CSCI 1951-W Sublinear Algorithms for Big Data **Fall 2020**

Lecture 3: Concentration Inequalities and Mean Estimation

Lecturer: Jasper Lee *Scribe: Ross Briden*

1 Overview

In this lecture, we'll quickly recap how an optimization algorithm that fails with constant probability can be adapted into a high probability one. Then, we'll introduce two concentration inequalities – **Hoeffding's inequality** and **Bernstein's inequality** – for analyzing the sample complexity of the sample mean for estimating the mean of an underlying distribution. We also state (without proof) a sample complexity lower bound that demonstrates the sample mean of independent random variables, in non-asymptotic regimes, is a poor estimator for their true mean, and introduce a more robust estimator known as the **Median of Means algorithm**.

2 Recap from Last Class

Suppose we have an optimization problem specified by an objective function $f : S \to \mathbb{R}$ subject to a set of constraints $\{C_1, \ldots, C_k\}$, where S is any set and $C_i \subseteq S$ for all constraints.

Definition 3.1 We say that $y \in S$ is a feasible solution to our optimization problem if $y \in \bigcap_{i=1}^{k} C_i$.

The objective of the optimization problem is to find a feasible solution x that either maximize or minimize $f(x)$. WLOG suppose we want to maximize f.

Now, suppose there exists an algorithm D that, with probability $\geq \frac{2}{3}$, returns an answer z so that $f(z) \geq \text{OPT} - \epsilon$ with a *promise* that $z \in \bigcap_{i=1}^{k} C_i$, where OPT is the global maximum of f. Notice that the promise ensures that the algorithm returns a value that is feasible if it succeeds.

Now, we can design a new algorithm D' that uses D as a subroutine to solve the optimization problem with high probability. In particular, given $\delta \in (0, 1)$, D' should run D for $n = \Theta(\log(\frac{1}{\delta}))$ iterations and return the largest feasible value returned by D. To see why D' is correct, first notice that the promise ensures that every time D succeeds, the solution returned by D is feasible. Therefore, the maximum of feasible solutions is also feasible and will also be greater than $\text{OPT} - \epsilon$ by definition. Hence, D' returns a valid solution if after n runs of D, at least one valid solution is returned by D. It then follows that the probability D' fails is $(1 - \frac{2}{3})^n = \frac{1}{3^n}$. Therefore,

$$\frac{1}{3^n} \leq \delta \iff -n\log(3) \leq \log(\delta) \iff n \geq \log\left(\frac{1}{\delta}\right) / \log(3)$$

So, if D' runs D for $\Theta(\frac{1}{\delta})$ iterations, D' will return a correct solution with probability at least $1 - \delta$, as desired.

This example, along with the examples we saw last lecture, demonstrates that many constant probability algorithms can be converted to high probability algorithms relatively easily. So, before designing a high probability algorithm, you should always consider whether a constant probability algorithm would be sufficient for your problem. Moreover, we can generalize to the following this heuristic:

Heuristic 3.2 *When designing high probability algorithms, the sample / query complexity $q(\delta)$ of your algorithm should be at least as good as $q(\delta) = O(q(\frac{1}{3})\log(\frac{1}{\delta}))$. Otherwise, a constant probability algorithm combined with a boosting technique (e.g. finding the median or max of solutions returned by a constant probability algorithm) will improve the δ dependence to the multiplicative $\log(\frac{1}{\delta})$ factor. Moreover, note that the $\log(\frac{1}{\delta})$ term is not tight for all problems; we can sometimes do better than this!*

3 Concentration Inequalities for Mean Estimation

First, we'll derive two concentration inequalities – Hoeffding's inequality and Bernstein's inequality – and use them to evaluate the sample complexity of the sample mean when estimating the true mean of a collection of i.i.d. random variables.

Remark 3.3 For the theorems below, we'll be assuming that all random variables are one dimensional.

Lemma 3.4 ***(Hoeffding's Lemma)*** *Let X be a real-valued random variable so that $\mathbb{P}(X \in [a,b]) = 1$ for some $a < b$ and $\mathbb{E}[X] = 0$. Then, $\mathbb{E}[e^{tX}] \leq e^{\frac{t^2(b-a)^2}{8}}$.*

Remark 3.5 The proof of Hoeffding's lemma is complex and involves some calculus, so it was not be covered in class. Nevertheless, the intuition is to notice $x \to e^{xt}$ is convex and thus Jensen's inequality can be applied.

