ECS 189A Sublinear Algorithms for Big Data **Fall 2024**

Lecture 2: Concentration Inequalities

Lecturer: Jasper Lee *Scribe: Alan Buckser, David Gee*

1 Lecture Schedule

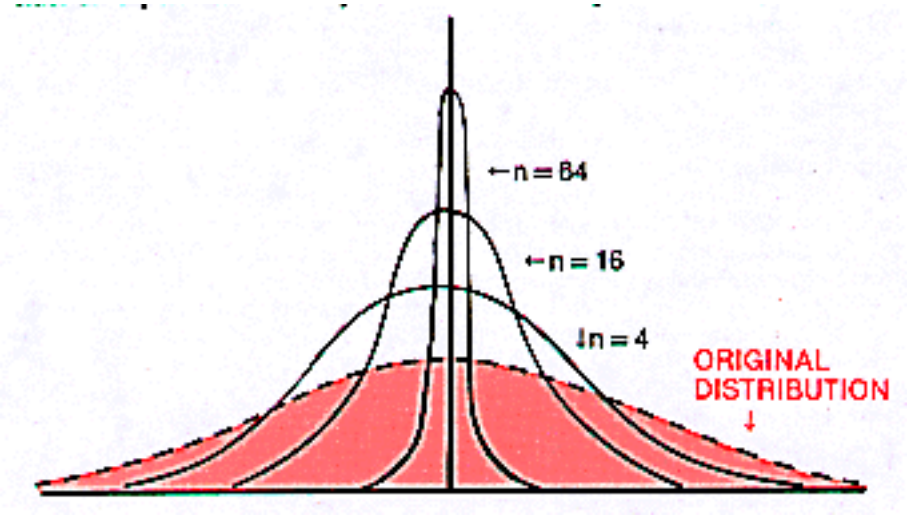
Lectures 2-3 are concentration inequalities.
Lectures 4-8 are doing specific problems.
Lectures 9-13 are property testing on discrete distributions (basically statistics).
The rest of course is streaming algorithms, where the sublinear resource is memory while reading data, not time.

2 Intuition

Over the course of this week, we will utilize randomness, see how to analyze and it, and hopefully by the end of the week see that randomness is well-behaved. "Aggregate behavior of lots of independent processes is typically well-behaved" (at least to some extent).

Consider $(X_1, \ldots, X_n) \leftarrow D \in \mathbb{R}$ independent random variables (r.v.) drawn from a random distribution D with finite mean μ and variance σ^2. Consider the sample mean $\bar{X}_n = \frac{1}{n}\sum X_i$, which aggregates the behavior of the X_i. There are two well-known theorems that describe this aggregate behavior.

- Law of Large Numbers (LLN): $\bar{X}_n \to \mu$ a.s. (almost surely)
- Central Limit Theorem (CLT): $\bar{X}_n \to \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$



(Lesson 1: Summary Measures of Data 2.1 - 9 - UT Health)

The diagram above shows how $\frac{1}{n}$ factor in the variance of the Gaussian approximation becomes spikier as n increases. We want to quantitatively describe how much this spikiness

exists for the actual $\bar{X}_n$, that is to describe how concentrated the probability mass of $\bar{X}_n$ is around its mean. To do this, we want to prove some concentration inequality of the form

$$\mathbb{P}(|\bar{X}_n - \mu| > \epsilon) \leq \text{small}(\epsilon, n).$$

We know a few such bounds.

Proposition 2.1 (Chebyshev's inequality)**.** *Consider a real valued random variable Y with finite mean μ and standard deviation σ.*

$$\mathbb{P}(|Y - \mu| \geq a) \leq \frac{\sigma^2}{a}$$

We can prove Chebyshev's inequality through Markov's inequality.

Proposition 2.2 (Markov's inequality)**.** *For a non-negative random variable X,*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

To prove Chebyshev's inequality with Markov's inequality, we take $X = |Y - \mu|^2$.

Proposition 2.3 (Applying Chebyshev's inequality to $\bar{X}_n$)**.** *Let $\bar{X}_n$ be the sample mean of n independent identically drawn (iid) random variables X_i, each with mean μ and variation σ^2. Then by Chebyshev's inequality,*

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon}.$$

To gauge the effectiveness of these inequalities, we compare them to the Gaussian approximation from the CLT. The Gaussian probability density function (PDF) of $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ is $\left(\sqrt{\frac{n}{2\pi\sigma^2}}\right) e^{-\frac{n\epsilon^2}{2\sigma^2}}$. Then the cumulative density function (CDF) is $\Theta(e^{-\frac{n\epsilon^2}{2\sigma^2}}) = e^{-\Theta n}$. Comparing this to the bound of Chebyshev's (which was $\Theta(1/n)$), we can see the concentration bound of Chebyshev's is very weak compared to the sample complexity of the CDF (linear decay vs exponential decay). We then want to find a stronger probability bound that matches the CDF's decay rate.

