol{X})].$$

For succinctness, we denote this $\boldsymbol{Y} \leq_{\mathrm{cx}} \boldsymbol{X}$.

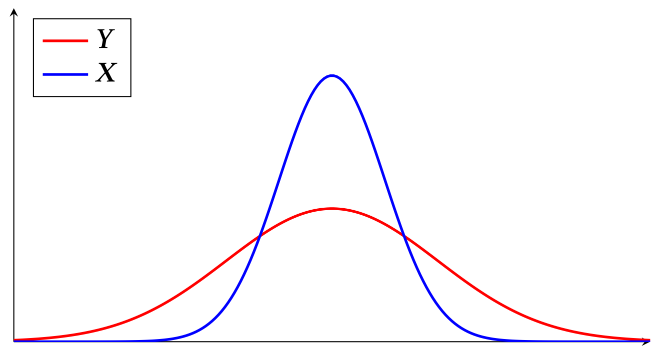


Figure 4: Example probability density functions for random variables X and Y such that $Y \leq_{\mathrm{cx}} X$.

Let us take a moment to understand what the convex order "looks like." In the proof of Theorem 2, we showed that $Y \leq_{\mathrm{cx}} X$ implies that $[X] = \mathbb{E}[Y]$, so X and Y should have the same "center." Beyond that, convex functions take on large values at the extremes, so X is more likely to take on values far from its mean. Figure 4 provides a visualization of the probability density function for a typical X, Y where $Y \leq_{\mathrm{cx}} X$. In general, I like to think of $Y \leq_{\mathrm{cx}} X$ as formalizing a notion in which Y "concentrates better" than X.

3 Couplings and their Relation to the Convex Order

One of the most direct ways of proving $Y \leq_{\mathrm{cx}} X$ is by demonstrating a coupling with the following properties.

Proposition 7 (Coupling implies convex ordering)**.** *If X and Y can be coupled $(\widehat{X}, \widehat{Y})$ so that $[\widehat{X} \mid \widehat{Y} = y] = y$ for all possible values of y then $Y \leq_{\mathrm{cx}} X$.*

Proof. This is via a simple application of Jensen's inequality. For any convex Φ,

$$\begin{aligned} [\Phi(\widehat{X})] &= \mathbb{E}_{\widehat{Y}}\left[\mathbb{E}[\Phi(\widehat{X}) \mid \widehat{Y}]\right] \\ &\geq \mathbb{E}_{\widehat{Y}}\left[\Phi(\mathbb{E}[\widehat{X} \mid \widehat{Y}])\right] && \text{(Jensen's inequality)} \\ &= \mathbb{E}_{\widehat{Y}}\left[\Phi(\widehat{Y})\right]. \end{aligned}$$

■

As an aside, the converse of Proposition 7 also holds: If $Y \leq_{\mathrm{cx}} X$, such a coupling is guaranteed to exist [Str65].[7] This converse is harder to prove than Proposition 7.

Back to our example concentration inequality, we will prove Lemma 5 by exhibiting such a coupling. This coupling is a slight simplification of the coupling used in [BHPS18] (see also the blog post of Dominic Yeo [Yeo]).

Proof of Lemma 5. We first describe the coupling. To sample $\widehat{X}$, we draw uniform and independent indices $\boldsymbol{i}_1, \dots, \boldsymbol{i}_n \overset{\text{iid}}{\sim} \mathrm{Unif}([m])$. Then, we set

$$\widehat{X} = \sum_{k \in [n]} \widehat{\boldsymbol{x}_k} \qquad \text{where} \quad \widehat{\boldsymbol{x}_k} = S_{\boldsymbol{i}_k}.$$

[7] Most computer scientists know of Strassen for his matrix multiplication algorithm, but this result is also foundational and often referred to as simply "Strassen's theorem."

To draw $\widehat{\boldsymbol{Y}}$, we do the following: Given $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_n$, let's say some $T \subseteq [m]$ is "valid" if $\boldsymbol{i}_k \in T$ for each $k \in [n]$. We sample $\widehat{\boldsymbol{T}}$ uniformly among all "valid" subsets of $[m]$ containing exactly n elements. Then, we set

$$\widehat{Y} = \sum_{j \in \widehat{T}} S_j.$$

To make things concrete, consider two examples where $m = 4$ and $n = 3$.

1. If, when determining $\widehat{\boldsymbol{X}}$, we happened to draw $\boldsymbol{i}_1 = 1, \boldsymbol{i}_2 = 2, \boldsymbol{i}_3 = 3$, there is only one valid choice for $\widehat{\boldsymbol{T}}$ (namely $\{1, 2, 3\}$) and so $\widehat{\boldsymbol{T}}$ is set to that choice with probability 1. This occurs whenever $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_n$ are distinct.
2. If we happened to draw a duplicate, say $\boldsymbol{i}_1 = 1, \boldsymbol{i}_2 = 1, \boldsymbol{i}_3 = 2$, then there is a "free slot." In this case, $\widehat{\boldsymbol{T}}$ is chosen uniformly between $\{1, 2, 3\}$ and $\{1, 2, 4\}$.

Next, we show that this is a valid coupling by arguing $\widehat{\boldsymbol{X}}$ has the same distribution as $\boldsymbol{X}$ and $\widehat{\boldsymbol{Y}}$ has the same distribution as $\boldsymbol{Y}$. For $\boldsymbol{X}$, this is quite easy. Clearly $\widehat{\boldsymbol{x}_k}$ and $\boldsymbol{x}_k$ have the same distribution (both are a uniform sample from S). Furthermore, since $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$ are independent and $(\widehat{\boldsymbol{x}_1}, \ldots, \widehat{\boldsymbol{x}_k})$ are independent, we have the following are equal in distribution,

$$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k) \overset{D}{=} (\widehat{\boldsymbol{x}_1}, \ldots, \widehat{\boldsymbol{x}_k}).$$

This implies that $\boldsymbol{X} \coloneqq \boldsymbol{x}_1 + \cdots + \boldsymbol{x}_k$ is equal in distribution to $\widehat{\boldsymbol{X}} \coloneqq \widehat{\boldsymbol{x}_1} + \cdots + \widehat{\boldsymbol{x}_n}$.

For $\boldsymbol{Y}$ the argument is a tad more delicate. Recall that $\boldsymbol{Y}$ is the sum of n points drawn uniformly and independently from S. This means that if we draw $\boldsymbol{T}$ uniformly among all size-n subsets of $[m]$, the distribution of $\sum_{j \in \boldsymbol{T}} S_j$ is equal to $\boldsymbol{Y}$'s distribution. Therefore, we need only argue that $\widehat{\boldsymbol{T}}$ is uniform over all size-n subsets of $[m]$. To prove this, note that $\widehat{\boldsymbol{T}}$ is guaranteed to be some size-n subset of $[m]$. Furthermore, by symmetry, all such subsets are equally likely, and so $\boldsymbol{Y}$ and $\widehat{\boldsymbol{Y}}$ are equal in distribution.

Lastly, we argue that for any choice of y,

$$\left[\widehat{\boldsymbol{X}} \mid \widehat{\boldsymbol{Y}} = Y\right] = Y.$$

To prove the above, suppose we condition on $\widehat{\boldsymbol{T}} = \widehat{T}$. Then, we know that $\boldsymbol{i}_k$ must be some element of $\widehat{T}$ (this is enforced based on how we drew $\widehat{\boldsymbol{T}}$). Furthermore, by symmetry, $\boldsymbol{i}_k$ is equally likely to be any element of $\widehat{T}$. Therefore,

$$\begin{aligned}
\left[\widehat{\boldsymbol{X}} \mid \widehat{\boldsymbol{Y}}\right] &= \sum_{k \in [n]} \mathbb{E}\left[\widehat{\boldsymbol{x}_k} \mid \widehat{\boldsymbol{Y}}\right] && \text{(Linearity of expectation)} \\
&= \sum_{k \in [n]} \mathbb{E}_{\widehat{\boldsymbol{T}} \mid \widehat{\boldsymbol{Y}}}\left[\mathbb{E}\left[\widehat{\boldsymbol{x}_k} \mid \widehat{\boldsymbol{T}}, \widehat{\boldsymbol{Y}}\right]\right] && \text{(Law of total expectation)} \\
&= \sum_{k \in [n]} \mathbb{E}_{\widehat{\boldsymbol{T}} \mid \widehat{\boldsymbol{Y}}}\left[\mathbb{E}\left[\widehat{\boldsymbol{x}_k} \mid \widehat{\boldsymbol{T}}\right]\right] && \text{($\widehat{\boldsymbol{Y}}$ is a deterministic function of $\widehat{\boldsymbol{T}}$)} \\
&= \sum_{k \in [n]} \mathbb{E}_{\widehat{\boldsymbol{T}} \mid \widehat{\boldsymbol{Y}}}\left[\mathbb{E}\left[S_{\boldsymbol{i}_k} \mid \widehat{\boldsymbol{T}}\right]\right] && \text{(Definition of $\widehat{\boldsymbol{x}_k}$)} \\
&= \sum_{k \in [n]} \mathbb{E}_{\widehat{\boldsymbol{T}} \mid \widehat{\boldsymbol{Y}}}\left[\frac{1}{n} \sum_{j \in \widehat{\boldsymbol{T}}} S_j\right] && \text{($\boldsymbol{i}_k$ is equally likely to be any element of $\widehat{\boldsymbol{T}}$)} \\
&= \sum_{k \in [n]} \mathbb{E}_{\widehat{\boldsymbol{T}} \mid \widehat{\boldsymbol{Y}}}\left[\frac{1}{n} \cdot \widehat{\boldsymbol{Y}}\right] = \widehat{\boldsymbol{Y}} && \text{(Definition of $\widehat{\boldsymbol{Y}}$.)}
\end{aligned}$$

■

4 Limitations of the Convex Order

I previously said that $Y \leq_{\text{cx}} X$ as formalizes a notion in which Y "concentrates better" than X. With the picture from Figure 4 as motivation, one could hope that something like the following holds.

Question 8. *Is it true that for any random variables $Y \leq_{\text{cx}} X$ each with mean μ that*

$$\Pr[|Y - \mu| \geq \tau] \leq \Pr[|X - \mu| \geq \tau]$$

for every $\tau > 0$?

Unfortunately Question 8 is false. Consider the examples,

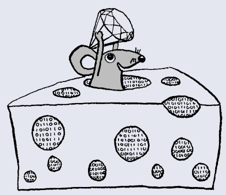
$$Y = \begin{cases} +1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2 \end{cases} \qquad \text{and} \qquad X = \begin{cases} +1/p & \text{with probability } p/2 \\ -1/p & \text{with probability } p/2 \\ 0 & \text{with probability } 1 - p. \end{cases}$$

It's not hard to show in this example that $Y \leq_{\text{cx}} X$ and that both X and Y have mean 0. Furthermore,

$$\Pr[|Y| \geq 1] = 1 \qquad \text{but} \qquad \Pr[|X| \geq 1] = p.$$

By setting p small, we observe that the probability Y is τ-far from its mean can be much larger than the probability X is τ-far from its mean, and so Question 8 is false.

Therefore, if we want a concentration inequality for X to extend to Y, we need it to have a proof structure similar to that of Hoeffding's inequality - only relying on a bound of $[\phi(X)]$ for some convex function ϕ. Fortunately, many concentrations are proven in this way and so extend from X to Y.



Sublinear Algorithms Hidden Gem 6

The Kearns-Saul Inequality

Lecturer: Abigail Gentle. *Scribe:* Sandeep Silwal

June 13, 2024

1 Introduction: the New Kid on the Block

Oftentimes, we must make use of concentration inequalities, so much so that we consider Markov, Chebyshev, Chernoff, Hoeffding, Bernstein, ... our dear friends. Well, move aside and make room for the new guest: Kearns-Saul (technically, it is named after two people). They are like the kid wearing Jordans on the court. Instead of school lunch, they packed McDonald's. Let's get to know them.

The story starts off with the following series of steps. We have a random variable X and we wish to to bound probability that X is larger than some parameter t. Then for any $\lambda > 0$, Markov's inequality implies

$$\Pr(X \geq t) = \Pr(e^{\lambda X} \geq e^{\lambda t}) \leq \mathbb{E}[e^{\lambda X}] \cdot e^{-\lambda t}.$$

Many concentration bounds follow by appropriately bounding the *moment-generating* function $\mathbb{E}[e^{\lambda X}]$ and choosing a suitable λ. Of central importance is the case of a Bernoulli distribution with mean p. Then the moment-generating function is

$$\mathbb{E}[e^{\lambda(X-p)}] = pe^{\lambda(1-p)} + (1-p)e^{-\lambda p}.$$

This is a gnarly expression, so we typically use the so called Hoeffding's lemma to bound it.

Lemma 1 (Hoeffding's lemma)**.** *Let Z be a bounded random variable with $Z \in [a, b]$. Then*

$$\mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}] \leq e^{\lambda^2(b-a)^2/8}.$$

In the Bernoulli case, the interval is $[0, 1]$ so we have

$$\mathbb{E}[e^{\lambda X}] = pe^{\lambda(1-p)} + (1-p)e^{-\lambda p} \leq e^{\lambda^2/8}.$$

In settings where p is small or large, the above bound is quite lossy since the RHS has no p dependence. The Kearns-Saul inequality provides a much better trade-off, as shown in the following plot and theorem.

Theorem 2 (Kearns-Saul)**.** *In the setting above where X is a centered Bernoulli(p) random variable,*

$$\mathbb{E}[e^{\lambda X}] \leq \exp\left(\frac{\lambda^2}{4} \cdot \frac{2p-1}{\log \frac{p}{1-p}}\right).$$

Note that at the extreme values of p, the blue curve comes quite close to the true black curve. For example when $p \to 0$, the $\log(p)$ in the denominator ensures that we get some additional savings. For example if $p \approx 2^{-k}$ for a large k, then Theorem 2 bounds the moment-generating function by $\exp(O(\lambda^2/k))$. The effect of this is that the 'variance' is smaller by a factor of k.

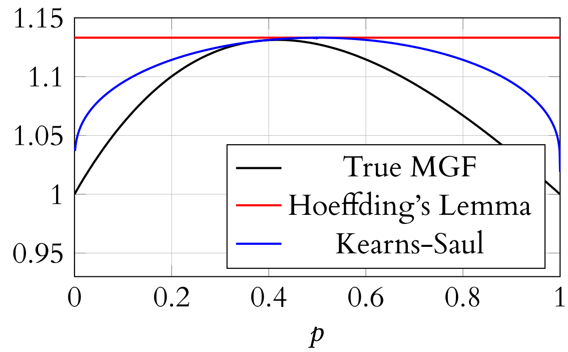


Figure 5: To see how these bounds compare, let's plot them as a function of p and take $\lambda = 1$ for simplicity.

1.1 An interesting application: The missing mass problem

Let's discuss an interesting application where Theorem 2 gives a much better bound than naively applying Chernoff bounds. These notes (and the talk) will follow the paper [BK13], with technical calculations skipped for simplicity. We refer to the paper [BK13] for full details.

The missing mass problem. Let $P = (p_1, p_2, \dots,)$ be a distribution over the positive integers. Take n samples from P and consider the 'missing mass', defined as the probability of all the items that we have not seen. More precisely, let X_i be the indicator for element i denoting that it is excluded among the n samples. We are interested in the concentration of the random variable

$$X = \sum_i p_i \cdot X_i.$$

Note that each $X_i \sim \text{Bernoulli}((1-p_i)^n)$. For simplicity, we will assume the X_i are independent in this section. This is not formally true, but since the X_i's are negatively associated (intuitively one being larger/smaller means the rest are smaller/larger respectively), it turns out it is OK to assume they are independent to derive a concentration result. To quote [BK13], "A basic fact about negative association is that it is "at least as good as independence" as far as concentration about the mean is concerned." For a formal reduction, see [MO03]. We will use Kearns-Saul to show the following.

Lemma 3. $\Pr(X \geq \mathbb{E}[X] + \alpha) \leq e^{-\Omega(\alpha^2 n)}$.

Let $\lambda > 0$ be a parameter we choose later and $Y_i = X_i - \mathbb{E}[X_i]$. We have

$$\begin{aligned}\Pr(X \geq \mathbb{E}[X] + \alpha) &\leq \Pr(\sum_i p_i(X_i - \mathbb{E}[X_i]) \geq \alpha) \\ &\leq \Pr(\exp(\lambda \cdot \sum_i p_i Y_i) \geq \exp(\lambda\alpha)) \\ &\leq \exp(-\lambda\alpha) \cdot \prod_{j \geq 1} \mathbb{E}[\exp(\lambda p_j Y_j)],\end{aligned}$$

where the last step follows from our assumption that X_i are independent and Markov's inequality. If we were to use Hoeffding's lemma, each term in the product would be bounded by $\exp(\lambda^2 p_j/8)$, giving

$$\Pr(X \geq \mathbb{E}[X] + \alpha) \leq \exp(-\lambda\alpha + \lambda^2/8),$$

which is bounded by $\exp(-\Omega(\alpha^2))$. If we were to use fancier versions of Chernoff bounds or Bernstein, it turns out they are also insufficient: we don't even seen the n dependency! For such calculations using other 'standard' concentration inequalities, see [MO03].

Now let's use Kearns-Saul. Letting $q_i = (1-p_i)^n$ and substituting Theorem 2 (since X_i are Bernoulli(q)) gives

$$\prod_{j\geq 1}\mathbb{E}[\exp(\lambda p_j Y_j)] \leq \prod_{j\geq 1}\exp\left(\frac{\lambda^2 p_i^2}{4}\cdot\frac{2q_i-1}{\log\frac{q_i}{1-q_i}}\right).$$

Bounding this expression is not easy, but after a lot of painful calculations in [BK13], we arrive at

$$\exp\left(\frac{\lambda^2 p_i^2}{4}\cdot\frac{2q_i-1}{\log\frac{q_i}{1-q_i}}\right) \leq \exp\left(\frac{\lambda^2 p_i}{4n}\right),$$

giving us

$$\prod_{j\geq 1}\exp\left(\frac{\lambda^2 p_i^2}{4}\cdot\frac{2q_i-1}{\log\frac{q_i}{1-q_i}}\right) \leq \exp(-\lambda^2/4n),$$

and hence

$$\Pr(X \geq \mathbb{E}[X]+\alpha) \leq \exp(-\lambda\alpha + \lambda^2/4n),$$

so Lemma 3 follows by setting $\lambda = \Theta(\alpha n)$. Phew!

2 An Approach without Kearns-Saul

We now discuss a more elementary approach that only relies on 'basic' concentration inequalities but gets us most of the way to what Kearns-Saul gives us. This part was not in the original talk, but I (scribe) thought it would be instructive to have it here.

Recall that Kearns-Saul gives $\Pr(X \geq \mathbb{E}[X]+\alpha) \leq e^{-\Omega(\alpha^2 n)}$. We instead show

$$\forall \alpha \geq \frac{(\log n)^2}{\sqrt{n}}, \quad \Pr(X \geq \mathbb{E}[X]+\alpha) \leq e^{-\Omega\left(\frac{\alpha^2 n}{(\log n)^2}\right)} + e^{-\Omega(\sqrt{n}\log n)}.$$

This range of α is not too restrictive, as even the Kearns-Saul inequality is only meaningful for $\alpha \geq 1/\sqrt{n}$.

Now let C be a large constant. For any i, we have $1 = 1\left[p_i \leq \frac{C\log(n)}{n}\right] + 1\left[p_i > \frac{C\log(n)}{n}\right]$, so we can decompose X as

$$X = \underbrace{\sum_i p_i X_i \cdot 1\left[p_i \leq \frac{C\log(n)}{n}\right]}_{X_1} + \underbrace{\sum_i p_i X_i \cdot 1\left[p_i > \frac{C\log(n)}{n}\right]}_{X_2}.$$

We will individually bound the probability that $X_1 \geq \mathbb{E}[X_1]+\alpha$ and $X_2 \geq \mathbb{E}[X_2]+\alpha$. Their sum upper bounds the probability that $X \geq \mathbb{E}[X]+2\alpha$.

For X_1, we use the standard McDiarmid's inequality. Changing any of the n samples can only change the sum X_1 by at most $C\log(n)/n$. Thus, McDiarmid's immediately implies that

$$\Pr(X_1 \geq \mathbb{E}[X_1]+\alpha) \leq \exp\left(-\frac{2\alpha^2}{n\cdot\left(\frac{C\log(n)}{n}\right)^2}\right) \leq e^{-\Omega\left(\frac{\alpha^2 n}{(\log n)^2}\right)}.$$

For X_2, we work a little harder. For $k \geq 0$, let

$$S_k = \left\{ i \mid p_i \in \left[\frac{C\log(n)\cdot 2^k}{n}, \frac{C\log(n)\cdot 2^{k+1}}{n}\right]\right\}.$$

Note there are at most $O(\log n)$ different k's. Let $\mathscr{E}_k$ be the event that more than $\sqrt{n}/2^k$ domain elements in S_k are missed in our n samples. We show that $\sum_k \Pr(\mathscr{E}_k)$ is very small momentarily, but assuming that none of the $\mathscr{E}_k$ happen, then we know that

$$X_2 \leq O\left(\sum_k \frac{\log(n) \cdot 2^k}{n} \cdot \frac{\sqrt{n}}{2^k}\right) = O\left(\frac{(\log n)^2}{\sqrt{n}}\right).$$

Now

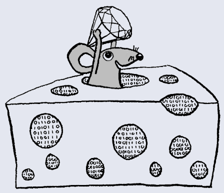
$$\Pr(\mathscr{E}_k) \leq \binom{|S_k|}{\sqrt{n}/2^k} \cdot \left(e^{-100\log(n)2^k}\right)^{\sqrt{n}/2^k},$$

since we first choose a set of $\sqrt{n}/2^k$ domain elements that are missed. Then note that the probability of missing any one of them is $(1 - p_i)^n \leq \exp(-np_i)$ and choose a very large C. These events are not independent, but missing any element makes it even less probable to miss another one by negative dependence. We have

$$\Pr(\mathscr{E}_k) \leq n^{\sqrt{n}/2^k} \cdot \left(\frac{1}{n^{1002^k}}\right)^{\sqrt{n}/2^k} \ll \frac{1}{n^{10\sqrt{n}}}.$$

So overall, $\sum_k \Pr(\mathscr{E}_k) \leq \frac{1}{n^{\sqrt{n}}}$. Thus if $\alpha \gg \frac{(\log n)^2}{\sqrt{n}}$, then we must always have $X_2 < \alpha$ if none of the $\mathscr{E}_k$ happen. Altogether, we have

$$\forall \alpha \geq \frac{(\log n)^2}{\sqrt{n}}, \quad \Pr(X \geq \mathbb{E}[X] + \alpha) \leq e^{-\Omega\left(\frac{\alpha^2 n}{(\log n)^2}\right)} + e^{-\Omega(\sqrt{n}\log n)}.$$



Sublinear Algorithms Hidden Gem 7

The Lovász-Simonovits Curve Technique

Lecturer: Akash Kumar. *Scribe:* Akash Kumar

June 11, 2024

Adapted from Akash Kumar's lecture notes at [Kum23].