Theorem 3.6 ***(Hoeffding's Inequality)*** *Let $\{X_i\}_{i=1}^n$ be independent random variables so that $\mathbb{P}(X_i \in [a,b]) = 1$ for some $a < b$, and let $\epsilon > 0$. Then:*

1. $\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \geq \epsilon) \leq e^{\frac{-2n\epsilon^2}{(b-a)^2}}$
2. $\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \leq -\epsilon) \leq e^{\frac{-2n\epsilon^2}{(b-a)^2}}$
3. $\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) \leq 2e^{\frac{-2n\epsilon^2}{(b-a)^2}}$

Proof. We will first show (1) and use it to easily prove (2) and (3). Now,

$$\begin{aligned}
\mathbb{P}(\overline{X}_m - \mathbb{E}[\overline{X}_n] \geq \epsilon) &\leq \inf_{t>0} \frac{\mathbb{E}[e^{t(\overline{X}_n - \mathbb{E}[\overline{X}_n])}]}{e^{t\epsilon}} \\
&= \inf_{t>0} \frac{\mathbb{E}[e^{\frac{t}{n}(\sum_{i=1}^n X_i - \mathbb{E}[X_i])}]}{e^{t\epsilon}} \\
&= \inf_{t>0} \frac{\prod_{i=1}^n \mathbb{E}[e^{\frac{t}{n}(X_i - \mathbb{E}[X_i])}]}{e^{t\epsilon}}
\end{aligned}$$

by Lemma (2.8) and since $X_1, \ldots, X_n$ are independent. By applying Hoeffding's lemma to this inequality, we have that

$$\begin{aligned}\inf_{t>0} \frac{\prod_{i=1}^n \mathbb{E}[e^{\frac{t}{n}(X_i - \mathbb{E}[X_i])}]}{e^{t\epsilon}} &\leq \inf_{t>0} \frac{\prod_{i=1}^n e^{\frac{(t^2/n^2)(b-a)^2}{8}}}{e^{t\epsilon}} \\ &= \inf_{t>0} \frac{e^{\frac{(nt^2/n^2)(b-a)^2}{8}}}{e^{t\epsilon}} \\ &= \inf_{t>0} e^{\frac{(t^2/n)(b-a)^2}{8} - t\epsilon}\end{aligned}$$

Recall that e^x is an increasing function, so the infimum of $e^{\frac{(t^2/n)(b-a)^2}{8} - t\epsilon}$ will be determined by the minimum value of $\frac{(t^2/n)(b-a)^2}{8} - t\epsilon$. Furthermore, this is a quadratic function, which implies that the minimum of $\frac{(t^2/n)(b-a)^2}{8} - t\epsilon$ will be $\frac{-y}{2x}$ where $x = \frac{(b-a)^2}{8n}$ and $y = \epsilon$. Therefore, the minimum occurs at $t = \frac{4n\epsilon}{(b-a)^2}$. Moreover, since $t > 0$, we can bound the infimum by $t = \frac{4n\epsilon}{(b-a)^2}$.

Now, it follows that

$$\inf_{t>0} e^{\frac{(t^2/n)(b-a)^2}{8} - t\epsilon} = e^{\frac{-2n\epsilon^2}{(b-a)^2}}$$

Hence, inequality (1) follows. By negating $\overline{X}_n$ and translating, we get inequality (2).

Finally, by combining inequalities (1) and (2), we have that

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) = \mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \geq \epsilon) + \mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \leq -\epsilon) \leq 2e^{\frac{-2n\epsilon^2}{(b-a)^2}}$$

which proves inequality (3). □

Corollary 3.7 *Suppose that $\{X_i\}_{i=1}^n$ are i.i.d. random variables so that $\mathbb{P}(X_i \in [a, b]) = 1$, with $a < b$, holds for all X_i. Then, given $\epsilon > 0$, the sample complexity of the sample mean required to estimate $\mathbb{E}[X]$ to an additive error ϵ while only failing with probability at most δ is $n = O(\frac{\sigma^2}{\epsilon^2} + \frac{M}{\epsilon}) \log(\frac{1}{\delta})$.*

Proof. Let $\epsilon > 0$. Now, the probability of failure for this problem is $\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon)$. Furthermore, by Hoeffding's inequality, we have that $\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) \leq 2e^{\frac{-2n\epsilon^2}{(b-a)^2}}$. Hence, it follows that

$$2e^{\frac{-2n\epsilon^2}{(b-a)^2}} \leq \delta \iff n \geq \frac{(b-a)^2}{2\epsilon^2} \log\left(\frac{2}{\delta}\right)$$