To do this, we first consider other moments.

Proposition 2.4 (Beefing up Chebyshev's for higher moments)**.** *For every $k \in \mathbb{N}$,*

$$\mathbb{P}(|Y - \mu| \geq a) \leq \frac{\mathbb{E}(|Y - \mu|^k)}{a^k}.$$

The proof of this is in homework 0. (Note that it had $|Y - \mu|^k \geq a^k$ in the probability, but $|Y - \mu| \geq a$ is equivalent to this.)

Corollary 2.5. $\mathbb{P}(|Y - \mu| \geq a) \leq \inf_{k \in \mathbb{N}} \left(\frac{\mathbb{E}(|Y-\mu|^k)}{a^k}\right)$

In theory, this is a tight upper bound that yields good results. However, we do not know how to choose the correct k to minimize the right hand side. To deal with this, we will look into Chernoff Bounds, which we will be using for the rest of the quarter.

3 Chernoff Bound for Bernoulli Random Variables

We start with the observation that $e^{t\cdot x} = 1 + tx + \frac{t^2x^2}{2!} + \frac{t^3x^3}{3!} + \dots$. Then by the linearity of expectation, applying the moment generating function of Y at t gives us $M_Y(t) = \mathbb{E}_Y(e^{tY}) = 1 + t\mathbb{E}(Y) + \frac{t^2\mathbb{E}(Y^2)}{2!} + \frac{t^3\mathbb{E}(Y^3)}{3!} + \dots$.

We will use $M_Y(t)$ instead of Chebyshev's inequality to prove the desired concentration bound.

Lemma 2.6. *For any r.v. $Y \in \mathbb{R}$ (with finite moment generating function (mgf)), $t > 0$,*

$$\begin{aligned}
\mathbb{P}(Y \geq b) &= P(e^{tY} \geq e^{tb}) \\
&\leq \frac{\mathbb{E}(e^{tY})}{e^{tb}} \, (\textit{by Markov}) \\
&\leq \inf_{t>0} \frac{\mathbb{E}(e^{tY})}{e^{tb}} \\
\mathbb{P}(Y \leq b) &= \mathbb{P}(e^{tY} \leq e^{tb}) \\
&\leq \frac{E(e^{tY})}{e^{tb}} \\
&\leq \inf_{t>0} \frac{\mathbb{E}(e^{tY})}{e^{tb}}
\end{aligned}$$

To apply this to $\bar{X}_n$, we let $Y = \bar{X}_n = \frac{1}{n}\sum X_i$ and $b = \mu + \epsilon$. There are then two more steps we must take to use this:

- We need to upper bound the MGF $\mathbb{E}(e^{tY}) = (\mathbb{E}(e^{t\frac{X}{n}}))^n$. (The proof of this equality is in homework 0.)
- We then need to compute this infimum (which requires finding which t minimizes this).

Theorem 2.7 (Chernoff bound for Poisson Trials[1])**.** *Let $X_i \leftarrow \text{Ber}(p_i)$ be independent random variables. Then $X = \sum X_i$ is the sample sum, which describes aggregate behavior, and $\mu = \sum p_i = \mathbb{E}X$. Note that because X is the sum not the average, μ contains the n from earlier.*

1. $\mathbb{P}(X \geq (1+\delta)\mu) \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu$
2. *For $\delta \in (0,1)$, $\mathbb{P}(X \geq (1+\delta)\mu) \leq e^{-\mu\delta^2/3}$.*

Note that 2 is a simpler but weaker version of 1.