1 Overview: Finding Sparse Cuts Using Random Walks

You can find a sparse cut in a graph by finding out the eigenvector $\boldsymbol{x}$ which is paired with the second smallest eigenvalue, λ_2 of the normalized Laplacian, $\boldsymbol{\mathcal{L}}$. This eigenvector based approach opens doors to what is possible and returns a cut which is guaranteed to have conductance within $O(\sqrt{\varphi(G)})$ factor of the optimum. Now, just like we do everything in life, we ask for more :-). The question that drives our discussion today is, can you recover this low conductance cut really quickly?

Of course, you need to spend time proportional to size of the set you are trying to find. So, as a loose benchmark, we ask: can you find an approximation $\widetilde{S}$ to the set S with smallest conductance which runs in time $O(|S|)$? As you can imagine, such a routine will be super handy if the set S is super small. Indeed, you expect real world networks to have small sized cuts. These sets with small conductance capture the real-life meaning of "a community". It is usually taken, in several disciplines, to refer to a set of vertices which is very "well-connected inside" and "loosely connected outside". First things first, let us make sure that the eigenvector based sweep approach cannot find this set in time proportional to only the size of this set. Indeed, you have to compute eigenvectors and even this requires you to write down the entire matrix $\boldsymbol{\mathcal{L}}$.

However, you might have heard of iterative methods to solve eigenvalue/eigenvector problems. Here, the entire idea is to find the eigenvector iteratively by doing power iteration on some appropriate matrix. Luckily for us, power iterations on $\boldsymbol{\mathcal{L}}$ bear a close resemblance to power iterations on the random walk matrix $\boldsymbol{M}$. Indeed, you might recall that these two matrices share the same eigenbasis for a (d-regular) graph and the eigenvalues of these two matrices also bear a linear relationship. So, it is fair game to expect power iterations on $\boldsymbol{M}$ to reveal information about results of running a power iteration on $\boldsymbol{\mathcal{L}}$. And power iterations on $\boldsymbol{M}$ can be simulated by just doing random walks! Does this mean we can expect a random walk based approach to be able to find sparse cuts in graphs? As you will see, in a truly remarkable feat, Spielman-Teng [ST13a] made this high level plan work. Just one final note in this (super) high level story. Perhaps this is not that unexpected. Indeed, you might expect that if you truly have a low conductance cut (say community of people who play candy crush on the Facebook graph – this is a low conductance cut, you guys don't have too many friends) then a random walk which is started at a vertex within the community is unlikely to ever leave this community. And thus after doing enough of these walks and collecting all the end-points reached, intuitively you should be able to find a low conductance cut. Ok, *enough chitchat*, let us setup the stage to work out the relevant math.

2 The Lovász-Simonovits Theorem

Consider a connected graph $G' = (V, E')$. Define a (directed) graph $G'' = (V, E'')$ where G'' is obtained by taking an edge $(u, v) \in E'$ and making two copies of it (one in each direction). Finally, define a (lazy) graph $G = (V, E)$ where G is obtained by adding d_u loops at vertex u where d_u is the outdegree of u. Let $|E| = 2m$ where m denotes the number of self-loops in this graph. Recall from last lecture that the distribution $\boldsymbol{\pi}$ where

$\boldsymbol{\pi}(v) = \deg(v)/2|E|$ satisfies the detailed balance condition for G and thus $\boldsymbol{\pi}$ is the stationary distribution for random walks on G. Let $\boldsymbol{\pi}$ denote the "current distribution". For a vertex $u \in V$ and $v \in N(u)$, the probability mass sent by u along the edge (u, v) with respect to the current distribution $\boldsymbol{\pi}$ equals $\frac{\boldsymbol{\pi}(u)}{\deg(u)}$. Since $\boldsymbol{\pi}(u) \propto d_u$ (where d_u is the degree of u in G' or the out-degree of u in G''), one notices that this distribution is uniform on all the edges of G. That is,

$$p(u, v) = \boldsymbol{\pi}(u)/\deg_G(u) = \boldsymbol{\pi}(u)/2d_u = 1/2m$$

for all directed edges out of u in G.

Next, take a distribution $\boldsymbol{p}$ (different from $\boldsymbol{\pi}$) supported on vertices of G. This distribution also induces a natural distribution $\boldsymbol{\rho}$ on $E(G)$. For any edge e directed out of $u \in V$, we have the probability mass about to be dispatched along the edge e (with respect to $\boldsymbol{p}$) equals

$$\boldsymbol{\rho}(e) = \frac{\boldsymbol{p}(u)}{d_G(u)}.$$

Note that this value, the probability dispatched by u along any edge out of u, does not depend on the other end-point of this edge. This is nice, as we now have a distribution on edges coming from $\boldsymbol{p}$ and the other from $\boldsymbol{\pi}$. One natural approach to show convergence of random walks to the distribution $\boldsymbol{\pi}$ on vertices is to instead show convergence to the distribution on edges induced by $\boldsymbol{\pi}$. We would like to show convergence time bounds (and find sparse cuts in G) by using this framework of thinking about these induced edge distributions instead. We already discussed an approach to bounding the mixing time in the last lecture (which was expressed in terms of the gap $\lambda_1(\boldsymbol{M}) - \lambda_2(\boldsymbol{M})$, where $\boldsymbol{M} = \boldsymbol{M}(G)$ denotes the random walk matrix on G). This approach tracked a single number: the ℓ_2 distance to the stationary at the current step. The key idea of Lovász and Simonovits was the suggestion that instead of tracking a single number, you could track an entire curve and use this information to bound the mixing time.

So, following Lovász-Simonovits, we now define a "greedy" potential function.

Definition 1. Let $\boldsymbol{p}$ denote a distribution on $V(G)$. Denote the distribution induced on $E(G)$ as $\boldsymbol{\rho}$. Sort the edges in decreasing order of mass dispatched along the eges (with respect to $\boldsymbol{p}$) to get

$$\boldsymbol{\rho}(e_1) \geq \boldsymbol{\rho}(e_2) \cdots \boldsymbol{\rho}(e_{2m}).$$

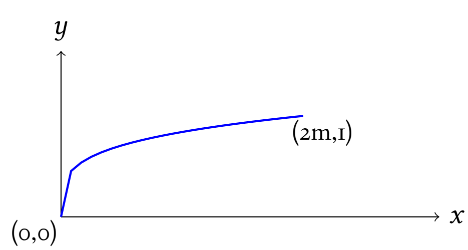
Then the discrete Lovász Simonovits potential, denoted $G_{\boldsymbol{p}}$ is a function defined as follows:

$$G_{\boldsymbol{p}}\{0, 1, 2, \cdots, 2m\} \to [0, 1] \text{ where } G_{\boldsymbol{p}}(x) = \sum_{i=1}^{x} \boldsymbol{\rho}(e_i).$$

We define the Lovász-Simonovits potential $g_{\boldsymbol{p}}$ as an extension of $G_{\boldsymbol{p}}$ to the domain $[0, 2m]$ by linearly interpolating $G_{\boldsymbol{p}}$ between two successive integer values $i, i + 1$ where $i \in \{0, 1, 2, \cdots, 2m - 1\}$. Often, we will also write $g_{\boldsymbol{p}}$ as $g_{\boldsymbol{\rho}}$.

We first make a few remarks to help digest this definition.

Remark 2. *Note that the potential defined above is truly a greedy potential. After all, it "greedily aggregates" the x heaviest elements with respect to $\boldsymbol{\rho}$ and sums them up which results in $g_{\boldsymbol{p}}(x)$. One thing you might find annoying is the use of x to count number of edges* [8]. *Finally, you note that the curve $g_{\boldsymbol{p}}$ is increasing (or non-decreasing) and concave (why?). Here is a pictorial illustration.*



[8] Get over it :-) It is not such a big thing.

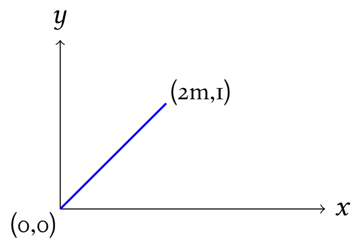
Note that $g_{\boldsymbol{p}}(0) = 0$ and $g_{\boldsymbol{p}}(2m) = 1$.

Next, I would like you to verify the following.

Problem 3. Let $\boldsymbol{\pi}$ denote the stationary distribution of random walks on $G = (V, E)$ as defined above. Show that

$$g_{\boldsymbol{\pi}}(x) = \frac{x}{2m}, \quad \forall 0 \le x \le 2m.$$

That is, it is given by the following picture.



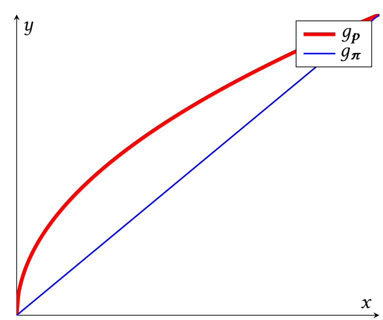
We now solve a problem through which we confirm an interesting fact. The Lovász-Simonovits potential gives some kind of special treatment to the stationary distribution, $\boldsymbol{\pi}$.

Claim 2. *For any distribution $\boldsymbol{p}$ on $V(G)$, the curve $g_{\boldsymbol{\pi}}$ lies pointwise below $g_{\boldsymbol{p}}$.*

Proof. Indeed, this problem is fairly straightforward to show if you understand the definition of Lovász-Simonovits curves and that they have this defining feature of being constructed by this greedy aggregation. You just note the following

- All curves start at $(0, 0)$ and end at $(2m, 1)$.
- You have $g_{\boldsymbol{p}}(1) > g_{\boldsymbol{\pi}}(1)$, that is the curve $g_{\boldsymbol{p}}$ starts out with a bigger slope.

Finally, you recall that all of these curves are concave. Thus, if the curve $g_{\boldsymbol{p}}$ were to ever dip below $g_{\boldsymbol{\pi}}$, it cannot reach all the way to $(2m, 1)$ without violating concavity. And this solves the problem above. The picture below illustrates this. ■



Alright, this is great. We now follow Lovász-Simonovits and ask: is it possible to argue convergence of random walks using this picture above? Namely, letting $\boldsymbol{p}_t^T = \boldsymbol{p}_{t-1}^T \boldsymbol{M}$ (where $\boldsymbol{M}$ denotes the random-walk matrix for G), is it conceivable that $g_{\boldsymbol{p}_t} \to g_{\boldsymbol{\pi}}$ as $t \to \infty$? As I already told you, the answer is yes. The rest of this lecture is focused on showing just that.

2.1 Convergence of Lovász-Simonovits curves

Let us start with a few notational simplications. First off, we will say LS instead of Lovász-Simonovits from now on. Additionally, we use another shorthand.

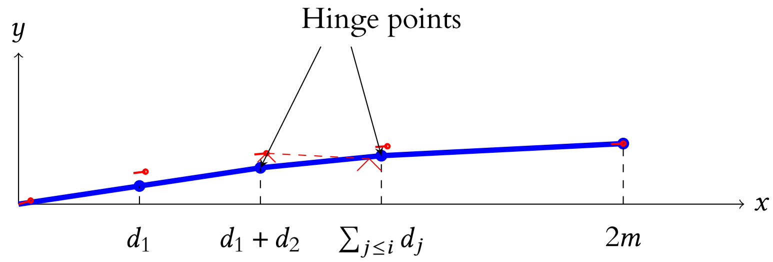


Figure 6: If g_{t-1} dominates g_t on all hinges of g_t then g_{t-1} cannot dip below g_t anywhere else without violating concavity.

Notation. Let $\boldsymbol{p}_0$ denote a point distribution of the form 1_v for some $v \in V$. We let $\boldsymbol{p}_t^T = \boldsymbol{p}_{t-1}^T \boldsymbol{M}$ and we will denote the t-step LS curve as $g_t := g_{\boldsymbol{p}_t}$.

We will prove Lemma 4 as a warmup towards showing this convergence of LS curves. It just asserts that the LS curves only fall downward with time (actually, the claim is *weaker*: it shows that the later LS curves don't rise above the former curves).

Lemma 4. *For every $t \geq 1$, for every $x \in [0, 2m]$, it holds that*

$$g_t(x) \leq g_{t-1}(x).$$

It would be helpful to make a definition before we proceed.

Definition 5. Let $\boldsymbol{p}$ denote a distribution on $V(G)$ and let $\boldsymbol{\rho}$ denote the corresponding distribution induced on edges. Arrange edges in the order

$$\boldsymbol{\rho}(e_1) \geq \boldsymbol{\rho}(e_2) \geq \cdots \geq \boldsymbol{\rho}(e_{2m})$$

where all the out-edges incident on a vertex $u \in V$ occur consecutively in this sequence (of course, these out edges occur in an arbitrary order). An integer $0 \leq x \leq 2m$ is called a hinge point with respect to $\boldsymbol{\rho}$ (or $\boldsymbol{p}$) if the x heaviest edges under $\boldsymbol{\rho}$ correspond to all the out edges (including loops) incident on some subset S of vertices. If $|S| = i$, the integer x is called the ith hinge point on $g_{\boldsymbol{\rho}}$. We call this order on edges $e_1, e_2, \cdots e_{2m}$ a hinge respecting order.

As you can see, the hinges are merely a convenient way to organize the edges with respect to the distribution $\boldsymbol{\rho}$. Note that any curve $g_{\boldsymbol{\rho}}$ has exactly $|V(G)|$ hinges. We first argue that it suffices to prove the following claim.

Claim 3. *Fix $t \geq 1$ and let $I \subseteq [2m]$ denote the subset of hinges on g_t. Suppose*

$$\forall x \in I, \text{ we have } g_t(x) \leq g_{t-1}(x).$$

Then for all $0 \leq x \leq 2m$, we have $g_t(x) \leq g_{t-1}(x)$.

Proof assuming Claim 3. Convexity (rather concavity) argument. We will show that if g_t is below g_{t-1} on two successive hinges of g_t, then it must be below g_{t-1} at all the intermediate points. To this end, recall that both the LS curves, g_{t-1} and g_t start at $(0, 0)$ and wind up at $(2m, 1)$. Consider two consecutive hinges on g_{t-1}, say x_i and x_{i+1} which are the ith and the $(i+1)$th hinge points. Take $y \in [x_{i-1}, x_i]$. Suppose for contradiction sake that $g_t(y) > g_{t-1}(y)$. We know $g_t(x_i) \leq g_{t-1}(x_i)$ and $g_t(x_{i-1}) \leq g_{t-1}(x_{i-1})$. This means there is a hinge y' of g_{t-1} which lies between x_{i-1} and x_i as well. But this would mean g_{t-1} is not concave in the interval $[y', x_i]$. Here is a picture that assists the proof. ■

So, to finish off proving Lemma 4, it suffices to show Claim 3. We do that now.

Proof of Claim 3. Call the distribution on $E(G)$ with respect to distribution $\boldsymbol{p}_t$ on vertices $\boldsymbol{\rho}$ and suppose $\boldsymbol{\rho}(e_1) \geq \boldsymbol{\rho}(e_2) \geq \cdots \geq \boldsymbol{\rho}(e_{2m})$ where the edges are ordered in some hinge respecting order. Write $e_i = (u_i, v_i)$ where several edges might share the same u_i. Let x denote a hinge point on $\boldsymbol{p}_t$ corresponding to some subset S of vertices.

$$\begin{aligned}
g_t(x) = \sum_{i=1}^{x} \boldsymbol{\rho}(e_i) = \sum_{i=1}^{x} \boldsymbol{\rho}(u_i, v_i) = \sum_{u \in S} \boldsymbol{p}_t(u) & \\
\overset{(1)}{=} \sum_{i=1}^{x} \boldsymbol{\rho}_{t-1}(\overleftarrow{e_i}) & \qquad \overleftarrow{e_i} = (v_i, u_i) \text{ denotes the reverse edge to } e_i \\
\overset{(2)}{\leq} g_{t-1}(x) & \qquad \text{by greedy property of LS curves}
\end{aligned}$$

A little more detail on the above derivation. Here, (1) holds because the mass $\boldsymbol{p}_t(u)$ deposited on each $u \in S$ is precisely the mass which reached those vertices along the in-edges to u at the previous step. For (2), note that to evaluate g_{t-1} at x you must grab x heaviest edges under $\boldsymbol{\rho}_{t-1}$ and add up the probabilities on those edges. Here, we are grabbing *some* x edges which are reversals of edges incident on S which need not be the x heaviest edges with respect to $\boldsymbol{\rho}_{t-1}$ and therefore this step follows. ∎

2.2 A brief interlude

We pause the main story for a while where the plan is to go beyond the warmup done in Lemma 4 and show that the curve should fall significantly from one step to another if you know something about the graph (like the graph expands). Through this brief interlude, we will first solidify our understanding of the LS potential function – which is now possible thanks to our little warmup proving Lemma 4. So, let us ask the following question.

Problem 6. Let $\boldsymbol{p}_t$ denote the tth step distribution on $V(G)$ and $\boldsymbol{\rho}_t$ denote the corresponding distribution induced on $E(G)$ where

$$\boldsymbol{\rho}(e_1) \geq \boldsymbol{\rho}(e_2) \geq \cdots \geq \boldsymbol{\rho}(e_{2m})$$

denotes a hinge respecting order with respect to $\boldsymbol{\rho}$. Then can you interpret $g_t(x)$ as solution to some natural ILP?

It should not surprise you that there is indeed such an ILP. Indeed, the LS curve aggregates edges greedily by heaviness with respect to $\boldsymbol{\rho}$. Taking inspiration from the fractional knapsack problem, and writing c_{e_i} for ith component of $\boldsymbol{c}$, we get:

$$\begin{aligned}
g_t(x) = \text{maximize } & \vec{\boldsymbol{c}} \cdot \vec{\boldsymbol{\rho}} \\
\text{subject to } & \quad 0 \leq c_{e_i} \leq 1 \;\; \forall 1 \leq i \leq 2m \\
& \sum_i c_{e_i} = x
\end{aligned}$$

For those among you who are not yet comfortable with ILPs, let me present this result in a different/more elementary way.

Lemma 7. *For every $c_1, c_2, \cdots, c_{2m}$ such that $0 \leq c_i \leq 1$,*

$$\sum_{i=1}^{2m} c_i \left(\boldsymbol{\rho}(e_i)\right) \leq g_t\Big(\sum_{i=1}^{2m} c_i\Big).$$

You should prove this on your own. I will give a hint: recall that LS curves are built greedily and in some sense the right hand side is the greedy maximum you get for the left hand side. Okay, this finishes the brief interlude, let's get back to the main story.

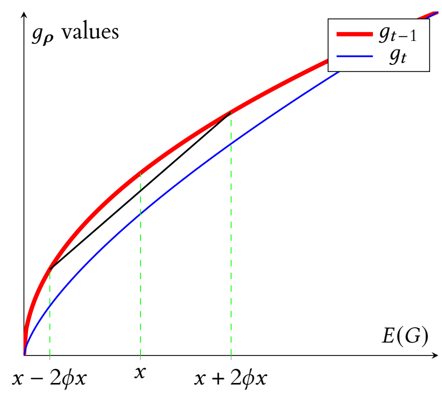
2.3 Resuming the LS convergence saga

There are two major results in this section – Theorem 8 and Theorem 9. Theorem 8 identifies a certain recurrence the LS potential functions satisfy which formalizes our intuition about the drops of the LS curve from one timestep to the next. Theorem 9 solves this recurrence to give a form more readily usable in applications. We start by stating Theorem 8.

Theorem 8. *Let $G = (V, E)$ be the graph defined before and suppose $\varphi(G) = \phi$. For every start distribution $\boldsymbol{p}_0$, for every $t \geq 1$, and every hinge point x of g_t, we have the following:*

1. *If $x \leq m$, then*
$$g_t(x) \leq \frac{g_{t-1}(x - 2\phi x) + g_{t-1}(x + 2\phi x)}{2}.$$
2. *If $x > m$, then*
$$g_t(x) \leq \frac{g_{t-1}(x - 2\phi(2m - x)) + g_{t-1}(x + 2\phi(2m - x))}{2}.$$

To understand what this theorem is saying, consider the following picture.