Thus, we should select $n = O\left(\frac{(b-a)^2}{\epsilon^2} \log\left(\frac{1}{\delta}\right)\right)$ to estimate the mean of $\{X_i\}_{i=1}^n$ to an additive error ϵ. □

Is this a good sample complexity? In other words, can we exploit any more information about $\{X_i\}_{i=1}^n$ to reduce the sample complexity of this problem? Yes, in many cases we can further reduce the sample complexity. To see why this is intuitively true, suppose you have a collection of i.i.d. random variables $X_1, \ldots, X_n$ that are drawn from a distribution that has high probability mass in the interval $[0, 1]$ and zero mass outside of this interval,

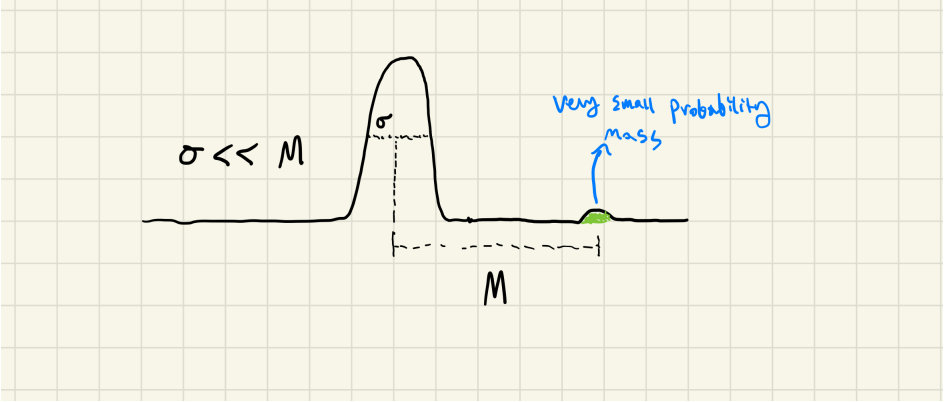

Figure 1: How outliers can hurt sample complexity when using Hoeffding's inequality. In this case, the interval $[a, b]$ must be large enough to contain the points located in the green probability mass, making $[a, b]$ quite large and thus increasing the sample complexity. In this case, M represents the distance from the mean of the distribution to the center of the outlier probability mass.

with the exception of one small interval $[\alpha, \beta]$ with non-zero probability mass and $1 \ll \alpha$. Then, since $\mathbb{P}(X_i \in [\alpha, \beta]) > 0$ for all X_i, b is forced to be greater than or equal to β (See Figure 1 for a visual example of this). However, if the mass in $[\alpha, \beta]$ is sufficiently small, then n samples will not see anything outside $[0, 1]$, making Hoeffding's inequality very loose. The variance is a much more robust way to measure the "width" of a distribution than the interval covering all its probability mass.

Hence, the robustness of variance motivates our next inequality: **Bernstein's Inequality**, which uses variance information to achieve a tighter bound.

Theorem 3.8 *(**Bernstein's Inequality**) Let $\{X_i\}_{i=1}^n$ be independent random variables so that $\mathbb{P}(|X_i - \mathbb{E}[X_i]| \leq M) = 1$ holds for all X_i for some $M \geq 0$. Also, let $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}[X_i]$. Then,*

$$\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n]) \leq \exp \frac{-n\epsilon^2}{2\sigma^2 + \frac{2M\epsilon}{3}}$$

Remark 3.9 The proof of this is quite complicated and was not covered in lecture.

Corollary 3.10 *Suppose that $\{X_i\}_{i=1}^n$ are i.i.d. random variables so that $\mathbb{P}(|X_i - \mathbb{E}[X_i]| \leq M) = 1$ holds for all X_i for some $M \geq 0$. Then, given $\epsilon > 0$, the sample complexity of the sample complexity required to estimate $\mathbb{E}[X]$ to an additive error ϵ while only failing with probability at most δ is $n = O(\frac{\sigma^2}{\epsilon^2} + \frac{M}{\epsilon}) \log(\frac{1}{\delta})$.*