Proof. Step 1: Upper bound the Moment Generating Function (mgf)

[1]Reminder that Poisson Trials are a series of trials where the probability of success varies from trial to trial.

$$\begin{aligned}\mathbb{E}[e^{tX_i}] &= p_i e^t + (1 - p_i) \\ &= 1 + p_i(e^t - 1) \\ &\leq e^{p_i(e^t-1)} \text{ (when } \sum p_i = \mu) \\ &\leq \prod_i e^{p_i(e^t-1)} = e^{\mu(e^t-1)}\end{aligned}$$

Step 2: Apply the Chernoff bound and compute infimum:

$$\begin{aligned}P(X \geq (1+\delta)\mu) &\leq \inf_{t>0} \frac{\mathbb{E}(e^{tX})}{e^{t(1+\delta)\mu}} \text{ (apply } \textit{Lemma 2.6}) \\ &\leq \inf_{t>0} \frac{e^{(e^t-1)\mu}}{e^{t(1+\delta)\mu}} \text{ apply step 1} \\ &= \left(\inf_{t>0} e^{(e^t-1)-t(1+\delta)}\right)^{\mu}\end{aligned}$$

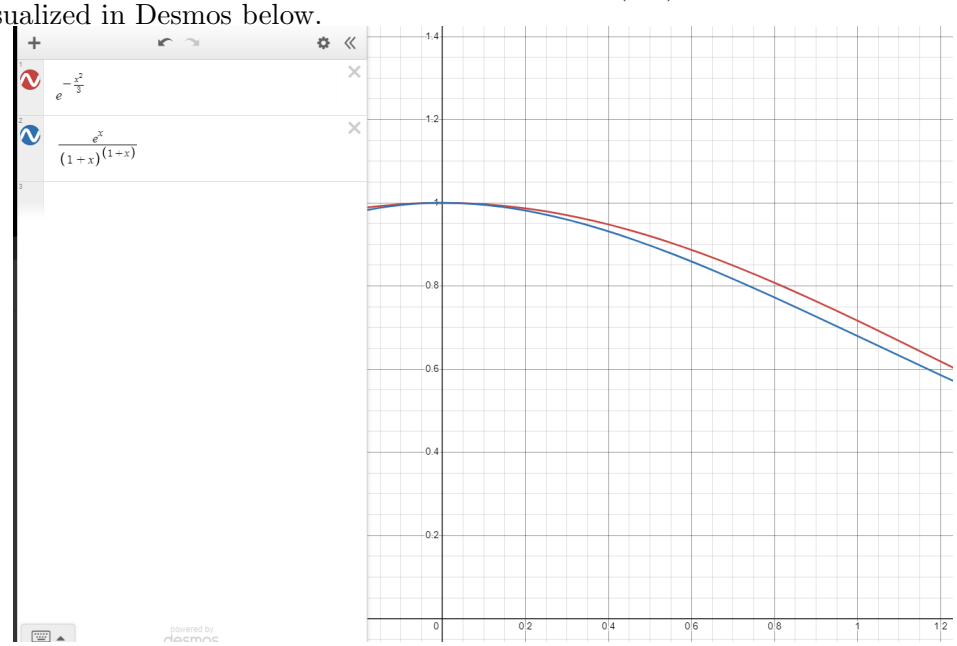
We minimize $e^t - t(1+\delta)$ by setting its derivative equal to 0 and solving for t: $\frac{d}{dt}(e^t - (1+\delta)t) = e^t - (1+\delta) = 0$, so $t = \ln(1+\delta)$.

Then

$$\begin{aligned}P(X \geq (1+\delta)\mu) &\leq \left(\inf_{t>0} e^{(e^t-1)-t(1+\delta)}\right)^{\mu} \\ &\leq \frac{e^{\delta}}{(1+\delta)^{1+\delta}}.\end{aligned}$$

This completes the proof of part 1.

For part 2, observe (via Wolfram Alpha) that $\frac{e^{\delta}}{(1+\delta)^{1+\delta}} \leq e^{-\delta^2/3}$ for $\delta \in (0, 1]$. This is visualized in Desmos below.



□

Theorem 2.8 (Theorem 2.7 but lower bound (L.B.))**.** *For any* $\kappa \in (0,1)$*,*

1. $\mathbb{P}(X \leq (1-\kappa)\mu) \leq \frac{e^{-\kappa}}{(1-\kappa)^{1-\kappa}}$
2. $\mathbb{P}(X \leq (1-\kappa)\mu) \leq e^{-\mu\kappa^2/2}$.

Note the 2 *in the denominator of the exponent, not* 3 *as in Theorem 2.7.*

Combining the previous two theorems, we get a 2-sided tail bound.

Theorem 2.9 (2-sided Chernoff bound for Poisson Trials)**.** *Consider the same setup as Theorem 2.7 and Theorem 2.8, holding* $\kappa \in (0,1)$*. Then* $\mathbb{P}(|X-\mu| \geq \kappa\mu) \leq 2e^{-\mu\kappa^2/3}$

What if we don't know μ, but only know $\mu^+ \geq \mu$? We can still derive some tail bounds.