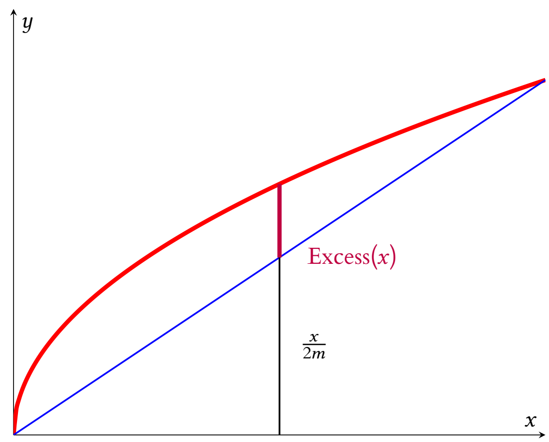
In this picture, we have two successive LS curves, g_{t-1} and g_t. What you notice is that the entire curve at time t drops significantly below the curve at time $t - 1$. And this drop is large if the conductance is large. Indeed, if the conductance is at least ϕ, you can draw chords of length $4\phi x$ below x such that at the point x, g_t is below this chord[9](!) This matches well with what we intended to show: if the conductance is large these curves not only fall, they fall *royally* and the entire sequence of curves rapidly heads towards the curve g_π. One shortcoming of this theorem is that it is not fully transparent what the rate of convergence is. The theorem below teases out this rate and provides the convergence guarantee in a form which is mathematically more transparent.

[9]For technical reasons, we show this only for hinge points on g_t. However, this is enough to show convergence as you will see in the proof of Theorem 9.

Theorem 9. *Let $G = (V, E)$ be a graph as defined above. Suppose $\varphi(G) \geq \phi$. Then, for every distribution $\boldsymbol{p}_0$, every $t \geq 1$ and every $x \in [0, 2m]$, we have*

$$g_t(x) \leq \min(\sqrt{x}, \sqrt{2m-x}) \left(1 - \frac{\phi^2}{2}\right)^t + \frac{x}{2m}.$$

We know that any distribution pointwise dominates the uniform distribution on the edges. In particular, this means the LS value of a distribution $\boldsymbol{\rho}$ on the edges at a point $0 \leq x \leq 2m$ is always at least $x/2m$ (which is the value of LS curve for uniform distribution at x). Theorem 9 asserts that if the conductance is large, this value is only slightly bigger (additively) and it is within an at most $Excess(x) \leq \sqrt{x}\left(1 - \frac{\phi^2}{2}\right)^t$ of $x/2m$ at time t provided the graph is an ϕ-expander. Let us have another picture to appreciate what this theorem asserts.



We delegate the proof of these last two theorems to the following two subsubsections.

Proof of Theorem 8

Proof. This proof is what we call a "mass reshuffling argument." I will only prove item 1. Item 2 follows by an analogous argument. Let x denote a hinge point of g_t. Then just like we did in proof of Lemma 4 (actually, proof of Claim 3), we have

$$\begin{aligned} g_t(x) &= \sum_{i=1}^{x} \boldsymbol{\rho}(e_i) \\ &= \sum_{i=1}^{x} \boldsymbol{\rho}(u_i, v_i) \\ &= \sum_{u \in S} \boldsymbol{p}_t(u) \\ &= \sum_{i=1}^{x} \boldsymbol{\rho}_{t-1}(\overleftarrow{e_i}) \quad \overleftarrow{e_i} \text{ denotes the reverse edge to } e_i \end{aligned} \tag{7}$$

We just skipped the last step where we (loosely) upper bounded the last expression in the above chain of equalities as $g_{t-1}(x)$. This time around, we know the graph is an ϕ-expander and we would like to exploit this fact to get a better bound. Let S denote the set of vertices corresponding to which x is the hinge point on g_t. That is, $\sum_{u \in S} \deg_G(u) = 2d_u = x$. Let

$$E_x = \{(u_i, v_i)\}_{u_i \in S} \text{ i.e., } \textit{set of out-edges or loops incident on some vertex in } S$$

and W is obtained by reversing them all. That is,

$$W = \{(v_i, u_i) : (u_i, v_i) \in E_x\}.$$

We divide this set of *reversed edges* W into two groups:

$$\begin{aligned} W_1 &= \{(v_i, u_i) : u_i, v_i \in S, u_i \neq v_i\} && \text{non loop edges fully in S} \\ W_2 &= \{(v_i, u_i) : u_i \in S, v_i \notin S\} \bigcup \{(u_i, u_i) : u_i \in S\} && \text{edges coming into S plus all loops} \end{aligned} \tag{8}$$

At this stage, you might want to try the following: write $\sum_{(u,v)\in W} \boldsymbol{\rho}_{t-1}(u,v) = \sum_{(u,v)\in W_1} \boldsymbol{\rho}_{t-1}(u,v) + \sum_{(u,v)\in W_2} \boldsymbol{\rho}_{t-1}(u,v)$. Note that $|W_2| \geq x/2 + \phi x$ and thus, $|W_1| \leq x/2 - \phi x$. By the greedy property of LS curves, you get

$$\sum_{(u,v)\in W_1} \boldsymbol{\rho}_{t-1}(u,v) \leq g_{t-1}(|W_1|) \leq g_{t-1}(x/2 - \phi x).$$

This is not quite what we want. We would like to "pull the one-half" out of this expression[10]. In particular, we would like to show the following:

$$\begin{aligned} \sum_{(v,u)\in W_1} \boldsymbol{\rho}_{t-1}(v,u) &\leq 1/2 \cdot g_{t-1}(x - 2\phi x) \\ \sum_{(v,u)\in W_2} \boldsymbol{\rho}_{t-1}(v,u) &\leq 1/2 \cdot g_{t-1}(x + 2\phi x) \end{aligned} \tag{9}$$

As you will see, our final proof ends up diverging from this plan. Let us press on, we will see where we diverge. Towards showing the first bound above, following Lovàsz and Simonovits, let us introduce a new set of edges. To this end, take an edge $e = (v, u) \in W_1$ and take a unique self loop e' at v and throw them both in a new set, W_1'. Note $|W_1'| = 2|W_1|$ and that $\boldsymbol{\rho}_{t-1}(e) = \boldsymbol{\rho}_{t-1}(e')$. In particular, this means $\sum_{e\in W_1'} \boldsymbol{\rho}_{t-1}(e) = 2\sum_{e\in W_1} \boldsymbol{\rho}_{t-1}(e)$. By the greedy rule again,

$$\sum_{e\in W_1} \boldsymbol{\rho}_{t-1}(e) = \frac{1}{2} \cdot \sum_{e\in W_1'} \boldsymbol{\rho}_{t-1}(e') \leq \frac{1}{2} \cdot g_{t-1}(|W_1'|)$$

Note that one can pull out the one-half factor which gives us the desired bound on the first term as the last term above satisfies $g_{t-1}(|W_1'|)|) \leq g_{t-1}(x - 2\phi x)$[11]. Lovàsz and Simonovits resist this temptation *(and this is where we diverge from the plan)* as the same argument cannot be done with W_2 as we only know $|W_2| \geq x/2 + \phi x$ and we cannot use monotonicity for showing such a bound. Instead, this will follow from a clever mass shuffling together with concavity arguments.

We again define a new set W_2' using the set W_2. Recall, W_2 contains loops and edges $e \in E(\overline{S}, S)$. As before, we take a non-loop edge $e = (v, u) \in W_2$ and we pair it with a unique loop e' at v. Next, we take a loop $e = (u, u) \in W_2$ and we pair it with a unique edge e' directed out of u[12]. We add all these e, e' pairs in W_2'. By an entirely similar argument as before, we get

$$\sum_{(v,u)\in W_2} \boldsymbol{\rho}_{t-1}(v,u) \leq \frac{1}{2} \cdot g_{t-1}(|W_2'|).$$

Thus, $g_t(x)$ lies below the "midpoint of a longer chord"[13] on g_{t-1} – namely the chord connecting

$$p_1 = \left(|W_1'|, g_{t-1}(|W_1'|)\right) \text{ and } p_2 = \left(|W_2'|, g_{t-1}(|W_2'|)\right).$$

[10]When life gives you lemons, make lemonade.

[11]By monotonicity of LS curves and recalling $|W_1'| \leq x - 2\phi x$.

[12]why is this pairing possible?

[13]If you are worried why x is the midpoint of this chord, try proving it yourselves. The proof is given in Claim 4 anyway.

And therefore, $g_t(x)$ will also lie below the "mid-point of a shorter chord" which connects

$$q_1 = (x - 2\phi x, g_{t-1}(x - 2\phi x)) \text{ and } q_2 = (x + 2\phi x, g_{t-1}(x + 2\phi x))\,.$$

∎

For those of you who want it, here is Claim 4 for completeness. Note that this claim is not standalone and it will only help if you read up proof of Theorem 8 above.

Claim 4. *Let x, W_1, W_2, W_1', W_2' be as defined in the above argument. Then*

$$|W_1'| + |W_2'| = x.$$

Proof. Recall that

$$\begin{aligned}
|W_1'| = 2|W_1| &= 2 \cdot \#\text{ of non-loop edges internal to S} \\
&= \#\text{ of non-loop edges internal to S} + \#\text{ loops internal to S} \\
&= x - |E(S, \overline{S})|
\end{aligned}$$

Since each non-loop edge has a reverse copy, $|E(S, \overline{S})| = |E(\overline{S}, S)|$ in our directed graph G. Next up, note that

$$\begin{aligned}
|W_2'| = 2|W_2| &= 2 \cdot \#\text{ loops internal to S} + 2 \cdot |E(\overline{S}, S)| \\
&= \#\text{ loops internal to S} + \#\text{ non-loop edges internal to S} + 2 \cdot |E(\overline{S}, S)| \\
&= x - |E(S, \overline{S})| + 2 \cdot |E(\overline{S}, S)| \\
&= x + |E(S, \overline{S})|
\end{aligned}$$

where recall the last step follows because $|E(S, \overline{S})| = |E(\overline{S}, S)|$. And therefore, it follows that x is the average of $|W_1'|$ and $|W_2'|$. ∎

Proof of Theorem 9

We introduce a new curve,

$$R_0(x) = \min(\sqrt{x}, \sqrt{2m - x}) + x/2m.$$

It is easy to check that $\forall x \in [0, 2m], g_0(x) \le R_0(x)$. We use define a new curve R_t where R_t's are obtained by using the same recurrence that LS curves obey. In particular, we define

$$R_t(x) = \begin{cases} \frac{1}{2}\left(R_{t-1}(x - 2\phi x) + R_{t-1}(x + 2\phi x)\right) & \text{if } x \le m \\ \frac{1}{2}\left(R_{t-1}(x - 2\phi(2m - x)) + R_{t-1}(x + 2\phi(2m - x))\right) & \text{Otherwise} \end{cases}$$

We will now show the following:

1. First, we show
$$\forall x, R_t(x) \le \min(\sqrt{x}, \sqrt{2m - x}) + x/2m.$$
2. Second, we show that $g_t(x) \le R_t(x)$ holds at all hinges of g_t.
3. Last, note that the above item implies that $g_t(x) \le R_t(x)$ holds for all x. Indeed, both the functions are concave and increasing[14]. Indeed, R_t cannot dip below g_t between two hinges of g_t without violating concavity.

[14] Q: Why are R_t's concave? This is a good exercise you should think about.

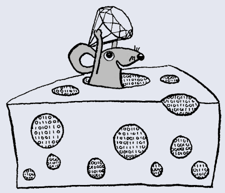
Thus, it suffices to show the first two items above. Let us start with the second item. We will show this only for hinge x of g_t where $x \le m$. We have,

$$\begin{aligned} g_t(x) &\le 1/2 \cdot (g_{t-1}(x - 2\phi x) + g_{t-1}(x + 2\phi x)) \\ &\le 1/2 \cdot (R_{t-1}(x - 2\phi x) + R_{t-1}(x + 2\phi x)) \\ &= R_t(x) \end{aligned}$$

Great. Now, let us show item 1 from our plan above. We again show this for $x \le m$ and leave the other part as an exercise. The right hand side is:

$$\begin{aligned} RHS &= \frac{1}{2}\left(R_{t-1}(x - 2\phi x) + R_{t-1}(x + 2\phi x)\right) \\ &\le \frac{1}{2} \cdot \left(1 - \frac{\phi^2}{2}\right)^{t-1} \cdot \left(\min(\sqrt{x - 2\phi x}, \sqrt{2m - (x - 2\phi x)}) + \min(\sqrt{x + 2\phi x}, \sqrt{2m - (x + 2\phi x)})\right) + x/2m \\ &\le \frac{1}{2}\left(\sqrt{x - 2\phi x} + \sqrt{x + 2\phi x}\right)\left(1 - \frac{\phi^2}{2}\right)^{t-1} + x/2m \\ &= \frac{1}{2}\left(1 - \phi^2/2\right)^{t-1} \cdot \sqrt{x}\left(\sqrt{1 - 2\phi} + \sqrt{1 + 2\phi}\right) + x/2m \\ &\le \sqrt{x}\left(1 - \phi^2/2\right)^{t} + x/2m \end{aligned}$$

You should verify the last step using Taylor expansions which tells you $1/2(\sqrt{1 - 2\phi} + \sqrt{1 + 2\phi}) \le 1 - \phi^2/2$ which holds for all $\phi \in (0, 1/2)$.



Sublinear Algorithms Hidden Gem 8

Poissonization

Lecturer: Maryam Aliakbarpour. *Scribe:* Ria Stevens

July 2, 2024

1 General Idea

Suppose you're given m fish, assigned randomly to n tanks. You want to study how many fish are in each tank, or determine when every tank will have at least one fish. However, you run into a problem: the number of fish in each tank is binomially distributed and not independent, making your life quite difficult. Luckily for you, if you're willing to get an approximate answer, Poissonization can save your day and your fish. **Poissonization** translates the dependent binomial random variables in your problem into independent Poisson random variables, which may be easier to analyze.

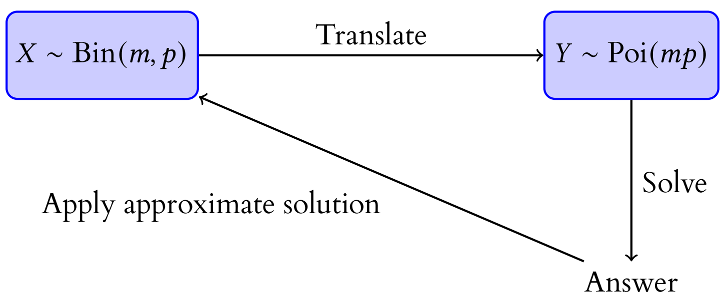


Figure 7: Pictorial Depiction of Poissonization

Why use Poissonization? Poissonization may be used to simplify an analysis by introducing independence. The following properties of the Poisson distribution may also be useful in many settings:

1. If $Y \sim \mathrm{Poi}(\lambda)$, then $\mathbb{E}[Y] = \mathrm{Var}\, Y = \lambda$
2. If $Y_1 \sim \mathrm{Poi}(\lambda_1)$ and $Y_2 \sim \mathrm{Poi}(\lambda_2)$, then $Y_1 + Y_2 \sim \mathrm{Poi}(\lambda_1 + \lambda_2)$

Why does Poissonization work? Recall the distributions involved in Poissonization:

$$\Pr_{X\sim\mathrm{Bin}(m,p)}[X = k] = \binom{m}{k} p^k (1-p)^{m-k}$$
$$\Pr_{Y\sim\mathrm{Poi}(mp)}[Y = k] = \frac{(mp)^k e^{-mp}}{k!}$$

In simple terms, Poissonization works because the Poisson distribution is a good approximation of the binomial distributions when $m \gg k$. We can see this by comparing the probability mass functions of both distributions:

$$\begin{aligned}
\Pr_{X\sim\mathrm{Bin}(m,p)}[X=k] &= \binom{m}{k} p^k\ (1-p)^{m-k} \\
&= \frac{m!}{(m-k)!\,k!}\, p^k\ (1-p)^{m-k} \\
&\approx \frac{m^k}{k!}\, p^k\ (1-p)^{m-k} \\
&\approx \frac{m^k}{k!}\, p^k\ (1-p)^{m} \\
&\approx \frac{m^k}{k!}\, p^k\ (e^{-p})^{m} \\
&= \frac{m^k}{k!}\, p^k\ e^{-mp} \\
&= \Pr_{Y\sim\mathrm{Poi}(mp)}[Y=k]
\end{aligned}$$

The first and second approximations in the above derivation hold when $m \gg k$.

2 Example: Balls and Bins

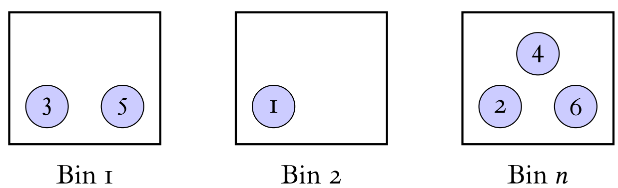


Figure 8: $m = 6$ balls distributed into $n = 3$ bins

Suppose we have m balls, distributed uniformly at random into n bins. We want to study how many balls are in each bin i. To do so, let X_i = # balls in bin i. Then, X_i is a binomial random variable:

$$X_i \sim \mathrm{Bin}\left(m, \frac{1}{n}\right). \tag{10}$$

However, the X_i's are not independent. To introduce independence to our setting, let:

$$Y_i \sim \mathrm{Poi}\left(\frac{m}{n}\right). \tag{11}$$

Now, while the X_i's are not independent, the Y_i's are! Further, as you may have deduced, we can approximate the X_i's using the Y_i's in certain applications.

Particularly, we have the following theorem, which states that the joint distribution of the Y_i's, conditioned on their sum being k, is identical to the joint distribution of the X_i's.

Theorem 1 (Joint Distribution)**.** *Let $X_i \sim \mathrm{Bin}(m, 1/n)$ for all $i \in [n]$. Let $Y_i \sim \mathrm{Poi}(m/n)$ for all $i \in [n]$. Then, the distribution of $(Y_1, \ldots, Y_n)$ conditioned on $\sum_{i=1}^n Y_j = k$ is identical to the distribution of $(X_1, \ldots, X_n)$.*

Proof. Let $k_1, \ldots, k_n \in \{0, \ldots k\}$ such that $\sum_{i=1}^n k_i = k$.

We want to show:

$$\Pr\left[Y_1 = k_1, \ldots, Y_n = k_n \;\middle|\; \sum_{i=1}^{n} Y_i = k\right] = \Pr\left[X_1 = k_1, \ldots, X_n = k_n\right]$$

First, note that the joint distribution of the X_i's is *multinomial*, with probability mass function:

$$\Pr[X_1 = k_1, \ldots, X_n = k_n] = \frac{k!}{k_1! \cdots k_n!} \cdot \prod_{i=1}^{n} \left(\frac{1}{n}\right)^{k_i} = \frac{k!}{\left(\prod_{i=1}^{n} k_i!\right) n^k}$$

Then, consider the joint conditional probability of the Y_i's:

$$\Pr\left[Y_1 = k_1, \ldots, Y_n = k_n \;\middle|\; \sum_{i=1}^{n} Y_i = k\right] = \frac{\Pr\left[Y_1 = k_1, \ldots, Y_n = k_n\right]}{\Pr\left[\sum_{i=1}^{n} Y_i = k\right]} = \frac{\prod_{i=1}^{n} \Pr\left[Y_i = k_i\right]}{\Pr\left[\sum_{i=1}^{n} Y_i = k\right]}$$

The second equality above holds because the Y_i's are independent. Further, as the sum of multiple independent Poisson random variables is also a Poisson random variable, we know $\sum_{i=1}^{n} Y_i \sim \mathrm{Poi}(m)$. We can thus concisely express and simplify the above probabilities using the probability mass function of the Poisson distribution:

$$\begin{aligned}
\Pr\left[Y_1 = k_1, \ldots, Y_n = k_n \;\middle|\; \sum_{i=1}^{n} Y_i = k\right] &= \left(\prod_{i=1}^{n} \frac{e^{-m/n}\,(m/n)^{k_i}}{k_i!}\right) \cdot \left(\frac{k!}{e^{-m} m^k}\right) \\
&= \frac{k!}{e^{-m}\, m^k} \left(\prod_{i=1}^{n} e^{-\frac{m}{n}}\right) \left(\prod_{i=1}^{n} (m/n)^{k_i}\right) \left(\prod_{i=1}^{n} \frac{1}{k_i!}\right) \\
&= \frac{k!}{\prod_{i=1}^{n} k_i!} \cdot \frac{1}{e^{-m}\, m^k} \cdot \left(e^{-\frac{m}{n}}\right)^n \cdot \left(\frac{m}{n}\right)^{\sum_{i=1}^{n} k_i} \\
&= \frac{k!}{\prod_{i=1}^{n} k_i!} \cdot \frac{1}{e^{-m}\, m^k} \cdot e^{-m} \cdot \left(\frac{m}{n}\right)^k \\
&= \frac{k!}{\left(\prod_{i=1}^{n} k_i!\right) n^k} \\
&= \Pr\left[X_1 = k_1, \ldots, X_n = k_n\right].
\end{aligned}$$

■

3 Generating Independence

Now, let's see how Poissonization may be used to generate independence where none previously existed.

Let P be a distribution over $[n]$. Let $p_i = \Pr_{S \sim P}[S = i]$. Assume we draw m samples $S_1, \ldots, S_m \sim P$.

Let $X_i^{(m)}$ be a random variable capturing the number of i's among S_j's. $X_i^{(m)}$ is a binomial random variable:

$$X_i = \sum_{j=1}^{m} \mathbb{1}\left[S_j = i\right] \sim \mathrm{Bin}(m, p_i). \tag{12}$$

However, as in the balls and bins example, the X_i's are *not* indepedent.

To generate independence among the number of samples in each "bin", we may use the following procedure:

1. Draw $\hat{m} \sim \text{Poi}(m)$. $\hat{m}$ is meant to approximate a fixed sample size m.
2. Draw $\hat{m}$ samples, $S_1, \ldots, S_{\hat{m}}$ from P.
3. Define $Y_i = \sum_{j=1}^{\hat{m}} \mathbb{1}\left[S_j = i\right]$.