Proof. Fix $\epsilon > 0$. Then, the probability of failure is $\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \geq \epsilon)$. By Bernstein's inequality, we have that $\mathbb{P}(\overline{X}_n - \mathbb{E}[\overline{X}_n] \geq \epsilon) \leq \exp \frac{-n\epsilon^2}{2\sigma^2 + \frac{2M\epsilon}{3}}$. Therefore, it follows that

$$\exp \frac{-n\epsilon^2}{2\sigma^2 + \frac{2M\epsilon}{3}} \leq \delta \iff \frac{n\epsilon^2}{2\sigma^2 + \frac{2M\epsilon}{3}} \geq \log\left(\frac{1}{\delta}\right) \iff n \geq \left(2\frac{\sigma^2}{\epsilon^2} + \frac{2}{3}\frac{M}{\epsilon}\right)\log\left(\frac{1}{\delta}\right)$$

Therefore, $n = O\left((\frac{\sigma^2}{\epsilon^2} + \frac{M}{\epsilon})\log(\frac{1}{\delta})\right)$, as desired. □

So, making essentially the same assumption that the random variables are bounded in an interval of width $O(M) = O(b-a)$, we have two possible sample complexities for the sample mean to choose from: $n = O\left(\frac{(b-a)^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$ (derived from Hoeffding) and $n = O\left((\frac{\sigma^2}{\epsilon^2} + \frac{M}{\epsilon})\log(\frac{1}{\delta})\right)$ (derived from Bernstein) when estimating the mean of independent random variables. It's therefore natural to ask if and when the second sample complexity is better than the first.

The fundamental insight is that when the standard deviation σ of the random variables is significantly smaller than M, in other words $\sigma \ll M$, the sample complexity provided by Bernstein's inequality is tighter since the bound provided by Hoeffding's inequality grows quadratically with M. This is intuitive because factoring in information about variance will allow us to better approximate the underlying behavior of the random variables, requiring fewer samples to approximate their mean. In the regime where $\sigma \approx M$, the two sample complexities are asymptotically the same, so Bernstein's inequality will provide minimal benefit over Hoeffding's inequality. Nevertheless, Bernstein's inequality provides a sample complexity at least as good as that given by Hoeffding's inequality because $\sigma^2 \leq O(M^2)$ always holds.

4 Comparison with CLT; Median of Means Method

In statistics, the sample mean of a collection of independent random variables is often used to estimate their true mean. Moreover, the central limit theorem states that the sample mean $\overline{X}_n$ of a collection of i.i.d. random variables $X_1, \ldots, X_n$ exhibits Gaussian-like behavior [1] as $n \to \infty$. Yet, as we'll see, in the non-asymptotic regime, when estimating $\mathbb{E}[\overline{X}_n]$ with high probability, $\overline{X}_n$ does not behave necessarily like a Gaussian – no matter what concentration inequality you use to bound the sample mean. The upside, however, is that in the constant probability regime (i.e. when the probability of error is fixed), $\overline{X}_n$ does provide Gaussian-like guarantees on sample complexity! Using this, we will introduce the **Median of Means** algorithm which estimates $\mathbb{E}[\overline{X}_n]$ with high probability while also giving Gaussian-like performance.

To start our analysis, recall that given i.i.d. Gaussian random variables $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, we have that

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) \approx \exp\frac{-n\epsilon^2}{2\sigma^2}$$

Therefore,

$$\exp\frac{-n\epsilon^2}{2\sigma^2} \leq \delta \iff n \geq 2\frac{\sigma^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)$$

So, the sample complexity of the sample mean is $n \approx 2\frac{\sigma^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)$ when $X_1, \ldots, X_n$ are Gaussian. Given this, we can show that the sample complexity derived from Bernstein's inequality is not Gaussian in all regimes:

Proposition 3.11 *$n = O((\frac{\sigma^2}{\epsilon^2} + \frac{M}{\epsilon})\log(\frac{1}{\delta}))$ is worse than $n = O(\frac{\sigma^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right))$ and thus the sample complexity of the sample mean given by Bernstein's inequality does not give Gaussian-behavior in all regimes.*

Proof. There are three distinct cases:

[1] A similar statement holds for independent but not identical random variables, requiring additional constraints on the moments of these random variables. See Lyapunov's CLT for more information.

1. **Case 1**: $\epsilon \gg \frac{\sigma^2}{M}$. Since $\frac{\sigma^2}{\epsilon^2}$ has an inverse square dependence on ϵ, we have that $\frac{M}{\epsilon} \gg \frac{\sigma^2}{\epsilon^2}$. Therefore, $O\left((\frac{\sigma^2}{\epsilon^2} + \frac{M}{\epsilon})\log(\frac{1}{\delta})\right)$ provides a worse sample complexity than $O\left(\frac{\sigma^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$ in this case. Note that this is the case that concretely shows that Bernstein's inequality does not give a Gaussian-like sample complexity for the sample mean.