Corollary 2.10. *For* $n \geq \mu^+ \geq \mu, \kappa \in (0,1)$*, we have* $\mathbb{P}(X \geq (1+\kappa)\mu^+) \leq e^{-\mu^+\kappa^2/3}$.

Proof. Let $X = \sum X_i, X_i \sim \text{Ber}(p_i)$.

Consider $Y_i \leftarrow \text{Ber}(q_i)$ s.t. $q_i \geq p_i$ and $\sum q_i = \mu^+$, and let $Y = \sum Y_i$. This technique is called a coupling argument.

Then

$$\begin{aligned}\mathbb{P}(X \geq (1+\kappa)\mu^+) &\leq \mathbb{P}(Y \geq (1+\kappa)\mu^+) \\ &\leq e^{-\mu^+\kappa^2/3} \text{ (Theorem 2.7)}\end{aligned}$$

□

4 Applications

We can use this Chernoff bound to improve the success probability of some algorithms with constant success probability by running those algorithms repeatedly and then somehow combining the results of the runs into some consensus result.

4.1 Application 1

Consider a decision algorithm $\mathcal{A}$ (outputs 0 or 1) that succeeds with probability $\geq \frac{2}{3}$. We want to construct another algorithm $\mathcal{A}'$ that succeeds with probability $1-\delta$.

Algorithm 1 $\mathcal{A}'$ is an algorithm that boosts the capability of a decision algorithm $\mathcal{A}$

1. Run $\mathcal{A}$ n times.
2. Take the majority vote.

4.2 Analysis of Algorithm 1 (figure out n)

Let E_i be an indicator for whether the ith run got the wrong answer.

$$\begin{aligned}\mathbb{E}(E_i) &= \mathbb{P}(\mathcal{A} \text{ wrong}) \leq \frac{1}{3} \\ \mathbb{P}(\mathcal{A}' \text{ fails}) &= \mathbb{P}\left(\sum E_i \geq \frac{n}{2}\right) \\ &\leq e^{-n/3\left(\frac{1}{2}\right)^2/3} \text{ (Corollary 2.10)} \\ &= e^{-\Theta(n)}\end{aligned}$$

We want the probability $\mathcal{A}'$ fails to be $\leq \delta$.

$$\begin{aligned}e^{+\Theta(n)} &\geq \frac{1}{\delta} \\ n &\geq \Theta\left(\log \frac{1}{\delta}\right)\end{aligned}$$

4.3 Application 2

Consider an algorithm $\mathcal{B}$ that outputs a real number in $[\star - \epsilon, \star + \epsilon]$ (with the goal to output $\star$) with probability $\geq \frac{2}{3}$. We want to construct another algorithm $\mathcal{B}'$ that succeeds at outputting a real number in $[\star - \epsilon, \star + \epsilon]$ with probability $1 - \delta$.

Algorithm 2 $\mathcal{B}'$ is an algorithm that boosts the capability of an estimation algorithm $\mathcal{B}$

1. Run $\mathcal{B}$ n times.
2. Take the median.

We take the median instead of the mean because there are no guarantees about outliers. Note this is a generalization of Algorithm 1 because the median of 0 or 1 is just majority vote.

4.4 Analysis of Algorithm 2 (figure out n)

Let E_i be an indicator for whether the ith run got the wrong answer (outputs less than $\star - \epsilon$).

$$\begin{aligned}\mathbb{E}(E_i) &= \mathbb{P}(\mathcal{B} \text{ wrong}) \leq \frac{1}{3} \\ \mathbb{P}(\mathcal{B}' \text{ fails}) &= \mathbb{P}\left(\sum E_i \geq \frac{n}{2}\right) \\ &\leq e^{-n/3\left(\frac{1}{2}\right)^2/3} \text{ (Corollary 2.10)} \\ &= e^{-\Theta(n)}\end{aligned}$$

We want the probability $\mathcal{B}'$ fails to be $\leq \delta$.

$$e^{+\Theta(n)} \geq \frac{1}{\delta}$$
$$n \geq \Theta\left(\log \frac{1}{\delta}\right)$$

Note that this complexity is totally ϵ agnostic.

5 A General Heuristic

An algorithm with query complexity $q(\delta)$ should aim for $q(\delta) \leq O\left(q\left(\frac{1}{3}\right) \cdot \log \frac{1}{\delta}\right)$. Note the $\log \frac{1}{\delta}$ blow up isn't always tight.