The Y_i's generated using this procedure *are* independent Poisson random variables with rate mp_i! Particularly, they're independent Poisson random variables, with $Y_i \sim \text{Poi}(mp_i)$.

Theorem 2. *Let Y_i be generated according to the above procedure. Then, $Y_i \sim \text{Poi}(mp_i)$.*

Proof. Let $k \in \mathbb{N}_0$. The probability mass function of Y_i can be expanded using the law of total probability as:

$$\begin{aligned}
\Pr[Y_i = k] &= \sum_{t=0}^{\infty} \Pr[\hat{m} = t] \ \Pr\left[Y_i = k \,\middle|\, \hat{m} = t\right] \\
&= \sum_{t=0}^{k-1} \Pr[\hat{m} = t] \ \Pr[Y_i = k \mid \hat{m} = t] + \sum_{t=k}^{\infty} \Pr[\hat{m} = t] \ \Pr[Y_i = k \mid \hat{m} = t]
\end{aligned}$$

Here, t represents the number of samples that we may draw from P. When $k > t$, we cannot draw k samples of any i, and the first term above must be 0. Further, when $\hat{m}$ is fixed at t, $Y_i \sim \text{Bin}(t, p_i)$, by Theorem 1. We can thus rewrite the above as:

$$\begin{aligned}
\Pr[Y_i = k] &= \sum_{t=k}^{\infty} \frac{e^{-m} \, m^t}{t!} \binom{t}{k} p_i^k \, (1 - p_i)^{t-k} \\
&= \frac{e^{-m} \, p_i^k}{k!} \sum_{t=k}^{\infty} \frac{m^t \, (1 - p_i)^{t-k}}{(t-k)!} \\
&= \frac{e^{-m} \, p_i^k \, m^k}{k!} \sum_{t=k}^{\infty} \frac{m^{t-k} \, (1 - p_i)^{t-k}}{(t-k)!} \\
&= \frac{e^{-m} \, (mp_i)^k}{k!} e^{m(1-p_i)} \\
&= \frac{e^{-mp_i} \, (mp_i)^k}{k!} \\
&= \Pr_{Y \sim \text{Poi}(mp_i)} [Y = k] \, .
\end{aligned}$$

■

4 Application: Functions of Joint Random Variables

Beyond analyzing distributions themselves, Poissonization offers a powerful tool for studying functions of random variables. In particular, using this technique, we can relate the expectation of a non-negative function applied to binomial random variables to expectation of the same function applied to Poisson random variables. The following theorem formalizes this idea.

Theorem 3. *Let $X_i^{(m)} \sim \text{Bin}(m, p_i)$ for all $i \in [n]$. Let $Y_i^{(m)} \sim \text{Poi}(mp_i)$ for all $i \in [n]$. Let $f(x_1, \ldots, x_n)$ be a non-negative function. Then,*

$$\mathbb{E}\left[f(X_1^{(m)}, \ldots, X_n^{(m)})\right] < \sqrt{2\pi m} \ \mathbb{E}\left[f(Y_1^{(m)}, \ldots, Y_n^{(m)})\right]$$

Proof. We begin by applying the law of total expectation to the expectation of $f(Y_1^{(m)}, \ldots, Y_n^{(m)})$, conditioning on all possible realizations of k–the number of "samples" drawn from the binomial distribution to mimic a Poisson distribution.

$$\begin{aligned}\mathbb{E}\left[f(Y_1^{(m)}, \ldots, Y_n^{(m)})\right] &= \sum_{k=0}^{\infty} \mathbb{E}\left[f(Y_1^{(m)}, \ldots, Y_n^{(m)}) \;\middle|\; \sum_{i=1}^{n} Y_i^{(m)} = k\right] \Pr\left[\sum_{i=1}^{n} Y_i^{(m)} = k\right] \\ &\geq \mathbb{E}\left[f(Y_1^{(m)}, \ldots, Y_n^{(m)}) \;\middle|\; \sum_{i=1}^{n} Y_i^{(m)} = m\right] \Pr\left[\sum_{i=1}^{n} Y_i^{(m)} = m\right] \end{aligned} \tag{13}$$

By Theorem 1, conditioned on $\sum_{i=1}^{n} Y_i^{(m)} = m$, each Y_i is distributed identically to X_i. Particularly, this implies:

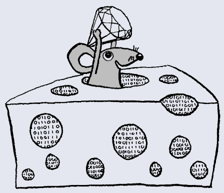
$$\mathbb{E}\left[f(Y_1^{(m)}, \ldots, Y_n^{(m)}) \;\middle|\; \sum_{i=1}^{n} Y_i^{(m)} = m\right] = \mathbb{E}\left[f(X_1^{(m)}, \ldots, X_n^{(m)})\right] \tag{14}$$

Further, because, prior to the conditioning, each $Y_i^{(m)}$ is a Poisson random variable with rate mp_i, their sum is also a Poisson random variable with rate $\sum_{i=1}^{n} mp_i = m$. This gives us a closed-form expression for the probability that this sum equals m:

$$\Pr\left[\sum_{i=1}^{n} Y_i^{(m)} = m\right] = \frac{m^m e^{-m}}{m!} > \frac{m^m e^{-m}}{\sqrt{2\pi m}\left(\frac{m}{e}\right)^m e^{\frac{1}{12m}}} = \frac{1}{\sqrt{2\pi m}\, e^{\frac{1}{12m}}} > \frac{1}{\sqrt{2\pi m}} \tag{15}$$

The first inequality is derived from Stirling's Approximation. Combining Equation 13, Equation 14 and Equation 15, we have the desired result:

$$\mathbb{E}\left[f(Y_1^{(m)}, \ldots, Y_n^{(m)})\right] > \frac{1}{\sqrt{2\pi m}}\, \mathbb{E}\left[f(X_1^{(m)}, \ldots, X_n^{(m)})\right]. \qquad \blacksquare$$



Sublinear Algorithms Hidden Gem 9

The Polarisation Lemma

Lecturer: Ari Biswas. *Scribe:* Ari Biswas

July 11, 2024

1 Problem Statement

Fix $n \in \mathbb{N}$. You are given sample access to two discrete distributions $X^{(0)}$ and $X^{(1)}$ over $\{0,1\}^n$, and promised that either $X^{(0)}$ and $X^{(1)}$ are "kinda" far, or they are "kinda" close in total variation distance; but *not* told which case holds. The goal is to come up with samples from "polarised" distributions[15] $Y^{(0)}$ and $Y^{(1)}$ (over possibly a larger domain), such that if the original distributions were "kinda" close, the polarised distributions are extremely close; and if the original distributions were "a little bit far", then the polarised distributions are very very far. More formally,

Goal

Given $0 \leq \beta < \alpha \leq 1$, where[a] $\alpha^2 > \beta$, come up with an efficient algorithm $\mathsf{Polarise}^{(\alpha,\beta)}$ which takes as input a parameter[b] $k \in \mathbb{N}$ and sample access to distributions $X^{(0)}$ and $X^{(1)}$ over $\{0,1\}^n$. It outputs a pair of samples from a new distribution $Y^{(0)}$ and $Y^{(1)}$ over a domain of size $\{0,1\}^{\mathrm{poly}(n)}$ such that

- If $\mathrm{dist}_{\mathsf{TV}}\left(X^{(0)}, X^{(1)}\right) \geq \alpha \implies \mathrm{dist}_{\mathsf{TV}}\left(Y^{(0)}, Y^{(1)}\right) \geq 1 - 2^{-k}$
- If $\mathrm{dist}_{\mathsf{TV}}\left(X^{(0)}, X^{(1)}\right) \leq \beta \implies \mathrm{dist}_{\mathsf{TV}}\left(Y^{(0)}, Y^{(1)}\right) \leq 2^{-k}$

[a] Setting $\alpha = 2/3$ and $\beta = 1/3$ satisfies this regime

[b] One can think of k being $\mathrm{poly}(n)$, and the algorithm's run time should be polynomial in k.

2 A First Attempt

Given we have sample access to the distributions, the simplest thing one might try is to take k independent samples from each distribution. That is to say, set

- Set $y^{(0)} = (x_1^{(0)}, \dots, x_k^{(0)}) \xleftarrow{\text{i.i.d}} X^{(0)}$ as a single sample $y^{(0)} \xleftarrow{\$} Y^{(0)}$.
- Set $y^{(1)} = (x_1^{(0)}, \dots, x_k^{(0)}) \xleftarrow{\text{i.i.d}} X^{(1)}$ as a single sample $y^{(1)} \xleftarrow{\$} Y^{(1)}$.

Using the direct product lemma (see Lemma 6), we have that if

$$\mathrm{dist}_{\mathsf{TV}}\left(X^{(0)}, X^{(1)}\right) \geq \alpha \implies \mathrm{dist}_{\mathsf{TV}}\left(Y^{(0)}, Y^{(1)}\right) \geq 1 - 2^{-\frac{\alpha^2 k}{2}}$$
$$\mathrm{dist}_{\mathsf{TV}}\left(X^{(0)}, X^{(1)}\right) \leq \beta \implies \mathrm{dist}_{\mathsf{TV}}\left(Y^{(0)}, Y^{(1)}\right) \leq k\beta$$

[15] Sahai and Vadhan [SV03] called the new distributions "amplified distributions". The name was changed in [Vad99] to reflect the fact that the polarisation lemma does not merely increase statistical distance.

So the lower bound on distance between the "little-bit far" distributions is amplified at an exponential rate, but the upper bound on the distance between the "little bit close" is also amplified (albeit, at a slower rate). Thus, we have a process by which we can amplify the distance between far distributions, but no way to shrink the distance between close distributions. Imagine that we had a black box way to shrink the distance between distributions. That is to say we had a shrinking algorithm defined as follows:

There is an efficient algorithm which we will call $\texttt{Shrinker}(X^{(0)}, X^{(1)}, k)$, that receives as input sample access to a pair distributions $X^{(0)}, X^{(1)}$ over $\{0,1\}^n$ and $k \in \mathbb{N}$. It outputs samples from a pair of distributions $(Y^{(0)}, Y^{(1)})$ over $(\{0,1\}^n)^k$ such that

$$\mathrm{dist}_{\mathsf{TV}}\left(Y^{(0)}, Y^{(1)}\right) = \mathrm{dist}_{\mathsf{TV}}\left(X^{(0)}, X^{(1)}\right)^k$$

It turns out that if we did have such a distance shrinking algorithm, we could interweave our amplifier and shrinker to get what we wanted. This leads to the Polarisation lemma.

Theorem 1 (Polarisation Lemma). *Fix $k \in \mathbb{N}$. Let $\alpha, \beta \in [0,1]$ such that $\alpha^2 > \beta$. There is a polynomial time (in k) function* $\mathsf{Polarise}^{(\alpha,\beta)}$*, which receives as input sample access to distributions $X^{(0)}$ and $X^{(1)}$, and an integer k, and which outputs samples from distributions $(Y^{(0)}, Y^{(1)})$ such that*

$$\mathrm{dist}_{\mathsf{TV}}\left(X^{(0)}, X^{(1)}\right) \geq \alpha \implies \mathrm{dist}_{\mathsf{TV}}\left(Y^{(0)}, Y^{(1)}\right) \geq 1 - 2^{-k}$$

$$\mathrm{dist}_{\mathsf{TV}}\left(X^{(0)}, X^{(1)}\right) \leq \beta \implies \mathrm{dist}_{\mathsf{TV}}\left(Y^{(0)}, Y^{(1)}\right) \leq 2^{-k}$$

Proof. Let $\lambda = \min\{\alpha^2/\beta, 2\} > 1$ (This is where we use $\alpha^2 > \beta$), and let $\ell = \lceil \log_\lambda 4k \rceil = O(\log k)$. Apply the $\texttt{Shrinker}$ algorithm to $(X^{(0)}, X^{(1)}, l)$ to get (X_0', X_1') such that

$$\mathrm{dist}_{\mathsf{TV}}\left(X^{(0)}, X^{(1)}\right) \geq \alpha \implies \mathrm{dist}_{\mathsf{TV}}\left(X_0', X_1'\right) \geq \alpha^\ell$$

$$\mathrm{dist}_{\mathsf{TV}}\left(X^{(0)}, X^{(1)}\right) \leq \beta \implies \mathrm{dist}_{\mathsf{TV}}\left(X_0', X_1'\right) \leq \beta^\ell$$

Let $m = \frac{\lambda^\ell}{2\alpha^{2\ell}} \leq \frac{1}{2\beta^\ell}$. (We use $\lambda^\ell \leq \alpha^{2\ell}/\beta^\ell$). Notice $m \leq \mathrm{poly}(k)$, since $\ell = O(\log k)$, and $\lambda \leq 2$, and $\alpha \in [0,1]$ is constant. Applying the direct product lemma (Lemma 6 below) we get $X_0'' = \otimes^m X_0'$ and $X_1'' = \otimes^m X_1'$.

$$\mathrm{dist}_{\mathsf{TV}}\left(X_0, X_1\right) \geq \alpha \implies \mathrm{dist}_{\mathsf{TV}}\left(X_0'', X_1''\right) \geq 1 - 2\exp\left(\frac{\lambda^\ell}{2\alpha^{2\ell}} \cdot \frac{(\alpha^\ell)^2}{2}\right) \geq 1 - 2e^{-k}$$

$$\mathrm{dist}_{\mathsf{TV}}\left(X_0, X_1\right) \leq \beta \implies \mathrm{dist}_{\mathsf{TV}}\left(X_0'', X_1''\right) \leq m\beta^\ell \leq \left(\frac{1}{2\beta^\ell}\right)\beta^\ell = \frac{1}{2}.$$

Apply the $\texttt{Shrinker}(X_0'', X_1'', k)$ again to get sample access to distributions $Y^{(0)}$ and $Y^{(1)}$.

$$\mathrm{dist}_{\mathsf{TV}}\left(X_0, X_1\right) \geq \alpha \implies \mathrm{dist}_{\mathsf{TV}}\left(Y^{(0)}, Y^{(1)}\right) \geq (1 - 2e^{-k})^k \geq 1 - 2ke^{-k} > 1 - 2^k$$

$$\mathrm{dist}_{\mathsf{TV}}\left(X_0, X_1\right) \leq \beta \implies \mathrm{dist}_{\mathsf{TV}}\left(Y^{(0)}, Y^{(1)}\right) \leq \frac{1}{2^k}. \qquad \blacksquare$$

Thus the task of polarising distributions can be reduced to finding an efficient shrinking algorithm.

In order to be able to come up with an efficient shrinking algorithm, we will analyse a seemingly unrelated problem, call the *statistical difference promise problem*. However, the analysis of this unrelated problem will lend insight into how to come up with the shrinking algorithm with properties described above.

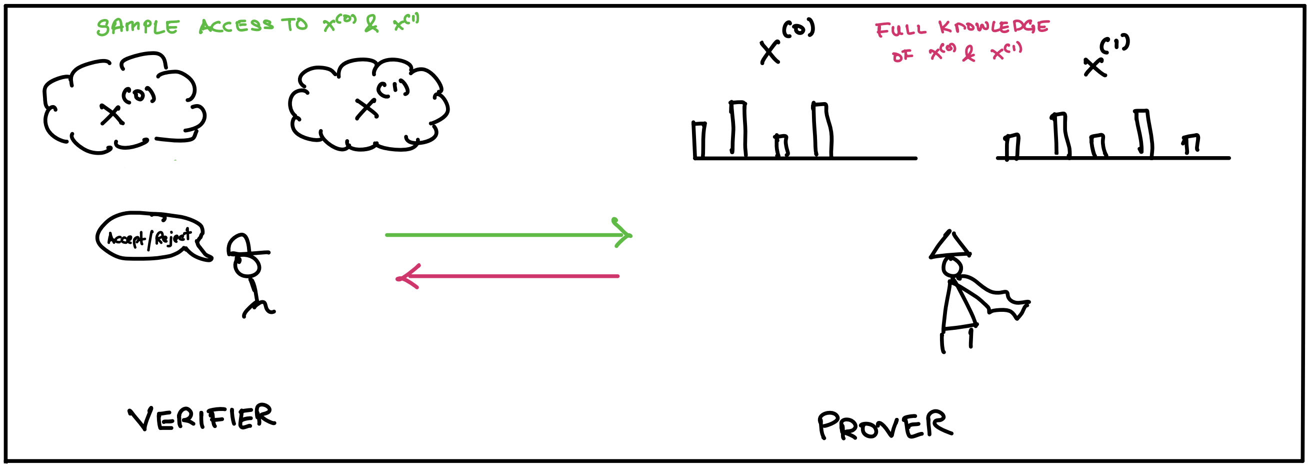


Figure 9: The verifier just like before has sample access to $X^{(0)}$ and $X^{(1)}$. In addition to that, the verifier can talk to an untrusted but computationally unbounded prover that knows the mass functions for $X^{(0)}$ and $X^{(1)}$. The prover and verifier exchange messages, and at the end of the interaction, the verifier outputs accept or reject.

3 The Statistical Difference Promise Problem

Once again, like in the above setting, you (who we refer to as the verifier) are given sample access to two distributions $X^{(0)}$ and $X^{(1)}$ with the same promise about being a little bit far, or a little close. In addition to this sample access, the verifier can communicate with an untrusted and computationally unbounded prover, who has full knowledge of the probability mass distribution of $X^{(0)}$ and $X^{(1)}$ (as described in Figure 9). The goal is to come up with a two party "next message" function[16] Π that specifies the messages the prover and the verifier must send each other. To specify what a prover must send at any given time step, Π takes as input the prover's random coins, its input (the mass function of the two distributions) and all the messages the verifier has sent so far. Similarly, to specify what the verifier must send at a given time step, Π takes as input the verifier's input (samples from the two distributions), the verifier's random coins and all the messages the prover has sent the verifier so far. Thus, the messages exchanged by the prover and the verifier are random, where the randomness is specified by the prover and verifier's private random coins, and the randomness induced by samples from $X^{(0)}$ and $X^{(1)}$.

Let $\Pi(P,V)$ denote the random variable which describes the messages[17] between a prover strategy P and verifier strategy V. Let $\mathsf{out}(\Pi(P,V))$ describe the verifier's final decision after seeing the all the prover's messages. We want Π to satisfy the following conditions:

- **Completeness**: Fix $\delta_c \in [0,1/3]$. If $\text{dist}_{\mathsf{TV}}\left(X^{(0)}, X^{(1)}\right) \geq \alpha$ and the prover follows Π exactly as prescribed, then after interacting with the prover, the verifier accepts with high probability. More formally,

$$\Pr_{\tau \leftarrow\!\!\$\, \Pi(P,V)}\left[V(\tau) = \mathsf{Reject}\right] \leq \delta_c$$

  δ_c is referred to as the completeness error of the protocol.
- **Soundness**: Fix $\delta_s \in [0,1/3]$. If the distributions are close i.e $\text{dist}_{\mathsf{TV}}\left(X^{(0)}, X^{(1)}\right) \leq \beta$, then no matter what strategy is employed by the prover, the verifier rejects with high probability.

$$\Pr_{\tau \leftarrow\!\!\$\, \Pi(P^*,V)}\left[V(\tau) = \mathsf{Accept}\right] \leq \delta_s$$

  where P^* denotes an arbitrary prover strategy (one that may or may not follow the specifications of Π).

[16] See Section 1 of https://people.seas.harvard.edu/~salil/cs221/spring10/lec17.pdf for a definition of a next-message function.

[17] In the literature, these messages are often referred to as the protocol transcript.

3.1 Protocol

Consider the protocol described in Figure 10. The verifier samples a bit b uniformly at random (Line 1), and draws a sample $z \xleftarrow{\$} X^{(b)}$ (Line 2). It then sends over z to the prover and asks the prover to guess which distribution z was sampled from. The prover looks at the mass functions of the two distributions and sets its guess c to the index of the distribution for which z is more likely (Lines 3–6). If the prover's guess was correct, the verifier accepts; otherwise it rejects the prover's claim (Lines 7–8).

	Verifier		**Prover**
1 :	$b \xleftarrow{\$} \{0,1\}$		
2 :	$z \xleftarrow{\$} X^{(b)}$	$\xrightarrow{\text{Send over } z}$	
3 :			Set $c = 0$, if $\Pr[X^{(0)} = z] > \Pr[X^{(1)} = z]$
4 :			Set $c = 1$, if $\Pr[X^{(0)} = z] < \Pr[X^{(1)} = z]$
5 :			Set $c \xleftarrow{\$} \{0,1\}$, if $\Pr[X^{(0)} = z] = \Pr[X^{(1)} = z]$
6 :		$\xleftarrow{\text{Send over } c}$	
7 :	If $c = b$ output Accept		
8 :	If $c \neq b$ output Reject		
9 :			

Figure 10: An interactive proof system for the statistical difference promise problem

Analysing completeness and soundness. When $\alpha = 1$, i.e. $\mathsf{Supp}(X^{(0)}) \cap \mathsf{Supp}(X^{(1)}) = \emptyset$, it is easy to see that $\delta_c = 0$, as the honest prover can perfectly distinguish between samples from either distribution. Similarly, when $\beta = 0$ i.e $X^{(0)} = X^{(1)}$, the prover can do no better than random guessing, so $\delta_s = \frac{1}{2}$. In essence this the nicest set of parameters for the above protocol.

To analyse soundness and completeness for a general $0 \leq \beta < \alpha \leq 1$, where $\alpha^2 > \beta$, it helps to take a different view of the definition of variation distance.