2. **Case 2**: $\epsilon = \Theta(\frac{\sigma^2}{M})$. Then, $M = \Theta(\frac{\sigma^2}{\epsilon})$. So then follows that $M \approx \frac{\sigma^2}{\epsilon}$. So by substituting this value of M into $O\left(\frac{\sigma^2}{\epsilon^2} + \frac{M}{\epsilon}\right)$, we have that

$$O\left(\frac{\sigma^2}{\epsilon^2} + \frac{M}{\epsilon}\right) \approx O\left(\frac{\sigma^2}{\epsilon^2}\right)$$

   Therefore, Bernstein's inequality gives roughly a Gaussian sample complexity for the sample mean in this case.

3. **Case 3**: $\epsilon \ll \frac{\sigma^2}{M}$. Since $\frac{\sigma^2}{\epsilon^2}$ has an inverse square dependence on ϵ, we have that $\frac{M}{\epsilon} \ll \frac{\sigma^2}{\epsilon^2}$. Therefore, since $\frac{\sigma^2}{\epsilon^2}$ dominates $\frac{M}{\epsilon}$, Bernstein's inequality gives the same sample complexity for the sample mean as the Gaussian case.

□

Since Bernstein's inequality does not give Gaussian-like performance in some regimes, it's pertinent to consider whether **any** concentration inequality can guarantee a Gaussian-like sample complexity for the sample mean. **Catoni 2012** provides a lower bound on the sample complexity for the sample mean, which answers this question negatively:

Theorem 3.12 ***(Catoni 2012)***[2] *Given a collection $X_1, \ldots, X_n$ of independent random variables, the sample mean $\overline{X}_n$ needs $\Omega(\frac{\sigma^2}{\epsilon^2\delta})$ samples, assuming that the second moment of $\overline{X}_n$ is finite.*

Given this result, we know that there exists a bad distribution $\mathcal{D}$ so that the sample mean requires at least $\Omega(\frac{\sigma^2}{\epsilon^2\delta})$ samples to estimate the mean of $\mathcal{D}$ to additive error ϵ, with probability at least $1-\delta$. Hence, in the high probability regime (when δ is not fixed), the sample complexity of the sample mean varies inverse linearly in δ, which blows up when δ is made small. Conversely, the Gaussian sample complexity for the sample mean grows at a much slower rate of $\log(\frac{1}{\delta})$. Therefore, the sample mean is a poor estimator of the true mean of a collection of independent random variables because it does not have Gaussian-like performance, as claimed by the CLT, for some distributions.

However, by examining the lower bound $\Omega(\frac{\sigma^2}{\epsilon^2\delta})$ closely, you'll notice that if we fix δ as a constant, the lower bound has the same form as the sample complexity of the Gaussian case! Namely, the Gaussian sample complexity for the sample mean $2\frac{\sigma^2}{\epsilon^2}\log(\frac{1}{\delta})$ differs from $\frac{\sigma^2}{\epsilon^2\delta}$ only by a constant factor when δ is a fixed constant. So, in the constant probability regime, the sample mean could be a Gaussian-like estimator. In fact, we can easily demonstrate that the sample mean has Gaussian sample complexity in this regime. If we assume $\{X_1, \ldots, X_n\}$ is a collection of i.i.d. random variables with variance σ^2, then

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2 n}$$

[2] https://arxiv.org/pdf/1009.2048.pdf

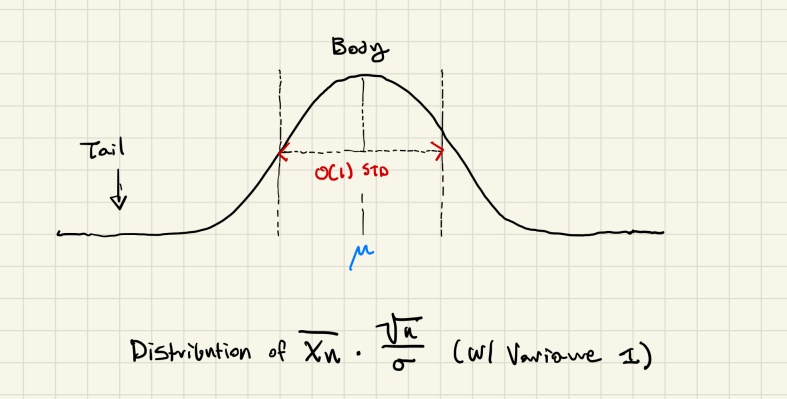

Figure 2: Concentration of the Sample Mean

by Chebyshev's inequality. Given this bound on $\overline{X}_n$, it follows that the sample complexity is

$$\frac{\sigma^2}{\epsilon^2 n} \leq \delta \iff n \geq \frac{\sigma^2}{\epsilon^2 \delta}$$