Definition 2 (Total Variation Distance Alternate Definition)**.** Given two distributions $X^{(0)}, X^{(1)}$ over the domain $\{0,1\}^n$, define sets $S^{(0)}$ and $S^{(1)}$ as

$$S^{(0)} \coloneqq \left\{x \in \{0,1\}^n : \Pr[X^{(0)} = x] > \Pr[X^{(1)} = x]\right\}$$
$$S^{(1)} \coloneqq \left\{x \in \{0,1\}^n : \Pr[X^{(1)} = x] > \Pr[X^{(0)} = x]\right\}$$

Then

$$\begin{aligned} \mathrm{dist}_{\mathsf{TV}}\left(X^{(0)}, X^{(1)}\right) &= \Pr_{x \xleftarrow{\$} X^{(0)}}\left[x \in S^{(0)}\right] - \Pr_{x \xleftarrow{\$} X^{(1)}}\left[x \in S^{(0)}\right] \\ &= \Pr_{x \xleftarrow{\$} X^{(1)}}\left[x \in S^{(1)}\right] - \Pr_{x \xleftarrow{\$} X^{(0)}}\left[x \in S^{(1)}\right] \end{aligned}$$

Proof sketch. Let $\delta(S) \coloneqq \Pr\left[X^{(0)} \in S\right] - \Pr\left[X^{(1)} \in S\right]$. We have $\delta = \max_S \delta(S)$. It is clear that for any S we can increase $\delta(S)$ by adding elements from $S^{(0)}$ to S. Similarly, $\delta(S)$ is reduced by removing elements from $S^{(1)}$ from S. $\delta(S)$ is unchanged by adding elements in the domain that have equal mass under $X^{(0)}$ and $X^{(1)}$. Thus, the maximum positive value for $\delta(S)$ can be achieved by setting $S = S^{(0)}$, and the minimum $\delta(S)$ is achieved by

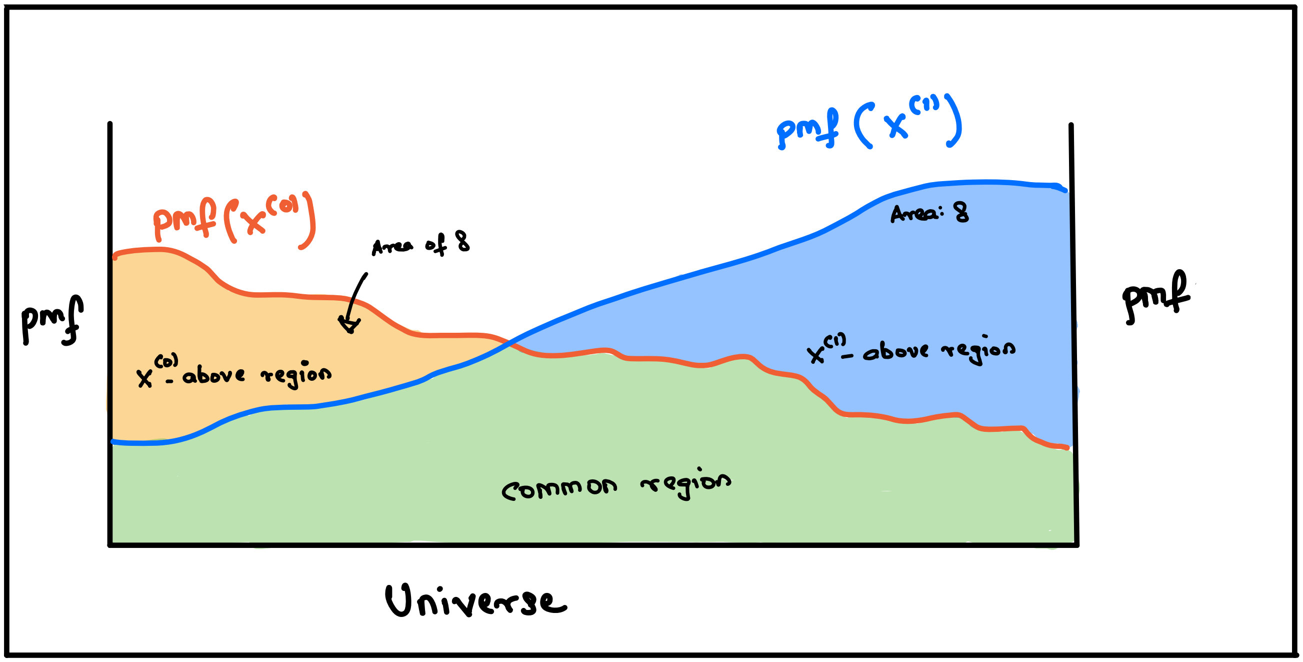


Figure 11: Statistical Difference As Area

setting $S = S^{(1)}$. The maximum positive value and the minimum negative value have the same magnitude as $\delta(\overline{S}) = -\delta(S)$ where $\overline{S}$ is the complement of S. ■

A mental experiment re-imagining the above protocol. Using the above alternate definition for total variation distance, we consider the picture depicted in Figure 11. Keeping this picture in mind, we obtain the following re-imagining of the statistical difference protocol.

The protocol described in Figure 12 is mental experiment only. Clearly, the verifier cannot actually implement this protocol as it does not know δ, and thus would fail at Line 1. However, observe that the distribution on b and z in the mental experiment protocol is identical to the protocol described in Figure 10. (The bit b is uniformly distributed in both protocols. Once b is specified, z is sampled according to $X^{(b)}$ in both protocols). Thus by analysing the soundness and completeness of the protocol in Figure 12, we are analysing the soundness and completeness of the protocol in Figure 10.

Notation: In Protocol 12, when we say sample a point in the common region or the above region, $(z, y) \xleftarrow{\$}$ region, we mean z is the associated x-coordinate and y is the associated y-coordinate of the point in 2-D space in Figure 11.

We have the soundness error δ_s as the probability the verifier V accepts when the two distributions are actually close i.e $\delta \coloneqq \text{dist}_{\text{TV}}\left(X^{(0)}, X^{(1)}\right) \le \beta$.

Verifier		.Prover
$d \leftarrow\!\!\$ \, \texttt{Bernoulli}(1-\delta)$		
If d is `HEADS`		
sample $(z, y) \leftarrow\!\!\$$ common region		
$b \leftarrow\!\!\$ \{0, 1\}$		
If d is `TAILS`		
$b \leftarrow\!\!\$ \{0, 1\}$		
If, $b = 0$, $(z, y) \leftarrow\!\!\$ X^{(0)}$–above region		
Else, sample $(z, y) \leftarrow\!\!\$ X^{(1)}$–above region		
	$\xrightarrow{\text{Send over } z,d}$	
		Set $c = 0$, if $\Pr[X^{(0)} = z] > \Pr[X^{(1)} = z]$
		Set $c = 1$, if $\Pr[X^{(0)} = z] < \Pr[X^{(1)} = z]$
		Set $c \leftarrow\!\!\$ \{0, 1\}$, if $\Pr[X^{(0)} = z] = \Pr[X^{(1)} = z]$
	$\xleftarrow{\text{Send over } c}$	
If $c = b$ output Accept		
If $c \neq b$ output Reject		

Figure 12: A mental experiment that re-describes the protocol described in Figure 10.

$$
\begin{aligned}
\delta_s &= \Pr\left[V \text{ accepts} \;\middle|\; \delta \leq \beta\right] \\
&= \Pr\left[\text{Prover guesses correctly}\right] \\
&= (1-\delta)\frac{1}{2} + \delta \cdot 1 && (16) \\
&= \frac{1+\delta}{2}. && (17)
\end{aligned}
$$

Equation (16) comes from the fact that with probability $1-\delta$, z sampled independent of b, so seeing z gives the prover no information about b. Thus, the best guessing strategy is random guessing. As all $z \in$ `common-region` are equally likely under $X^{(0)}$ and $X^{(1)}$, the prover's strategy specified in Figure 12 is random guessing for when $d =$ `HEADS`. With probability δ, the prover's strategy described in Figure 12 perfectly distinguishes all between if $X^{(0)}$–above region and the $X^{(1)}$–above region. Thus, the soundness and completeness error of the protocol in Figure 12, and therefore that of the protocol in Figure 10 is $\delta_s = \frac{1+\delta}{2}$ and $1 - \delta_s = \frac{1-\delta}{2}$ respectively.

> *At this point one might ask what was the point of analysing this proof system? My eventual goal was to shrink the distance between distributions. That task does not involve any prover. What did all this analysis really give me?*
>
> Observe that the statistical difference δ is directly related to the completeness error and soundness error of the games described above, which is directly related to the prover's ability to guess b correctly. In fact the statistical distance δ is *exactly* equal to the probability that prover can deterministically guess b (without relying on any random guessing). **Thus, one way to shrink the distance between two distributions is to shrink the prover's probability of deterministically guessing b correctly.**

One strategy to reduce the probability of the prover guessing correctly is to repeat the game described in

Figure 10 (or equivalently, Figure 12) k times sequentially and ask the prover to win the game *every* time. That is the verifier samples $(b_1, z_1), \ldots, (b_k, z_k)$ in each round of the game, and sends to the prover $(z_1, \ldots, z_k)$. Here each round is one instantiation of the protocol described in Figure 10. The prover replies with $(c_1, \ldots, c_k)$ and wins only if $c_i = b_i$ for all $i \in [k]$. The probability that this happens is exactly equal to δ^k (as the coin d would have to land `TAILS` every time in Figure 12).

Note that this repeated game is identical to playing the following game: The verifier sends $(z_1, \ldots, z_k)$ to the prover and instead of replying with each b_i, the prover is asked to send back the answer to $\oplus_{i=1}^k b_i$, where $\oplus$ represents the bit wise XOR operation. Note that even if a single z_i was unrelated to b_i (i.e is from the common region), the prover cannot predict that bit b_i without relying on random guessing[18] Therefore, it cannot predict $\oplus_{i=1}^k b_i$ better than random guessing either. Thus the task of guessing $\oplus_{i=1}^k b_i$ without relying on random guessing is equivalent to task of winning every round of the repeated game without relying on random guessing. This probability is exactly equal to δ^k. In other words, two distributions on k-tuples obtained by conditioning exclusive ORs to be 0 and 1 respectively, as in the above protocol has statistical difference δ^k.

Lemma 3 (XOR Lemma)**.** *Given a pair of distributions X_0, X_1 over $\{0,1\}^n$ and $k \in \mathbb{N}$ as input, there exists a polynomial time computable function that outputs a pair of distributions (Y_0, Y_1) such that*

$$\mathrm{dist}_{\mathsf{TV}}(Y_0, Y_1) = \mathrm{dist}_{\mathsf{TV}}(X_0, X_1)^k$$

Specifically, Y_0, Y_1 are defined as follows:

Y_0: Uniformly select $(b_1, \ldots, b_k) \in \{0,1\}^k$ such that $\oplus_{i=1}^k b_i = 0$, and set $Y_0 = (X_{b_1} \otimes \cdots \otimes X_{b_k})$.

Y_1: Uniformly select $(b_1, \ldots, b_k) \in \{0,1\}^k$ such that $\oplus_{i=1}^k b_i = 1$, and set $Y_1 = (X_{b_1} \otimes \cdots \otimes X_{b_k})$.

Proof sketch. The proof comes from the description above. For the prover to distinguish between Y_0 and Y_1 without random guessing, it must be able to guess $\oplus_{i=1}^k b_i$ without relying on random guessing. As the statistical difference between the distributions is exactly equal to the prover's probability of guessing without relying on randomness, we get the final claim. ∎

Given access to input distributions X_0 and X_1, the XOR lemma described above describes a procedure for drawing samples from Y_0 and Y_1 such that $\mathrm{dist}_{\mathsf{TV}}(Y_0, Y_1) = \mathrm{dist}_{\mathsf{TV}}(X_0, X_1)^k$. This is exactly the `Shrinker` algorithm we described earlier. The introduction of the prover really allows us to build intuition as to why mixing samples from both distributions $X^{(0)}$ and $X^{(1)}$ using the XOR property makes distributions close. If one already knew this fact, there is a linear algebraic proof of the above lemma without the use of a prover in [Vad99, Chapter 3, Page 41].

4 Useful Facts

Fact 4. *Given two vectors $\vec{X}$ and $\vec{Y}$, $\vec{X} \otimes \vec{Y}$ represents their tensor product (outer product). Then* $||\vec{X} \otimes \vec{Y}||_1 = ||\vec{X}||_1 \cdot ||\vec{Y}||_1$

Lemma 5 (Repetition Amplifies Distance Between Distributions)**.** *Let $X = (X_0, X_1)$ be a joint distribution where X_0 and X_1 are independent. Similarly, define $Y = (Y_0, Y_1)$. Then we have*

$$\mathrm{dist}_{\mathsf{TV}}(X, Y) \leq \mathrm{dist}_{\mathsf{TV}}(X_0, Y_0) + \mathrm{dist}_{\mathsf{TV}}(X_1, Y_1)$$

[18] Here by not relying on random guessing, we mean that the prover receives $d =$ `TAILS` from the verifier, and their strategy in Figure 12 always leads to the correct guess.

Proof.

$$\begin{aligned}
\mathrm{dist}_{\mathsf{TV}}(X,Y) &= \frac{1}{2}\left|X_0 \otimes X_1 - Y_0 \otimes Y_1\right| && (18)\\
&\le \frac{1}{2}\left|X_0 \otimes X_1 - Y_0 \otimes X_1\right| + \frac{1}{2}\left|Y_0 \otimes X_1 - Y_0 \otimes Y_1\right| && (19)\\
&= \frac{1}{2}\left|X_1 \otimes (X_0 - Y_0)\right| + \frac{1}{2}\left|Y_0 \otimes (X_1 - Y_1)\right| && (20)\\
&= |X_1|\,\frac{1}{2}\left|X_0 - Y_0\right| + |Y_0|\,\frac{1}{2}\left|X_1 - Y_1\right| && (21)\\
&= 1 \cdot \mathrm{dist}_{\mathsf{TV}}(X_0, Y_0) + 1 \cdot \mathrm{dist}_{\mathsf{TV}}(X_1, Y_1) && (22)
\end{aligned}$$

Equation (19) from the triangle inequality and the last equality comes as X_1 and Y_0 are discrete probability distributions. Equation (21) comes from Fact 4. ■

Lemma 6 (Direct Product Lemma)**.** *Let $X^{(0)}, X^{(1)}$ be two discrete distributions over $\{0,1\}^n$ such that $\mathrm{dist}_{\mathsf{TV}}\left(X^{(0)}, X^{(1)}\right) = \delta$. For all $k \in \mathbb{N}$, write $\otimes^k X^{(i)}$ for the distribution of k independent samples drawn from $X^{(i)}$. Then we have*

$$1 - 2 \cdot \exp\left(-\frac{k\delta^2}{2}\right) \le \mathrm{dist}_{\mathsf{TV}}\left(\otimes^k X^{(0)}, \otimes^k X^{(1)}\right) \le k\delta\,.$$

Proof. **Upper Bound:** The upper bound follows directly from Lemma 5 by replacing X_0 and X_1 in the lemma as i.i.d samples from $X^{(0)}$. Likewise, replace Y_0 and Y_1 in the lemma with i.i.d samples from $X^{(1)}$. Then use induction on k.

Lower Bound: Let $\delta \coloneqq \mathrm{dist}_{\mathsf{TV}}\left(X^{(0)}, X^{(1)}\right)$ and let $S^* \subseteq \{0,1\}^n$ where $\Pr[X^{(0)} \in S^*] - \Pr[X^{(1)} \in S^*] = \delta$. Define $p = \Pr[X^{(1)} \in S^*]$. Then $\Pr[X^{(0)} \in S^*] = p + \delta$. Thus, here we have two Bernoulli random variables with mean p and $p+\delta$. By the Chernoff Bound (additive version):

1. Let $Z_i^{(0)} = \mathbb{1}\{x_i \in S^*\}$ where x_i denotes the i'th sample from $X^{(0)}$. The probability that <u>at most</u> $k(p + \frac{\delta}{2})$ fraction of samples $x_1, \ldots, x_k \xleftarrow{\text{i.i.d}} X^{(0)}$ are in set S^* is equivalent to

$$\begin{aligned}
\Pr\left[\sum_{i \in [k]} Z_i^{(0)} \le k\left(p + \frac{\delta}{2}\right)\right] &= \Pr\left[\frac{1}{k}\sum_{i \in [k]} Z_i^{(0)} - \delta \le p - \frac{\delta}{2}\right] && (23)\\
&\le \exp\left(-\frac{k}{2}\delta^2\right) && (24)
\end{aligned}$$

2. Let $Z_i^{(1)} = \mathbb{1}\{x_i \in S^*\}$ where x_i denotes the i'th sample from $X^{(1)}$. The probability that <u>at least</u> $k(p + \frac{\delta}{2})$ fraction of samples $x_1, \ldots, x_k \xleftarrow{\text{i.i.d}} X^{(1)}$ lie in S^* is at most $\exp(-\frac{k}{2}\delta^2)$.

$$\begin{aligned}
\Pr\left[\sum_{i \in [k]} Z_i^{(1)} \ge k\left(p + \frac{\delta}{2}\right)\right] &= \Pr\left[\frac{1}{k}\sum_{i \in [k]} Z_i^{(1)} - p \ge \frac{\delta}{2}\right] && (25)\\
&\le \exp\left(-\frac{k}{2}\delta^2\right) && (26)
\end{aligned}$$

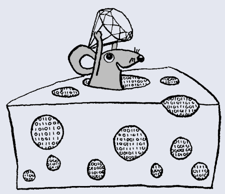
Define $\tilde{S}$ as a set of all k-tuples such that:

$$\tilde{S} = \left\{(z_1, \ldots, z_k) \in (\{0,1\}^n)^k : \text{at least } k(p + \frac{\delta}{2}) \text{ fraction of the components of the } k\text{-tuple lie in } S^*\right\}$$

Thus we have,

$$\begin{aligned} \mathrm{dist}_{\mathsf{TV}}\left((\otimes_{i=1}^k Z_i^{(0)}, \otimes_{i=1}^k Z_i^{(1)})\right) &\geq \Pr[\otimes_{i=1}^k Z_i^{(0)} \in \tilde{S}] - \Pr[\otimes_{i=1}^k Z_i^{(1)} \in \tilde{S}] && (27) \\ &\geq 1 - 2\exp(-\frac{k}{2}\delta^2) && (28) \end{aligned}$$

Equation (27) comes from the definition of statistical difference. Equation (28) comes from items 1 and 2 above. ■



Sublinear Algorithms Hidden Gem 10

Simple Sampling Lower Bounds via Information Theory

Lecturer: Krzysztof Onak. *Scribe:* Kaiwen He

July 14, 2024

1 Goals

The goal of this talk will be to introduce basic Information Theory tools and see how they can be applied to obtain sample lower bounds for the following with high constant probability.

1. Distinguishing a coin with heads–tails bias $(\frac{1}{2}-\epsilon, \frac{1}{2}+\epsilon)$ from $(\frac{1}{2}+\epsilon, \frac{1}{2}-\epsilon)$. We will show that this requires $\Omega(\epsilon^{-2})$ coin tosses.

2. Learning a discrete distribution on $[n] = \{1, 2, \ldots, n\}$ up to total variation distance ϵ. We will show that this will require $\Omega(n/\epsilon^2)$ samples.

2 Information Theory Review

Definition 1 (Entropy)**.** The Entropy of a random variable X on a finite set S s.t. $\Pr[X = i] = p_i$ is defined as

$$H(X) := \sum_{i \in S} p_i \log(1/p_i)\,.$$

Entropy is a measure of how unpredictable a variable is. The entropy gives us the "number of bits needed to specify the output of a variable."

Note: $H(x) \in [0, \log |S|]$, where the the lower bound is achieved by a distribution that outputs a fixed value, and the upper bound is achieved by a uniform distribution over the elements of S.

Definition 2 (Mutual Information)**.** The Mutual information between two random variables X, Y is

$$I(X; Y) := H(X) + H(Y) - H(X, Y)$$

where $H(X, Y)$ is the entropy of the joint variable.

The mutual information measures "how much X and Y have in common". If X and Y are independent then $I(X; Y) = 0$, and if $X = Y$ then $I(X; Y) = H(X) = H(Y)$.

The definitions of entropy and mutual information are well defined when we condition on some random variable set to a specific value (i.e. $H(X|Y = y)$). This allows us to define conditional entropy and conditional mutual information on the variable Z in general:

$$H(X|Z) := \sum_{i \in S} \Pr[Z = i] H(X|Z = i) = H(X) - I(X; Z) \tag{29}$$

$$I(X;Y|Z) := \sum_{i \in S} \Pr[Z=i] I(X;Y|Z=i)$$

Distinguishing between two biased coins or learning a distribution will require the following fact:

Theorem 3. *Let X be a Bernoulli random variable. Then $H(X) \geq 1 - \delta \implies \min\{\Pr[X=0], \Pr[X=1]\} \geq \frac{1}{2} - \sqrt{2\delta}$.*

If the entropy is close to 1, then we can't predict the variable's value with probability much better than $\frac{1}{2}$.