So, if we fix δ as a constant, then the sample complexity of the sample mean becomes $n = O(\frac{\sigma^2}{\epsilon^2})$, which is Gaussian. Therefore, in the constant probability regime, the sample mean is a good estimator of the true mean of $X_1, \ldots, X_n$.

The above observations can alternatively be summarized by Figure 2, which shows a "bump" denoting the distribution of the sample mean normalized to variance 1, namely $\frac{\sqrt{n}}{\sigma}\overline{X}_n$. We can divide the bump into two regimes: the "body" which is within a constant number of standard deviations from the mean, and the "tail" which is the rest of the distribution. Thus, by Chebyshev's inequality, the body is always Gaussian-like, in a big-O sense. The tail, on the other hand, can be heavy (having a lot of mass, decaying very slowly) and badly-behaved, from Catoni's lower bound.

Moreover, since the sample mean provides a good constant probability estimate for the true mean, we can combine it with the median trick from last lecture to design a high probability algorithm known as **Median of Means**:

Algorithm 1: Median of Means

---

input : $X_1, \ldots, X_n$ samples where $n = O(\frac{\sigma^2}{\epsilon^2}\log(\frac{1}{\delta}))$

output: A mean estimate μ so that $|\mu - \mathbb{E}[\overline{X}_n]| \geq \epsilon$ with probability less than δ

steps:

1. Divide samples in to $m = \Theta(\log(\frac{1}{\delta}))$ groups
2. Compute sample mean S^i for each group i, where each group is of size $O(\frac{\sigma^2}{\epsilon^2})$
3. Output median of $\{S_i\}_{i=1}^m$

In the median of means algorithm, the sample mean of each group S^i is within ϵ of the true mean with probability $\geq \frac{2}{3}$. This is done by selecting a group size of $\frac{3\sigma^2}{\epsilon^2} = O(\frac{\sigma^2}{\epsilon^2})$ in Step 2 of the algorithm. Thus, using the median technique from last lecture, the success probability of the entire algorithm is at least $1 - \delta$.

Remark 3.13 In conclusion, the median of means algorithm should be preferred over the sample mean for estimating the mean of a collection of independent random variables with high probability, as it provides Gaussian-like performance that is unattainable with the sample mean in general.

5 Conclusions

- Constant probability algorithms are incredibly useful and can be used to build efficient high probability algorithms; moreover, when designing high probability algorithms, a constant probability counterpart should serve as a baseline for performance.
- While concentration inequalities are very useful, they are only an analysis technique and shouldn't be blindly applied.
- Don't forget about Chebyshev's inequality, which can be powerful (and tight) when used correctly with other tools.

6 Additional Content

Theorem 3.14 ***(McDiarmid's Inequality)*** *Let $\{X_i\}_{i=1}^n$ be independent random variables, where each $X_i : \Omega \to S_i$ for some set S_i. Consider $f : \prod_{i=1}^n S_i \to \mathbb{R}$, and suppose that f is C_i-lipschitz in the i-th coordinate of the input for all i. Then, for any $\epsilon > 0$,*

$$\mathbb{P}(f(x_1, \ldots, x_n) - \mathbb{E}[f] \geq \epsilon) \leq \exp \frac{-2\epsilon^2}{\sum_{i=1}^n C_i^2}$$

Remark 3.15 By C_i-lipschitz, we mean that given a $C_i \in \mathbb{R}_{\geq 0}$ and any two points $x = (x_1, \ldots, x_i, \ldots, x_n), x' = (x_1, \ldots, x_i', \ldots, x_n) \in \prod_{i=1}^n S_i$ that differ by only their i-th corrdinate, we have that

$$|f(x) - f(x')| \leq C_i$$

Remark 3.16 Notice the connection between McDiarmid's inequality and Hoeffding's inequality. Simply set $f(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i$ and you'll get back Hoeffding's inequality.