3 Example 1: Distinguishing Two Types of Biased Coins

Question 4. *Suppose we have a coin that is biased, so that HEAD occurs either with probability at least $\frac{1}{2}+\epsilon$, or at most $\frac{1}{2}-\epsilon$. How many times do we need to flip the coin before we can distinguish between these two cases?*

Let $X \in \{-1, 1\}$ be a uniformly distributed random variable. We are unable to observe the value of X directly. We are only able to observe k independent coin toses $T_1, ..., T_k$ such that

$$\Pr[T_i = 0] = \frac{1}{2} - \epsilon X$$
$$\Pr[T_i = 1] = \frac{1}{2} + \epsilon X, .$$

How many times do we need to toss the coin before we can determine X (i.e. the bias of the coin) with high probability? Formally, what is the minimum number $k := k(\epsilon)$ such that there exists a function $f : \{0,1\}^k \to \{\pm 1\}$ which satisfies

$$\Pr_{X, T_1, \ldots, T_k}[f(T_1, \ldots, T_k) = X] \geq 3/4\,?$$

We will answer this question by computing the entropy of X conditional on the observed coin tosses $T_1, T_2, \ldots, T_k$:

Lemma 5. *$H(X|T_1, T_2, ..., T_k) \geq 1 - Ck\epsilon^2$, where C is some universal constant.*

We prove the lemma below, but first let's use it to get the lower bound. Let E denote the event that $T_1, \ldots, T_k$ take values $t_1, \ldots, t_k$ such that $H(X \mid T_1 = t_1, \ldots, T_k = t_k) \geq 1 - 100Ck\epsilon^2$. Using Lemma 5 together with Markov's inequality, we have $\Pr[E] \geq 99/100$. Then by applying Theorem 3, for any function $f : \{0,1\}^k \to \{\pm 1\}$ we get

$$\begin{aligned}\Pr_{X, T_1, \ldots, T_k}[f(T_1, \ldots, T_k) \neq X \mid E] &\geq \min\{\Pr[X=1 \mid E], \Pr[X=-1 \mid E]\} \\ &\geq \frac{1}{2} - \sqrt{200Ck\epsilon^2}\,.\end{aligned}$$

Hence if $k < \frac{1}{25} \cdot \frac{1}{200C\epsilon^2}$, then $f(T_1, \ldots, T_k) = X$ with probability at most

$$\Pr[\neg E] + \Pr[f(T_1, \ldots, T_k) = X \mid E] \leq \frac{1}{100} + \frac{1}{2} + \sqrt{200Ck\epsilon} \leq \frac{51}{100} + \frac{1}{5} = \frac{71}{100}\,.$$

Therefore:

Theorem 6. *The minimum number k of coin tosses T_i required to determine X with probability at least $3/4$ is at least $k = \Omega(\epsilon^{-2})$.*

Proof of Lemma 5.

The proof will proceed by induction on the number of coin tosses. We first show that $I(X;T_1) \leq C\epsilon^2$ for $\epsilon \in (0, \frac{1}{4})$. Define $p \coloneqq 1/2+\epsilon$ and $q \coloneqq 1/2-\epsilon$. Then:

$$\begin{aligned}
H(T_i|X) &= p\log(1/p) + q\log(1/q) \\
&= p\log\frac{2}{1+\epsilon} + q\log\frac{2}{1-\epsilon} \\
&= \frac{1}{2}\log\frac{4}{1-\epsilon^2} + \epsilon\log\frac{1-\epsilon}{1+\epsilon} \\
&= 1 + \frac{1}{2}\log(1+\frac{\epsilon^2}{1-\epsilon^2}) + \epsilon\log\left(1-\frac{2\epsilon}{1+\epsilon}\right) \\
&\geq 1 + \epsilon\log\left(1-\frac{2\epsilon}{1+\epsilon}\right) \\
&\geq 1 + \epsilon C'\left(-\frac{2\epsilon}{1+\epsilon}\right) \quad (C' \text{ constant}) \qquad (30) \\
&\geq 1 - C\epsilon^2 \quad (C \text{ constant}).
\end{aligned}$$

Equation (30) is obtained via Taylor series. We now prove the inductive step and show that $I(X;T_1,...,T_k) \leq kC\epsilon^2$. Assume $I(X;T_1,...,T_i) \leq iC\epsilon^2$. Then

$$\begin{aligned}
I(X;T_1,...,T_{i+1}) &= I(X;T_1,...,T_i) + I(X;T_{i+1}) - I(X;T_1,...,T_i;T_{i+1}) \\
&\leq iC\epsilon^2 + C\epsilon^2 - I(X;T_1,...,T_i;T_{i+1}),
\end{aligned}$$

where $I(X;T_1,...,T_i;T_{i+1})$ is the *interaction information*, a generalization of mutual information. The interaction information has the following identity:

$$I(X;Y;Z) = I(X;Y) - I(X;Y|Z)$$

The issue is that this the interaction information can be negative so we cannot immediately get an upper bound from setting it to be zero.

Example 7. Let Y, Z be uniform on $\{0,1\}$ and define $X \coloneqq Y \oplus Z$ (whose marginal distribution is therefore also uniform on $\{0,1\}$). Then $I(X;Y;Z) = -1$.

Fortunately:

$$I(X;T_1,..,T_i;T_{i+1}) = I(T_1,..,T_i;T_{i+1}) - I(T_1,..,T_i;T_{i+1}|X)$$

and $I(T_1,..,T_i;T_{i+1}|X) = 0$ because, when we know the bias X, the coin tosses are independent. So

$$I(X;T_1,...,T_{i+1}) \leq (i+1)C\epsilon^2 .$$

Put this together to finish the proof of Lemma 5:

$$H(X|T_1,...,T_k) = H(X) - I(X;T_1,...,T_k) \geq 1 - Ck\epsilon^2 .$$

4 Example 2: Learning Distributions on $[n]$

Question 8. *Suppose there is an unknown probability distribution D over $[n]$. How many random samples do we need to draw from D before we can output a distribution D' with $\mathrm{dist}_{\mathsf{TV}}(D,D') < \epsilon$?*

More formally, we say an algorithm A learns distributions over $[n]$ with $m := m(n, \epsilon)$ samples if, for every distribution D over $[n]$, when we take a random multiset $S := \{S_1, \ldots, S_m\}$ with each S_i drawn independently from D, the algorithm $A(S)$ with input S produces a distribution D' such that

$$\Pr_S \left[\mathrm{dist}_{\mathsf{TV}}(D, D') < \epsilon\right] \geq 9/10 .$$

Learning a distribution can be done with $O(n/\epsilon^2)$ samples by outputting the empirical distribution of the collected samples S. The analysis holds via a Chernoff and union bound. We will show that this is optimal. To prove the lower bound, we will define a class of distributions that is hard to learn:

Definition 9 (Hard distributions)**.** For simplicity we let n be even. Define the uniformly distributed random variable

$$X = (X_1, X_2, ..., X_{n/2}) \in \{\pm 1\}^{n/2} . \tag{31}$$

Similar to the coin toss example, we will define probability distributions over $[n]$ that are a function of X. Write D_X for the distribution over $[n]$ where each $j \in [n]$ occurs with probability p_j, defined such that, for each $i \in [n/2]$,

$$p_{2i-1} := \frac{1}{n}(1 - 20X_i\epsilon)$$
$$p_{2i} := \frac{1}{n}(1 + 20X_i\epsilon) .$$

D_X partitions $[n]$ into $n/2$ pairs and the probability of each pair is $2/n$. Each pair behaves like a biased coin, where the direction of bias is determined by an unknown X_i. Our goal is to show that in order to output a distribution D' close to D_X, the algorithm has to learn most of the X_i's. From the biased coin example, we know that learning a single X_i requires $\Omega(1/\epsilon^2)$ coin flips. Each sample from D_X can be thought of as a single coin flip from one of the $n/2$ coins, so, intuitively, we expect that learning $\Omega(n)$ coins should require $\Omega(n/\epsilon^2)$ samples. This is what we will prove:

Theorem 10. *At least $\Omega(n/\epsilon^2)$ samples are required to learn an unknown distribution D_X, of the form defined in Definition 9.*

There are three steps in the proof:

1. Assuming the learning algorithm A outputs a distribution D', construct an approximation X' for X and show that it must be close in Hamming distance.
2. Give an upper bound on the mutual information between X and the sample S.
3. Translate the mutual information bound into a bound on the number of samples required to approximate X.

Step 1: Constructing an approximation X' of X from D'.

Fix any $X \in \{\pm 1\}^{n/2}$. Now let D' be any probability distribution over $[n]$, and write q_i for the probability of i in D'. Define

$$X'_i = \begin{cases} -1 & \text{if } q_{2i-1} \geq q_{2i} \\ 1 & \text{if } q_{2i-1} \geq q_{2i} \end{cases} \tag{32}$$

We will show that if D' is close to D_X in TV distance, then X' is close to X in Hamming distance. Write $\mathrm{Ham}(X, Y)$ for the Hamming distance between two strings $X, Y \in \{\pm 1\}^{n/2}$. Then:

Lemma 11. *For any $X \in \{\pm 1\}^{n/2}$ and any distribution D' over $[n]$,*

$$\mathrm{dist}_{\mathsf{TV}}(D_X, D') \geq \frac{20\epsilon}{n} \mathrm{Ham}(X, X') .$$

Proof. Let p_i, q_i be the probabilities of i in D_X, D' respectively, for each $i \in [n]$. Consider a $j \in [n/2]$ s.t. $X_j \neq X'_j$. Without loss of generality, suppose $X_j = -1$ and $X'_j = 1$. Then $p_{2j-1} = \frac{1}{n}(1 + 20\epsilon) \geq p_{2j} = \frac{1}{n}(1 - 20\epsilon)$, and $q_{2j-1} \leq q_{2j}$ so $q_{2j-1} \leq p_{2j-1}$ and $q_{2j} \geq p_{2j}$. Hence,

$$\begin{aligned} |p_{2j-1} - q_{2j-1}| + |p_{2j} - q_{2j}| &= p_{2j-1} - q_{2j-1} + q_{2j} - p_{2j} \\ &\geq \frac{20\epsilon}{n} + \frac{20\epsilon}{n} = \frac{40\epsilon}{n}. \end{aligned}$$

Then

$$\mathrm{dist}_{\mathsf{TV}}(D, D') = \frac{1}{2} \sum_{j=1}^{n/2} |p_{2j-1} - q_{2j-1}| + |p_{2j} - q_{2j}| \geq \frac{1}{2} \cdot \mathrm{Ham}(X, X') \cdot \frac{40\epsilon}{n} .$$ ■

Corollary 12. *If* $\mathrm{dist}_{\mathsf{TV}}(D, D') \leq \epsilon$, *then* $\mathrm{Ham}(X, X') \leq \frac{n}{20}$

$\mathrm{Ham}(X, X') \leq \frac{n}{20}$ implies that the algorithm has learned at least 9/10 of the coordinates of X. The plan in the next section is to show that with only $o(n/\epsilon^2)$ samples, learning at least 9/10 of the coordinates is unlikely.

Step 2: Mutual information upper bound.

Consider a set of k independent samples $S_1, ..., S_k$ from D. We can think of each sample S_i as a pair $(J_i, T_i) \in [n/2] \times \{0, 1\}$, where $S_i = 2J_i - T_i$. The J_i is distributed independently and uniformly on $[n/2]$. The T_i is distributed with probabilities $\frac{1}{2} - 20\epsilon, \frac{1}{2} + 20\epsilon$ assigned in an unknown order determined by X_i.

Lemma 13. *For $\epsilon \leq 1/80$, $I(X; S_1, .., S_k) \leq Ck\epsilon^2$ for some constant $C > 0$.*

Proof. We will prove the theorem by induction on the number of samples k, similarly to the coin toss. In the base case we have one sample. We will show more generally that for every $i \in [k]$, $I(X; S_i) \leq C\epsilon^2$.

$$\begin{aligned} I(X; S_i) &= H(S_i) - H(S_i|X) \\ &= \log(n) - (H(T_i|J_i; X) + H(J_i; X)) \\ &= \log(n) - (H(T_i|J_i; X) + \log(n/2)) \\ &= 1 - H(T_i|J_i; X) . \end{aligned}$$

$H(T_i|J_i; X)$ reduces exactly to the analysis for a biased coin in the previous section. Hence,

$$I(X; S_i) = 1 - H(T_i|J_i; X) \leq 1 - (1 - C\epsilon^2) = C\epsilon^2 .$$

For the induction step, we want to show $I(X; S_1, .., S_k) \leq Ck\epsilon^2$. The proof follows similarly to the inductive step for the coin toss example. Assume $I(X; S_1, ..., S_i) \leq Ci\epsilon^2$.

$$\begin{aligned} I(X; S_1, ..., S_{i+1}) &= I(X; S_1, ..., S_i) + I(X; S_{i+1} - I(X; S_1, ..., S_i; S_{i+1}) \\ &\leq C(i+1)\epsilon^2 - I(X; S_1, ..., S_i; S_{i+1}) \\ &= C(i+1)\epsilon^2 - (I(S_1, ..., S_i; S_{i+1}) - I(S_1, ..., S_i; S_{i+1}|X)) \\ &= C(i+1)\epsilon^2 - I(S_1, ..., S_i; S_{i+1}) \\ &\leq C(i+1)\epsilon^2 . \end{aligned}$$ ■

Step 3: Sample lower bound.

Using the chain rule for mutual information.

$$I(X;S_1,...,S_k) = \sum_{i=1}^{n/2} I(X_i;S_1,...,S_k|X_1,...,X_{i-1})\,. \tag{33}$$

For each i,

$$\begin{aligned} I(X_i;S_1,...,S_k|X_1,...,X_{i-1}) &= I(X_i;S_1,...,S_k) - I(X_i;S_1,...,S_k;X_1,...,X_{i-1}) \\ &= I(X_i;S_1,...,S_k) - (I(X_i;X_1,...,X_{i-1}) - I(X_i;X_1,...,X_{i-1}|S_1,...,S_k))\,. \end{aligned}$$

$I(X_i;X_1,...,X_{i-1}) = 0$ because each X_i is sampled independently. $I(X_i;X_1,...,X_{i-1}|S_1,...,S_k) = 0$ because each X_i is independent for any set of samples. Hence,

$$I(X;S_1,...,S_k) = \sum_{i=1}^{n/2} I(X_i;S_1,...,S_k) \le Ck\epsilon^2\,. \tag{34}$$

By Markov's inequality, for a random $i \sim [n/2]$,

$$\Pr_{i\sim[n/2]}\left[I(X_i;S_1,..S_k) \ge \frac{20Ck\epsilon^2}{n}\right] \le \frac{1}{10}\,. \tag{35}$$

So, for at least $\frac{n}{2}\cdot\frac{9}{10}$ values of i, we have $I(X_i;S_1,...,S_k) \le \frac{20Ck\epsilon^2}{n}$ and therefore

$$H(X_i|S_1,...,S_k) \ge H(X_i) - I(X_i;S_1,...,S_k) \ge 1 - \frac{20Ck\epsilon^2}{n}\,. \tag{36}$$

If $k \le \frac{n}{16000C\epsilon^2}$, for those i's:

$$H(X_i|S_1,...,S_k) \ge 1 - \frac{1}{8000} \tag{37}$$

By Markov's inequality, with probability $9/10$, over the setting of $S_1,\ldots,S_k$, the conditional entropy of X_i is still at least $1-\frac{1}{800}$. Let E_i be the event that the entropy of X_i conditional on the values of $S_1,\ldots,S_k$, is at least $1-\frac{1}{800}$, so

$$\forall i \in [n/2]\ :\qquad \Pr[E_i] \ge 9/10\,.$$

Recall that, given $S = \{S_1,S_2,\ldots,S_k\}$, our algorithm $A(S)$ outputs a distribution D', from which we define $X' \in \{\pm1\}^{n/2}$ as in Equation (32). Using Theorem 3, the probability that $X'_i = X_i$ conditional on E_i is at most

$$\Pr[X'_i = X_i \mid E_i] \le \frac{1}{2} + \sqrt{2\cdot\frac{1}{800}} = \frac{1}{2} + \frac{1}{20}\,.$$

Therefore

$$\Pr[X'_i = X_i] \le \Pr[\neg E_i] + \Pr[X'_i = X_i \mid E_i] \le \frac{1}{10} + \frac{11}{20} = \frac{13}{20}\,.$$

This gives us a lower bound for the expected Hamming distance between X and X' if $k \le \frac{n}{16000C\epsilon^2}$:

$$\mathbb{E}[\mathrm{Ham}(X,X')] \ge \frac{n}{2}\cdot\frac{9}{10}\cdot\frac{7}{20} = \frac{63}{400}n$$

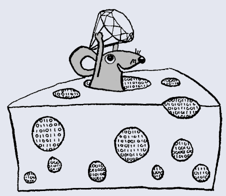
We claim that $\mathrm{Ham}(X,X') > \frac{n}{20}$ with probability greater than $\frac{1}{10}$. Otherwise, we would have

$$\mathbb{E}[\mathrm{Ham}(X,X')] \le \frac{n}{20}\cdot\frac{9}{10} + \frac{n}{2}\cdot\frac{1}{10} = \frac{19}{200}n$$

which is not possible because $19/200 < 63/400$. Therefore, any algorithm that uses at most $\frac{n}{16000\epsilon^2}$ samples, outputs with probability strictly greater than $\frac{1}{10}$ a distribution D' s.t.

$$\mathrm{dist}_{\mathsf{TV}}(D, D') > \frac{20\epsilon}{n} \cdot \frac{n}{20} = \epsilon$$

Hence algorithms that succeed with probability at least $\frac{1}{10}$ must use $> \frac{n}{16000\epsilon^2}$ samples. This completes the proof of Theorem 10.



Sublinear Algorithms Hidden Gem 11

Tightly Relating L^1, and L^∞ Norms of Polynomials

Lecturer: Vipul Arora. *Scribe:* Vipul Arora

July 19, 2024

In this talk, we saw a new way of tightly upper bounding the maximum norm of a polynomial over the solid cube $[-1,1]^n$ in terms of its L^1 norm over the same domain, using classical results from real analysis.

1 Introduction

We begin by considering the class of polynomials of individual degree d with real coefficients:

$$\mathscr{P}_d \triangleq \left\{ p(\boldsymbol{x}) \triangleq p(x_1,\ldots,x_n) = \sum_{\alpha_1=0}^{d} \cdots \sum_{\alpha_n=0}^{d} \underbrace{c_{(\alpha_1,\ldots,\alpha_n)}}_{\triangleq c_{\boldsymbol{\alpha}} \in \mathbb{R}} x_1^{\alpha_1} \ldots x_n^{\alpha_n} : \mathbb{R}^n \to \mathbb{R} \right\}.$$

For any bounded subset $S \subsetneq \mathbb{R}^n$, for any $q \in \mathbb{R}_{\geq 1}$, and any $f : \mathbb{R}^n \to \mathbb{R}$, the respective L^q norms are defined as:

$$\|f\|_{S,q} \triangleq \|f\|_{S,L^q} = \sqrt[q]{\int_S |f(\boldsymbol{x})|^q d\boldsymbol{x}} < \infty.$$

$$\|f\|_{S,\infty} \triangleq \|f\|_{S,L^\infty} = \lim_{q\to\infty} \|f\|_{S,L^q} = \sup_{\boldsymbol{x}\in S}\{|f(\boldsymbol{x})|\}.$$

Trivially, one may observe $\|f\|_{S,q} \leq \|f\|_{S,\infty} \cdot |S|^{1/q}, \forall q \in \mathbb{R}_{\geq 1}$. But for general functions, determining a similar relation in the other direction is hard. For e.g. for the Kronecker delta function

$$\delta_{i,j} \triangleq \begin{cases} 0, \text{ if } i \neq j \\ 1, \text{ if } i = j \end{cases},$$

$\|\delta_{i,j}\|_{[-1,1]^2,\infty} = \|\delta_{i,j}\|_{\mathbb{R},\infty} = 1$, but $\|\delta_{i,j}\|_{[-1,1]^2,q} = \|\delta_{i,j}\|_{\mathbb{R},q} = 0, \forall q \geq 1$. Similarly, for the Dirac Delta function

$$\delta(x) \triangleq \begin{cases} +\infty, \text{ if } x = 0 \\ 0, \text{ if } x \neq 0 \end{cases},$$

$\|\delta\|_{[-1,1],\infty} = \|\delta\|_{\mathbb{R},\infty} = +\infty,$ but $\|\delta\|_{[-1,1],1} = \|\delta\|_{\mathbb{R},1} = 1$. In these cases, such a relation doesn't even exist.

1.1 The results

For $\mathscr{C}_n \triangleq [-1,1]^n \subsetneq \mathbb{R}^n$, we will relate $\|p\|_{\mathscr{C}_n,\infty}$, and $\|p\|_{\mathscr{C}_n,1}$, for any $p \in \mathscr{P}_d$.

Theorem 1 (Upper Bound, [ABB⁺24, Theorem 1.9]). *For all $p \in \mathscr{P}_d$, $\|p\|_{\mathscr{C}_n,\infty} \leq (2d)^{2n}\|p\|_{\mathscr{C}_n,1}$. More generally, for any $1 \leq q \in \mathbb{R}$,*

$$\|p\|^q_{\mathscr{C}_n,\infty} \leq (d^2 2^{q+1})^n \|p\|^q_{\mathscr{C}_n,q}.$$

Theorem 2 (Lower Bound, [ABB+24, Proposition 1.10])**.** *For every odd $d \in \mathbb{N}$, there exists a family $\{(\mathcal{P}_d \ni) f_n : \mathbb{R}_n \to \mathbb{R}\}_{n\in\mathbb{N}}$, such that*

$$\frac{\|f_n\|_{\mathcal{C}_n,\infty}}{\|f_n\|_{\mathcal{C}_n,q}} \geq \max\left\{\left(\frac{d}{2\sqrt{2}}\right)^{2n}, \left(\frac{dq+1}{2}\right)^{n}\right\}.$$

For tight relations, think of $d \to \infty$ as the asymptotic parameter, while $n \geq 1$ is to be treated as a constant. So, proving tightness between L^q, and L^∞ norms is still open.

1.2 Useful tools from the univariate world

We will use the following classical results from real analysis, for proving Theorem 1:

Theorem 3 (Markov Brothers' Inequality, 1890, [Mar90])**.** *Let $g : \mathbb{R} \to \mathbb{R}$ be a polynomial of degree d, and $g' : \mathbb{R} \to \mathbb{R}$ be its derivative. Then,*

$$\|g'\|_{[-1,1],\infty} \leq d^2 \|g\|_{[-1,1],\infty}.$$

Theorem 4 (Lagrange's Mean Value Theorem (Cauchy, 1823), [Rud76, Theorem 5.10])**.** *Let $a < b \in \mathbb{R}$, and $g : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$, and differentiable on (a, b). Then, there exists a point $z \in (a, b)$ such that*

$$g(b) - g(a) = (b - a)g'(z).$$

For proving Theorem 2, we use some useful properties of Legendre polynomials[19]:

Definition 5 (Legendre polynomials)**.** A family of orthogonal polynomials $\{P_m : [-1, 1] \to [-1, 1]\}_{m\in\mathbb{N}}$, indexed by their degree m, such that $P_0(x) = 1$, for all $x \in [-1, 1]$, $P_m(1) = 1$, for all m, and for all $i \neq j$,

$$\int_{-1}^{1} P_i(x)P_j(x)dx = 0. \tag{38}$$

As they form an orthogonal basis for $\mathcal{P}_m$, for any univariate $f \in \mathcal{P}_{<m}$, with $f(x) \equiv \sum_{i=0}^{s<m} c_i P_i(x)$, we get

$$\int_{-1}^{1} f(x)P_m(x)dx = \sum_{i=0}^{s<m} c_i \cancelto{0}{\int_{-1}^{1} P_i(x)P_m(x)dx} = 0.$$

Their derivatives attain their respective extrema at the end points of the interval $[-1, 1]$:

$$\|P'_m\|_{[-1,1],\infty} = |P'_m(\pm 1)| = \frac{m(m+1)}{2}. \tag{39}$$

2 Proving the Upper Bound

We will present a proof sketch of Theorem 1, via induction on n. Using Theorems 3 and 4, we may observe:

Observation 6. *Let $(\mathcal{P}_d \ni) g : [-1, 1] \to \mathbb{R}$, and $x^* \in [-1, 1]$ such that $|g(x^*)| = \|g\|_{[-1,1],\infty}$. Then for every $x \in [-1, 1]$, such that $|x - x^*| \leq 1/(2d^2)$, $|g(x)| \geq |g(x^*)|/2$.*

[19] For more details, see the excellent resource on special functions: [AS64]

Proof. By Theorem 4 there exists a z between x and x^* such that $g(x)-g(x^*) = (x-x^*)g'(z)$, i.e. $|g(x)-g(x^*)| = |x-x^*||g'(z)| \leq |x-x^*| \cdot \|g'\|_{[-1,1],\infty}$. Coupled with Theorem 3, i.e. $\|g'\|_{[-1,1],\infty} \leq d^2\|g\|_{[-1,1],\infty}$, and the fact that $|x-x^*| \leq 1/(2d^2)$, we get

$$|g(x)-g(x^*)| \leq \|g'\|_{[-1,1],\infty}/(2d^2) \leq \|g\|_{[-1,1],\infty}/2 = |g(x^*)|/2.$$

Using triangle inequality, we get $|g(x)| \geq |g(x^*)| - |g(x)-g(x^*)| \geq |g(x^*)|/2$, completing the argument. ■

- Let $\boldsymbol{a} \in \mathbb{R}^n$, consider the axis-parallel line $\mathrm{L}_{\boldsymbol{a},\boldsymbol{e}_i} \triangleq \{\boldsymbol{a} + t\boldsymbol{e}_i : t \in \mathbb{R}\}$, passing through $\boldsymbol{a}$, in the direction $\boldsymbol{e}_i$.
- For any $p \in \mathscr{P}_d$, the restriction of p to this line $p_{\mathrm{L}_{\boldsymbol{a},\boldsymbol{e}_i}}(t) \triangleq p(\boldsymbol{a}+t\boldsymbol{e}_i) : \mathbb{R} \to \mathbb{R}$ is a univariate degree-d polynomial. (Essentially, we are fixing $\boldsymbol{x}_{-i}$ in $p(\boldsymbol{x})$.)
- For every $\boldsymbol{a} \in \mathscr{C}_n$, by Observation 6, there exists a line segment $J \subset \mathrm{L}_{\boldsymbol{a},\boldsymbol{e}_i} \cap \mathscr{C}_n$ of length $|J| \geq 1/(2d^2)$ such that for every $\boldsymbol{y} \in J$,
$$|p(\boldsymbol{y})| = |p_{\mathrm{L}_{\boldsymbol{a},\boldsymbol{e}_i}}(\boldsymbol{y})| \geq |p_{\mathrm{L}_{\boldsymbol{a},\boldsymbol{e}_i}}(\boldsymbol{a})|/2 = |p(\boldsymbol{a})|/2.$$
- Let $\boldsymbol{y}^* \in \mathrm{L}_{\boldsymbol{a},\boldsymbol{e}_i} \cap \mathscr{C}_n$ be such that $|p(\boldsymbol{y}^*)| = \|p_{\mathrm{L}_{\boldsymbol{a},\boldsymbol{e}_i}}\|_{\mathrm{L}_{\boldsymbol{a},\boldsymbol{e}_i}\cap\mathscr{C}_n,\infty}$. Then, for every $\boldsymbol{y} \in J \triangleq \{\boldsymbol{y} \in \mathrm{L}_{\boldsymbol{a},\boldsymbol{e}_i} \cap \mathscr{C}_n : \|\boldsymbol{y}^* - \boldsymbol{y}\|_2 \leq 1/2d^2\}$,
$$|p(\boldsymbol{y})| \geq |p(\boldsymbol{y}^*)|/2 = \|p_{\mathrm{L}_{\boldsymbol{a},\boldsymbol{e}_i}}\|_{\mathrm{L}_{\boldsymbol{a},\boldsymbol{e}_i}\cap\mathscr{C}_n,\infty}/2 \geq |p(\boldsymbol{a})|/2.$$
Note, at least half of $\{\|\boldsymbol{y}^* - \boldsymbol{y}\|_2 \leq 1/2d^2\}$ will be in $\mathscr{C}_n$. So, $|J| \geq 1/(2d^2)$.

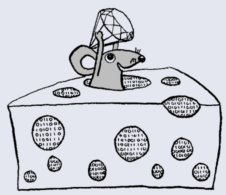
We are now ready to set up the induction.

- Let $\boldsymbol{x}^* \in \mathscr{C}_n$ be such that $|p(\boldsymbol{x}^*)| = \|p\|_{\mathscr{C}_n,\infty}$. We start with $\boldsymbol{a} = \boldsymbol{x}^*$, and the line $\mathrm{L}_{\boldsymbol{x}^*,\boldsymbol{e}_1}$.
- (Base Case) We have proved the existence of the line segment $J_1 \subset \mathrm{L}_{\boldsymbol{x}^*,\boldsymbol{e}_1} \cap \mathscr{C}_n$ of length at least $1/(2d^2)$, such that for every $\boldsymbol{y}_1 \in J_1$, $|p(\boldsymbol{y}_1)| \geq |p(\boldsymbol{x}^*)|/2 = \|p\|_{\mathscr{C}_n,\infty}/2$.
- Now consider the lines $\mathrm{L}_{\boldsymbol{y}_1,\boldsymbol{e}_2}$, for every $\boldsymbol{y}_1 \in J_1$. We get a set J_2 of line segments, of 2-dimensional measure at least $1/(2d^2)^2$, such that for every $\boldsymbol{y}_2 \in J_2$,
$$|p(\boldsymbol{y}_2)| \geq |p(\boldsymbol{y}_1)|/2 \geq \|p\|_{\mathscr{C}_n,\infty}/2^2.$$
- But these line segments are unique, and hence can be translated along $\boldsymbol{e}_2$ to form a 2-dimensional cuboid of 2-dimensional measure at least $1/(2d^2)^2$.
- Repeating this along all $\boldsymbol{e}_i$'s, we get an n-dimensional cuboid $J_n \subset \mathscr{C}_n$ of n-dimensional measure at least $1/(2d^2)^n$, such that for every $\boldsymbol{y}_n \in J_n$, $|p(\boldsymbol{y}_n)| \geq \|p\|_{\mathscr{C}_n,\infty}/2^n$. So, for every $q \geq 1$, we have
$$\|p\|^q_{\mathscr{C}_n,q} = \int_{\mathscr{C}_n} |p(\boldsymbol{y}_n)|^q d\boldsymbol{y}_n \geq \int_{J_n} |p(\boldsymbol{y}_n)|^q d\boldsymbol{y}_n \geq \int_{J_n} \frac{\|p\|^q_{\mathscr{C}_n,\infty}}{2^{nq}} d\boldsymbol{x} = \frac{\|p\|^q_{\mathscr{C}_n,\infty}}{2^{nq}} |J_n| \geq \frac{\|p\|^q_{\mathscr{C}_n,\infty}}{(d^2 2^{q+1})^n}.$$
- With $q = 1$, we get $\|p\|_{\mathscr{C}_n,\infty} \leq (2d)^{2n}\|p\|_{\mathscr{C}_n,1}$.

3 Proof of the Lower Bound

We saw a simpler lower bound instance:

- Consider $(\mathscr{P}_d \ni) h : \mathbb{R}_n \to \mathbb{R}$, defined as $h(\boldsymbol{x}) = \prod_{i=1}^n x_i^d$. Observe $\|h\|_{\mathscr{C}_n,\infty} = 1$, whereas for any $q \geq 1$
$$\|h\|^q_{\mathscr{C}_n,q} = \int_{\mathscr{C}_n} \prod_{i=1}^n |x_i^d|^q d\boldsymbol{x} = \prod_{i=1}^n \int_{-1}^1 |x_i^{dq}| dx_i = \left(2\int_0^1 x^{dq} dx\right)^n = \left(\frac{2}{dq+1}\right)^n \|h\|_{\mathscr{C}_n,\infty}.$$
$$\implies \left(\frac{d+1}{2}\right)^n \|h\|_{\mathscr{C}_n,1} \leq \|h\|_{\mathscr{C}_n,\infty} \leq (2d)^{2n}\|h\|_{\mathscr{C}_n,1}.$$
- In fact there exists a family $\{f_n \in \mathscr{P}_d\}_{n\in\mathbb{N}}$ with a better lower bound, for odd d:
$$\left(\frac{d}{2\sqrt{2}}\right)^{2n} \|f_n\|_{\mathscr{C}_n,1} \leq \|f_n\|_{\mathscr{C}_n,\infty} \leq (2d)^{2n}\|f_n\|_{\mathscr{C}_n,1}.$$
- We skipped this proof. For the complete proofs of Theorems 1 and 2, please refer to [ABB+24, Section 5].



Sublinear Algorithms Hidden Gem 12

Chaining; or, How to Get Better Union Bounds

Lecturer: Samson Zhou. *Scribe:* Shivam Nadimpalli

July 23, 2024

1 Introduction

The following (very broad) question is the motivation for this talk:

Question 1. *Given a collection of random variables $\{Z_t\}_{t\in T}$ indexed by a set $T \subseteq \mathbb{R}^n$, can we say that $\sup_{t\in T} Z_t$ is small with high probability?*

One frequently encounters this situation in life (see Section 2 of [Nel16] for examples across theoretical computer science[20]), and a standard approach is to use a union bound together with appropriate concentration inequalities. This approach can sometimes be too loose, and the topic of this talk—*chaining*—is a technique that allows us to "beat the union bound." Much of this talk is based on [Nel16];[21] alternatively, see [vH16, Ver18, Tal22] for more on chaining, its applications, and related topics.

2 A Concrete Setup

Throughout this note, we will write $B \subseteq \mathbb{R}^n$ for the unit ℓ_2-ball centered at the origin, i.e.

$$B := \{x \in \mathbb{R}^n : \|x\|_2 \le 1\}.$$

Let $T \subseteq \mathbb{R}^n$ be an arbitrary symmetric set (i.e. $t \in T$ whenever $-t \in T$), and consider the collection of jointly Gaussian random variables $\{Z_t\}_{t\in T}$ defined as

$$Z_t := \langle \boldsymbol{g}, t\rangle \qquad \text{where } \boldsymbol{g} \sim N(0, I_n).$$

Here $N(0, I_n)$ denotes the n-dimensional standard Gaussian distribution. Such a collection $\{Z_t\}_{t\in T}$ is sometimes called a *Gaussian process*. We can now ask the motivating question from above in this setting:

Question 2. *What upper bound can one prove on $\mathbb{E}\left[\sup_{t\in T}\langle \boldsymbol{g}, x\rangle\right]$, where $\boldsymbol{g} \sim N(0, I_n)$?*

(As an aside, the quantity $\mathbb{E}[\sup_{t\in T}\langle \boldsymbol{g}, t\rangle]$ is known as the *Gaussian width* of the set T and arises naturally in the context of dimensionality reduction and compressed sensing, cf. Chapter 7 of [Ver18].) Below, we will consider four methods that give increasingly refined answers to this question.

[20] In particular, the streaming example mentioned during the talk is discussed in Section 2.4 of [Nel16].
[21] An accompanying talk is available here.

Example 3 (A running example)**.** A concrete example to keep in mind is the set $T^\star := \{(x, 0, \ldots, 0) : |x| \le 1\}$. It is easy to see that

$$\mathop{\mathbb{E}}_{g\sim N(0,I_n)}\left[\sup_{t\in T^\star}\langle g, t\rangle\right] = \mathop{\mathbb{E}}_{g\sim N(0,I_n)}[|g_1|] = \sqrt{\frac{2}{\pi}}. \tag{40}$$

This will serve as a benchmark for each of the methods to follow.

2.1 Method 1: A union bound

Note that the random variable $\sup_{t\in T}\langle g, t\rangle$ is always non-negative thanks to the symmetry of T, and so

$$\begin{aligned}
\mathop{\mathbb{E}}_{g\sim N(0,I_n)}\left[\sup_{t\in T^\star}\langle g, t\rangle\right] &= \int_{u=0}^{\infty}\Pr_{g\sim N(0,I_n)}\left[\sup_{t\in T}\langle g, t\rangle \ge u\right] du \\
&= \int_{u=0}^{u^*}\Pr_{g\sim N(0,I_n)}\left[\sup_{t\in T}\langle g, t\rangle \ge u\right] du + \int_{u=u^*}^{\infty}\Pr_{g\sim N(0,I_n)}\left[\sup_{t\in T}\langle g, t\rangle \ge u\right] du \\
&\le u^* + \int_{u=u^*}^{\infty}\Pr_{g\sim N(0,I_n)}\left[\sup_{t\in T}\langle g, t\rangle \ge u\right] du \\
&\le u^* + \int_{u=u^*}^{\infty}\sum_{t\in T}\Pr_{g\sim N(0,I_n)}[\langle g, t\rangle \ge u]\, du \qquad (41) \\
&\le u^* + |T|\exp(-(u^*)^2/2) \qquad (42)
\end{aligned}$$

where Equation (41) follows via a union bound, and Equation (42) relies on the fact that $g \cdot t$ is distributed according to $N(0, \|t\|_2^2)$ together with a standard tail bound for univariate Gaussians. Setting $u^* := \sqrt{2\ln|T|}$ then gives

$$\mathop{\mathbb{E}}_{g\sim N(0,I_n)}\left[\sup_{t\in T}\langle g, t\rangle\right] \le O\left(\sqrt{\ln|T|}\right). \tag{43}$$

Note that this bound is meaningless when the set T is infinite, as is the case for our running example $T^* = \{(x, 0, \ldots, 0) : |x| \le 1\}$ (cf. Equation (40)).

2.2 Method 2: A union bound and an ε-net

A natural way to improve the previous method (especially in the context of our running example T^*) is to combine it with an ε-net of the set, defined (in this setting) as follows:

Definition 4 (ε-Net)**.** A set S_ε is an ε-net for the set T if for every $t \in T$ there exists $\pi(t) \in S_\varepsilon$ such that $\|t - \pi(t)\|_2 \le \varepsilon$.

We can thus write

$$\begin{aligned}
\mathop{\mathbb{E}}_{g\sim N(0,I_n)}\left[\sup_{t\in T}\langle g, t\rangle\right] &= \mathop{\mathbb{E}}_{g\sim N(0,I_n)}\left[\sup_{t\in T}\langle g, \pi(t)\rangle + \langle g, t-\pi(t)\rangle\right] \\
&\le \mathop{\mathbb{E}}_{g\sim N(0,I_n)}\left[\sup_{\pi(t)\in S_\varepsilon}\langle g, \pi(t)\rangle\right] + \mathop{\mathbb{E}}_{g\sim N(0,I_n)}\left[\sup_{t\in T}\langle g, t-\pi(t)\rangle\right] \\
&\le O\left(\sqrt{\log|S_\varepsilon|}\right) + \mathop{\mathbb{E}}_{g\sim N(0,I_n)}\left[\sup_{t\in T}\langle g, t-\pi(t)\rangle\right] \qquad (44) \\
&\le O\left(\sqrt{\log|S_\varepsilon|}\right) + O(\varepsilon\sqrt{n}), \qquad (45)
\end{aligned}$$

where Equation (44) follows from Equation (43) and Equation (45) follows via Cauchy–Schwarz and the fact that $\mathbb{E}[\|g\|_2] = O(\sqrt{n})$. Note that when $\varepsilon = 0$, we have $S_\varepsilon = T$ and we recover Equation (43) from Equation (45).

How does Equation (45) fare with respect to the correct bound (Equation (40)) for the set T^*? It is easy to check that it gives us an upper bound of

$$O\left(\sqrt{\log\left(\frac{1}{\varepsilon}\right)} + \varepsilon\sqrt{n}\right),$$

which is $O(\sqrt{\log n})$ for $\varepsilon = 1/\sqrt{n}$. While this beats the union bound, it is still far from the correct answer we know thanks to Equation (40).

One downside of this method in general is that it requires you to have good estimates on the size of an ε-net of your set.

2.3 Method 3: Chaining (Dudley's inequality)

Rather than just using one ε-net, we will instead use a (countably) infinite number of ε-nets, each at a different scale. Let $\{T_k\}_{k=1}^{\infty}$ be a sequence of ε-nets of T such that

$$T_k \text{ is a } (2^{-k})\text{-net of } T \text{ with } T_0 = \{0^n\}.$$

Note that T_0 is a 1-net of T since $T \subseteq B$. Let $\pi_k(t)$ be the point in T_k that is closest to t; note that $\|t - \pi_k(t)\|_2 \leq 2^{-k}$. We can think of $\pi_k(t) \to t$ as $k \to \infty$.

For $t \in T$, we can now write the inner product $\langle g, t\rangle$ as a telescoping sum:

$$\langle g, t\rangle = \langle \pi_0(t), g\rangle + \sum_{k=1}^{\infty} \langle g, \Delta_k(t)\rangle, \quad \text{where} \quad \Delta_k(t) := \pi_k(t) - \pi_{k-1}(t).$$

Recall that $\pi_0(t) = 0^n$, and so

$$\mathop{\mathbb{E}}_{\boldsymbol{g}\sim N(0,I_n)}\left[\sup_{t\in T}\langle \boldsymbol{g}, t\rangle\right] \leq \sum_{k=1}^{\infty} \mathop{\mathbb{E}}_{\boldsymbol{g}\sim N(0,I_n)}\left[\sup_{t\in T}\langle \boldsymbol{g}, \Delta_k(t)\rangle\right]. \tag{46}$$

At first glance, it seems like we are still taking a supremum over the set T, but note that it is actually a supremum over a much smaller set, namely the set of all possible pairs $(\pi_k(t), \pi_{k-1}(t))$. In particular, this set has size at most

$$|T_k| \cdot |T_{k-1}| \leq |T_k|^2.$$

It is also easy to check using the triangle inequality that

$$\begin{aligned}
\|\Delta_k(t)\|_2 &= \|\pi_k(t) - \pi_{k-1}\|_2 \\
&\leq \|\pi_k(t) - t\|_2 + \|\pi_{k-1}(t) - t\|_2 \\
&\leq \frac{3}{2^k}.
\end{aligned}$$

In particular, $\langle \boldsymbol{g}, \Delta_k(t)\rangle$ is distributed according to a univariate Gaussian with standard deviation at most $3/2^k$. These two observations, along with the "union bound" method, readily imply that

$$\mathop{\mathbb{E}}_{\boldsymbol{g}\sim N(0,I_n)}\left[\sup_{t\in T}\langle \boldsymbol{g}, \Delta_k(t)\rangle\right] \leq \frac{O(1)}{2^k}\sqrt{\log|T_k|},$$

and combining this with Equation (46) gives the following bound:

$$\mathop{\mathbb{E}}_{g\sim N(0,I_n)}\left[\sup_{t\in T}\langle g,t\rangle\right] \le O(1)\cdot\sum_{k=1}^{\infty}2^{-k}\sqrt{\log|T_k|}. \tag{47}$$

Equation (47) is known as (the discrete version of) Dudley's inequality (cf. Chapter 8 of [Ver18]).

Dudley's inequality finally allows us to obtain a tight (up to constant factors) bound on $\mathbb{E}[\sup_{t\in T^\star}\langle g,t\rangle]$ for our running example $T^* = \{(x,0,\ldots,0) : |x|\le 1\}$:

$$\begin{aligned}\mathop{\mathbb{E}}_{g\sim N(0,I_n)}\left[\sup_{t\in T^*}\langle g,t\rangle\right] &\le O(1)\cdot\sum_{k=1}^{\infty}2^{-k}\sqrt{\log(2^{k+1})}\\ &= O(1)\cdot\sum_{k=1}^{\infty}2^{-k}\cdot\sqrt{k+1}\\ &= O(1).\end{aligned}$$

2.4 Method 4: Generic chaining

While we have succeeded in obtaining a tight bound on $\mathbb{E}[\sup_{t\in T^\star}\langle g,t\rangle]$, it is worth mentioning a powerful sharpening of the chaining method due to Talagrand which is known as the *generic chaining* [Tal22]. We will give a very brief overview of Talagrand's bound; the interested reader is referred to [vH16, Nel16, Ver18] for accessible introductions.

Recall that the chaining method considered a sequence of nets $\{T_k\}_{k=0}^{\infty}$ at different scales (i.e. for $\varepsilon = 2^{-k}$ for $k=0,1,2,\ldots$) and controlled the size of the net T_k at each scale. The *generic chaining* proceeds via the natural "inverse" problem: it fixes the cardinality of the set T_k at different scales and controls the quality of T_k as a net. Namely, let $\{T_k\}_{k=0}^{\infty}$ be a sequence of subsets of T such that

$$|T_0| = 1, \quad |T_k|\le 2^{2^k}, \quad \text{and} \quad T_0\subseteq T_1\subseteq\ldots\subseteq T.$$

Note that each T_k is a $(\sup_{t\in T}\|t-T_k\|_2)$-net of T (where $\|t-T_k\|_2 := \inf_{s\in T_k}\|t-s\|_2$).

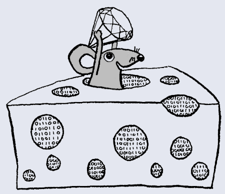
We can now re-express Dudley's inequality (Equation (47)) as

$$\mathop{\mathbb{E}}_{g\sim N(0,I_n)}\left[\sup_{t\in T}\langle g,t\rangle\right] \le O(1)\cdot\sum_{k=1}^{\infty}\sup_{t\in T}\|t-T_k\|_2\cdot 2^{k/2}. \tag{48}$$

Talagrand's improvement of Dudley's inequality is the following:

$$\mathop{\mathbb{E}}_{g\sim N(0,I_n)}\left[\sup_{t\in T}\langle g,t\rangle\right] \le O(1)\cdot\inf_{\{T_k\}\text{ admissible}}\;\sup_{t\in T}\sum_{k=1}^{\infty}\|t-T_k\|_2\cdot 2^{k/2}. \tag{49}$$

While the difference between Equations (48) and (49) may appear minor, it turns out that the R.H.S. of Equation (49) is also a lower bound (up to a multiplicative constant) on $\mathop{\mathbb{E}}_{g\sim N(0,I_n)}[\sup_{t\in T}\langle g,t\rangle]$; this is the content of Talagrand's celebrated *majorizing measures* theorem [Ver18, Tal22].



Sublinear Algorithms Hidden Gem 13

Klein-Plotkin-Rao Decomposition

Lecturer: Shweta Jain. *Scribe:* Sabyasachi Basu

July 25, 2024

1 Introduction

The Klein-Plotkin-Rao decomposition [KPR93] of graphs was a fundamental result in the graph partitioning literature. It was found in the context of studying flow-based problems. The single commodity maxflow-mincut duality is well known and is a cornerstone of theory research. However, no such equality holds in case of multicommodity flow, leading to a long line of research on approximation ratio for mincuts in this situation.

The first result in this context was the celebrated result of Leighton and Rao [LR88], who found an upper bound of $\log n$ on the ratio, and an approximation algorithm that achieved this gap.

Klein, Plotkin and Rao (hereafter referred to as KPR) go a step further and show the following: you can get this ratio down to merely $O(\log k)$ in minor excluded graph families. The key helper theorem they use is (very crudely) of the following flavor: if your graph can be broken into pieces such that they have low diameter in the *original* graph, and cross piece flows cannot have massive value, then that gives you a $O(\log k)$ approximation ratio, where k is just the number of commodities.

The low cross-piece cost is achieved by a decomposition that partitions the graph with very few edges being removed. The challenge lies in showing that such a partition can also achieve low diameter pieces. KPR show that this is possible in minor excluded graph families: they do an iterative BFS on the graph, and show that by doing r successive searches and partitions, whatever you achieve must either have low diameter, or contain a $K_{r,r}$ minor.

2 The Algorithm and Challenges

The algorithm itself is simple, and we can explicitly describe it first.

Algorithm 1 KPR Decomposition(G)

1: $\mathcal{T} :=$ BFS tree of G starting at a random vertex v
2: Partition V into α equivalence classes: the k-th class contains vertices at the kth level (mod α) of $\mathcal{T}$.
3: Remove all edges going from level $i \to i+1$ for some fixed 'good' (TBD) i. Call this set E_i.
4: Let $G' = (V, E \setminus E_i)$ and let $\{G_1\}$ be the connected components of G'.
5: Repeat steps 1 to 4 for each connected component iteratively $r+1$ times. Call the outputs at step j $\{G_j\}$
6: **return** $\{G_r\}$

The main theorem about the decomposition states the following:

Theorem 1. *If G excludes a $K_{r,r}$ minor, then the outputs $\{G_r\}$ have weak diameter $O(r^2\alpha)$.*

It is not hard to see that this decomposition can be done with 'low' cost. The equivalence classes on vertices impose equivalence classes on edges (assign each edge to its higher ranked endpoint), and since these are disjoint classes, at least one of them has at most a $1/\alpha$ fraction of the edges as an immediate consequence of the pigeon-hole principle. Let one such class be our 'good' i in step 3 of the algorithm. Then, iterating r times, we remove at most an r/α fraction of the edges.

It is much less obvious that this will give us the bound on the weak diameter. Consider, for example, a $r-1$ dimensional hypercube. Hypercubes are minor excluded graph families, so the theorem implies that we can partition it into low diameter pieces with just r levels of a BFS. Consider the following exercise, where we walk through the implementation of this algorithm on the hypercube. Starting at some vertex, we construct the BFS tree and then remove edges from the same equivalence class. Given the nice structure of the hypercube, what this does is partition the vertex set into annuli of equidistant vertices from our seed. If our annulus contains the seed vertex, we are done (in fact, it is then actually a ball, and not an annulus) because the vertices all have short paths through the seed vertex. However, suppose we take an annulus of vertices at distance $k \log n$ and $2k \log n$ from our seed vertex. Then, one can verify that any two vertices in this annulus could be as far as $O(n)$ distance away from each other.

What the KPR decomposition implies is that even if you simply repeat this procedure just $r+1$ times, you will achieve a bound on the weak diameters. In fact, that is exactly what we will do in the hypercube annuli: we take a low width annulus along one dimension, but find that the annulus could be arbitrarily large in any of the $(r-1)$ dimensions. We therefore repeat this in each direction, and that is sufficient! It is useful to think of each of the r's as a separate axis even when the graph is not a hypercube.

The formal proof goes via induction. It shows that if your final graph after r iterations is $K_{r,0}$ free, then at the $(r-t)$-th iteration, it will be $K_{r,t}$ free. You repeat this r times, and you get to a $K_{r,r}$.

3 Proof Sketch

We consider only one series of nested subgraphs, because that is sufficient to prove the bound on weak diameters achieved on all pieces. We use a distance parameter δ and construct a series of at most $(r+1)$ subgraphs. We flesh out the details of the algorithm explicitly, and carry out the analysis. Consider a rooted BFS-tree from a random vertex v, $\mathcal{T}_1$. We repeatedly slice this tree till we have satisfactory diameter bounds.

If $\mathcal{T}_1$ has δ levels or fewer, we stop. If not, select an arbitrary set of δ consecutive levels. These may not contain the root, and thus the subgraph induced by these levels may be arbitrarily large. Consider a connected component of this graph, G_2. We construct $\mathcal{T}_2$, and continue this process a total of $r+1$ times to get G_{r+1}.

Assume that the graph has two vertices that are very far apart. The aim of this section will be to show that if the distance is large enough that even after $(r+1)$ striations we are left with large pieces, then that long path itself will induce a $K_{r,r}$ minor in the graph. The following lemma, again a direct consequence of the pigeon-hole principle, states this explicitly:

Lemma 2. *Assume that there exist two nodes in G_{r+1} at distance at least $(r-1)(4(r+1)\delta+1)$ in G_1. Then there is a set of r distinct vertices at pairwise distance of at least $4(r+1)\delta+1)$.*

This is going to be the basis of our induction. Note that these r nodes immediately induce a $K_{r,0}$ minor; these vertices form an independent set (if not, then there is one violating pair of vertices that do not follow the pairwise lower bound of Lemma 2). The induction does the following:

1. Following from some base case, we consider some case where we have a $K_{r,t}$ minor. We will show that one can construct sufficiently many long disjoint paths from these supernodes.
2. We will ensure that the paths are long enough such that we can split the path into three parts: the first part is condensed to the existing supernode; the second part remains a super edge; the third part (from each existing super node) can be condensed to form a new supernode in the nested BFS trees.

3.1 Induction

We code the two parts of the bipartite minor with colors red and blue for ease of reference. We assume that the base case creates r red supernodes initialized at the vertices $v_1, \ldots, v_r$ as shown in Lemma 2. We assume the

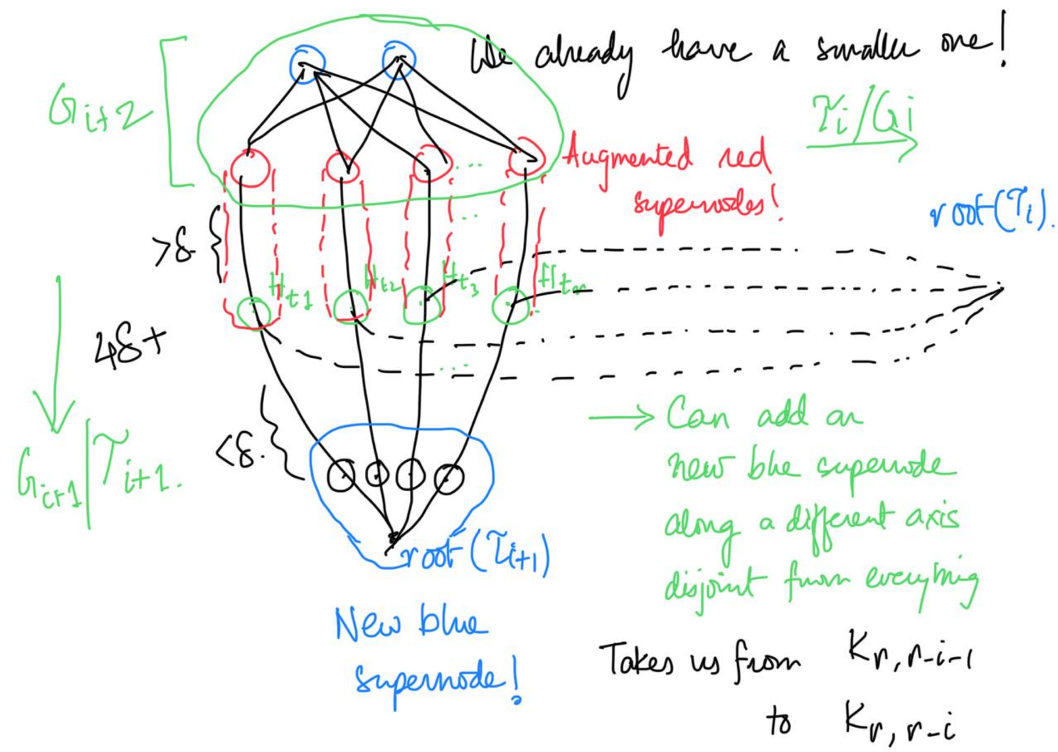


Figure 13: An illustration of the induction step. We use the structure of $\mathcal{T}_i$ to induct on the minor found in $\mathcal{T}_{i+1}$.

following:

1. There is a $K_{r,r-t}$ minor in G_{t+1} such that there are pairwise disjoint paths of length exactly 4δ in G_t leading from each red supernode to the seed vertex in $\mathcal{T}_t$.
2. The set of $(2\delta + 1)$–st node on each such path from the red supernodes are all pairwise at a distance $4t\delta$ from each other in G_1. Call this set H_t.

These disjoint paths form our minor; in the original paper, the authors refer to them as 'tails'. Let the end of each tail form the set $\hat{H}_t$. Assume these conditions hold for a $t = i + 1$; the inductive step will show that it also holds for $t = i$ and forms a $K_{r,r-i}$ from a $K_{r,r-i-1)}$. How do we construct a larger minor from the smaller one? Given access to a $K_{r,r-i-1)}$, we must find a new candidate for a blue supernode such that:

1. It has r disjoint paths going to the r existing red supernodes.
2. These new paths also do not intersect with the existing red–blue paths.

The tails are very good candidates for this! The first condition is perhaps easier to see. You take any red supernode in a graph G_k, and you retrace a path to its seed in the BFS tree $\mathcal{T}_{k-1}$. We call these the *ancestor paths*. For every red supernode, these paths are disjoint. By the first condition of the induction, they are also long. This allows us to find the set H_t, which has many vertices that are pairwise far away. The minor construction goes as follows:

1. For each red supernode, by the induction hypothesis, there is a tail in $\mathcal{T}_{i+1}$ of length 4δ. Pick the middle point: it is at distance at least δ from each red super node since G_{i+1} is at most δ levels in $\mathcal{T}_i$.
2. Any two midpoints are also at pairwise distance at least $4(i + 1)\delta$. By the same logic, the end of the tail is also distance at least δ from the midpoint.

3. Thus, from each midpoint, there is a 'moat' of δ levels to the existing red supernodes, and a connected subgraph of G_{i+1}. We *expand* the red supernodes to swallow the middle node from each corresponding tail, and we contract everything beyond the moat on the other side, to a single new supernode. This is possible because they are connected via the root in $\mathcal{T}_{i+1}$. It is easy to verify that this satisfies the disjointness properties sufficiently by going through the calculations carefully.

4. From each middle node in the tail (which we have just swallowed), we will next launch r new tails in the tree $\mathcal{T}_i$. The same process will go through and you will be able to get the $K_{r,r}$ minors eventually.

Thus, we have found a minor when the pairwise distances are $O(r^2)$. The exact calculations and constants are given in the full paper [KPR93], but the main argument goes through as detailed above.

Bibliography

[ABB⁺24] Vipul Arora, Arnab Bhattacharyya, Mathews Boban, Venkatesan Guruswami, and Esty Kelman. Outlier Robust Multivariate Polynomial Regression, 2024. https://arxiv.org/abs/2403.09465. arXiv:2403.09465.

[AS64] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, Ninth Dover reprint, Tenth GPO print edition, 1964. https://www.cs.bham.ac.uk/~aps/research/projects/as/resources/AandS-a4-v1-2.pdf.

[AS00] Noga Alon and Joel Spencer. *The Probabilistic Method*. John Wiley & Sons, 2000.

[BCDM97] Dany Breslauer, Artur Czumaj, Devdatt P. Dubhashi, and Friedhelm Meyer auf der Heide. Transforming comparison model lower bounds to the parallel-random-access-machine. *Information Processing Letters*, 62(2):103–110, 1997.

[BHPS18] Anna Ben-Hamou, Yuval Peres, and Justin Salez. Weighted sampling without replacement. *Brazilian Journal of Probability and Statistics*, 2018. doi:10.1214/17-BJPS359.

[BK13] Daniel Berend and Aryeh Kontorovich. On the concentration of the missing mass. *Electronic Communications in Probability*, 2013. doi:10.1214/ECP.v18-2359.

[BRY14] Piotr Berman, Sofya Raskhodnikova, and Grigory Yaroslavtsev. L_p-Testing. In *Proceedings of the ACM SIGACT Symposium on Theory of Computing (STOC)*, 2014. doi:10.1145/2591796.2591887.

[CW20] Clément L Canonne and Karl Wimmer. Testing data binnings. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM)*, 2020. doi:10.4230/LIPIcs.APPROX/RANDOM.2020.24.

[DFL⁺22] Constantinos Daskalakis, Maxwell Fishelson, Brendan Lucier, Vasilis Syrgkanis, and Santhoshini Velusamy. Multi-item nontruthful auctions achieve good revenue. *SIAM Journal on Computing*, 2022. doi:10.1137/22M1471742.

[DKL24] Ilias Diakonikolas, Daniel M Kane, and Sihan Liu. Testing closeness of multivariate distributions via ramsey theory. In *Proceedings of the ACM SIGACT Symposium on Theory of Computing (STOC)*, 2024. doi:10.1145/3618260.3649657.

[DKN15] Ilias Diakonikolas, Daniel M Kane, and Vladimir Nikishkin. Optimal algorithms and lower bounds for testing closeness of structured distributions. In *Proceedings of the IEEE Symposium on Foundations of Computer Science (FOCS)*, 2015. doi:10.1109/FOCS.2015.76.

[DP12] Devdatt P. Dubhashi and Alessandro Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, USA, 2012.

[ER50] Paul Erdős and Richard Rado. A combinatorial theorem. *Journal of the London Mathematical Society*, 1(4):249–255, 1950.

[FGHH25] Yuting Fang, Mika Göös, Nathaniel Harms, and Pooya Hatami. Constant-cost communication is not reducible to k-hamming distance. In *Proceedings of the ACM SIGACT Symposium on Theory of Computing (STOC)*, 2025.

[FHHH24] Yuting Fang, Lianna Hambardzumyan, Nathaniel Harms, and Pooya Hatami. No complete problem for constant-cost randomized communication. In *Proceedings of the ACM SIGACT Symposium on Theory of Computing (STOC)*, 2024.

[Fis04] Eldar Fischer. On the strength of comparisons in property testing. *Information and Computation*, 189(1):107–116, 2004. doi:10.1016/j.ic.2003.09.003.

[FRGH+23] Aris Filos-Ratsikas, Yiannis Giannakopoulos, Alexandros Hollender, Philip Lazos, and Diogo Poças. On the complexity of equilibrium computation in first-price auctions. *SIAM Journal on Computing*, 2023. arXiv:https://doi.org/10.1137/21M1435823, doi:10.1137/21M1435823.

[Gas] William Gasarch. Applications of Ramsey theory to computer science. https://www.cs.umd.edu/~gasarch/TOPICS/ramsey/ramsey.html. Accessed: July 2, 2024.

[Gol14] Oded Goldreich. On Multiple Input Problems in Property Testing. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM)*, LIPIcs. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2014. doi:10.4230/LIPIcs.APPROX-RANDOM.2014.704.

[Got20] Anthony Gottlieb. The man who thought too fast. https://www.newyorker.com/magazine/2020/05/04/the-man-who-thought-too-fast, 2020. Accessed Aug 3, 2024.

[HHT14] Jason Hartline, Darrell Hoy, and Sam Taggart. Price of anarchy for auction revenue. In *Proceedings of the ACM Conference on Economics and Computation (EC)*. ACM, 2014. doi:10.1145/2600057.2602878.

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 1963. doi:10.1080/01621459.1963.10500830.

[KPR93] Philip Klein, Serge A. Plotkin, and Satish Rao. Excluded minors, network decomposition, and multicommodity flow. In *Proceedings of the ACM SIGACT Symposium on Theory of Computing (STOC)*. ACM, 1993. doi:10.1145/167088.167261.

[Kum23] Akash Kumar. Spectral graph theory lecture notes, 2023. URL: https://sites.google.com/view/kumarakash/teaching/cs-774-spectral-graph-theory.

[LR88] Tom Leighton and Satish Rao. An approximate max-flow min-cut theorem for uniform multicommodity flow problems with applications to approximation algorithms. In *Proceedings of the IEEE Symposium on Foundations of Computer Science (FOCS)*, 1988. doi:10.1109/SFCS.1988.21958.

[Mar90] Andrey Andreyevich Markov. On a question by D. I. Mendeleev. *Zap. Imp. Akad. Nauk. St. Petersburg*, 62:1–24, 1890. https://history-of-approximation-theory.com/fpapers/markov4.pdf.

[Mis20] Cheryl Misak. The remarkable life of philosopher Frank Ramsey. https://blog.oup.com/2020/02/the-remarkable-life-of-philosopher-frank-ramsey/, 2020. Accessed August 3, 2024.

[MO03] David McAllester and Luis Ortiz. Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research*, 4(Oct):895–911, 2003. doi:10.1162/1532443041424292.

[MR02] Michael Molloy and Bruce Reed. *Graph Coloring and the Probabilistic Method*, volume 23 of *Algorithms and Combinatorics*. Springer, 2002.

[Nel16] Jelani Nelson. Chaining introduction with some computer science applications. *Bulletin of the EATCS*, (120), 2016.

[Ram30] Frank P. Ramsey. On a problem of formal logic. *Proceedings of the London Mathematical Society*, 2(1):264–286, 1930.

[Rud76] Walter Rudin. *Principles of Mathematical Analysis*. McGraw Hill, 1976.

[SS07] Moshe Shaked and J George Shanthikumar. *Stochastic orders*. Springer, 2007. doi:10.1007/978-0-387-34675-5.

[ST13a] Daniel A. Spielman and Shang-Hua Teng. A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning. *SIAM Journal on Computing*, 2013. doi:10.1137/080744888.

[ST13b] Vasilis Syrgkanis and Eva Tardos. Composable and efficient mechanisms. In *Proceedings of the ACM SIGACT Symposium on Theory of Computing (STOC)*. ACM, 2013. doi:10.1145/2488608.2488635.

[Str65] Volker Strassen. The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, 36(2):423–439, 1965. doi:10.1214/aoms/1177700153.

[SV03] Amit Sahai and Salil Vadhan. A complete problem for statistical zero knowledge. *Journal of the ACM (JACM)*, 50(2):196–249, 2003.

[Tal22] Michel Talagrand. *Upper and Lower Bounds for Stochastic Processes: Decomposition Theorems*, volume 60. Springer Nature, 2022. doi:10.1007/978-3-030-82595-9.

[Vad99] Salil Pravin Vadhan. *A Study Of Statistical Zero-Knowledge Proofs*. PhD thesis, Massachusetts Institute Of Technology, 1999.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018. doi:10.1017/9781108231596.

[vH16] Ramon van Handel. *Probability in high dimension*. Lecture Notes (Princeton University), 2016. URL: https://web.math.princeton.edu/~rvan/APC550.pdf.

[Waj17] David Wajc. Negative association: definition, properties, and applications, 2017. Manuscript, available from https://goo.gl/j2ekqM.

[Yeo] Dominic Yeo. Hoeffding's inequality and convex ordering. https://eventuallyalmosteverywhere.wordpress.com/tag/sampling-without-replacement/. Accessed: 2024-06-12.