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

Lecture 3: Concentration Inequalities and Median of Means

*Lecturer: Jasper Lee* *Scribe: Dasha Buka, Ravi Kini*

# 1 Overview

In the previous class, we used concentration inequalities to boost the success of decision algorithms.

**What if we tried to boost the success of an optimizer?**

**Definition 3.1** (Optimizer)**.** An algorithm which seeks to maximize an objective function $f(x)$ over variable $x$ in constraint set $C$, i.e. finding $\arg\max_{x\in C} f(x)$.

Suppose algorithm $\mathcal{A}$ returns $x_{\text{out}}$ such that $f(x_{\text{out}}) \geq \text{OPT} - \epsilon$ with probability $\geq 2/3$, where $\text{OPT} = \max\limits_{x\in C}(f(x))$. Like our previous constructions, we wish to construct $\mathcal{A}'$ that runs $\mathcal{A}$ $O(\log \frac{1}{\delta})$ times and succeeds with probability $1-\delta$.

```
Algorithm 1 Algorithm $\mathcal{A}'$
  1. Run $\mathcal{A}$ $O(\log \frac{1}{\delta})$ times.
  2. Take the maximum of these outputs.
```

The probability that $\mathcal{A}'$ fails is equal to the probability that all of the runs output results such that $f(x_{\text{out}}) < \text{OPT} - \epsilon$, as the maximum output will also not be at least $\text{OPT} - \epsilon$. This occurs with probability at most $\left(1-\frac{2}{3}\right)^n = \frac{1}{3^n}$, which we bound by $\delta$.

$$\delta = \frac{1}{3^n} \tag{1}$$

$$n = \frac{\log 1/\delta}{\log 3} \tag{2}$$

Running $\mathcal{A}$ $O(\log \frac{1}{\delta})$ times as a part of $\mathcal{A}'$ then produces a correct solution with probability $1-\delta$, as desired.

# 2 More concentration inequalities

**Theorem 3.2** (Hoeffding's inequality)**.** *Let $\{X_i\}_{i=1}^n$ be independent random variables such that $\Pr(X_i \in [a_i, b_i]) = 1$, with $M_i = b_i - a_i$. Then,*

1. $\Pr(\overline{X}_n - \mathbb{E}(\overline{X}_n) \geq \epsilon) \leq e^{-2n\epsilon^2/M^2}$
2. $\Pr(\overline{X}_n - \mathbb{E}(\overline{X}_n) \leq -\epsilon) \leq e^{-2n\epsilon^2/M^2}$
3. $\Pr(|\overline{X}_n - \mathbb{E}(\overline{X}_n)| \geq \epsilon) \leq 2e^{-2n\epsilon^2/M^2}$.

*where* $M = \max_i(M_i)$.

We can take advantage of the following lemma, which we will not prove, to complete the Chernoff bound in our proof for the first inequality in Theorem 3.2. The second bound arises from negating the random variable, and exploiting symmetry, and the third bound is just a union bound of the first two statements.

**Lemma 3.3** (Hoeffding's lemma)**.** *If random variable* $x_i$ *bounded in* $[a_i, b_i]$ *with probability* 1*, then* $\mathbb{E}(e^{tX_i}) \leq e^{t^2(b-a)^2/8}$.

*Proof.* For the proof, we will first bound the moment generation function of $\overline{X}_n$ over $t$, which we take from Hoeffding's lemma. Then, we apply the generic Chernoff bound ($P(X > a) \leq \inf\limits_{t>0} \frac{\mathbb{E}(e^{tX})}{e^{ta}}$), letting $X = \overline{X}_n - \mathbb{E}(\overline{X}_n)$ and $a = \epsilon$.

$$\begin{aligned}
\Pr(\overline{X}_n - \mathbb{E}(\overline{X}_n) > \epsilon) &\leq \inf_{t>0} \frac{\mathbb{E}(e^{t(\overline{X}_n - \mathbb{E}(\overline{X}_n))})}{e^{t\epsilon}} = \inf_{t>0} \frac{\mathbb{E}(e^{\frac{t}{n}(\sum X_i - \mathbb{E}(X_i))})}{e^{t\epsilon}} \\
&\leq \inf_{t>0} \frac{\prod \mathbb{E}(e^{\frac{t}{n}(X_i - \mathbb{E}(X_i))})}{e^{t\epsilon}} \\
&\leq \inf_{t>0} \frac{\prod e^{\frac{t^2}{n^2}(b_i - a_i)^2/8}}{e^{t\epsilon}} \\
&\leq \inf_{t>0} \frac{e^{\frac{t^2}{n}(b-a)^2/8}}{e^{t\epsilon}} \\
&\leq \inf_{t>0} e^{\frac{t^2}{n}(b-a)^2/8 - t\epsilon}
\end{aligned}$$

We have the quadratic $t^2 \frac{(b-a)^2}{8n} - t\epsilon$ in the exponent, which will be minimum at $t = \frac{4n\epsilon}{(b-a)^2}$. Plugging this value of $t$ back in as the right hand side and letting $M = b - a$ yields Hoeffding's inequality. □

We now try to find the sample complexity for getting an error of at most $\epsilon$ with probability $1 - \delta$ using the sample mean. The probability of the error exceeding $\epsilon$, by Hoeffding's inequality, is $e^{-2n\epsilon^2/M^2}$, which we bound with $\delta$.

$$\log 2/\delta = \frac{2n\epsilon^2}{M^2} \tag{3}$$

$$n = \frac{M^2}{2\epsilon^2} \log 2/\delta \tag{4}$$

However, this bound is not always tight, even though the bound depends only on the ratio $M/\epsilon$ and is (correctly) invariant to the same rescaling on $M$ and $\epsilon$. In any situation where most of the mass is concentrated in one area with some small outlier, the bound is not tight, as although $M$ is large, making $n$ large, the number of samples needed for the sample mean to estimate the true mean is practically much smaller.

We can improve this with the next inequality.

**Theorem 3.4.** *(Bernstein's inequality) Let* $\{X_i\}$ *be independent random variables such that* $\Pr(|X_i - \mathbb{E}(X_i)| \leq M) = 1$ *and* $\sigma^2 = \operatorname{Var}(\overline{X}_n)$*, then*

$$\Pr(|\overline{X}_n - \mathbb{E}(X_n)| > \epsilon) \leq 2e^{-n\epsilon^2/(2\sigma^2 + 2M\epsilon/3)}$$

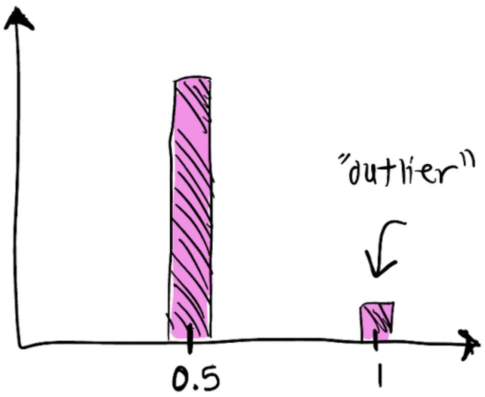

Figure 1: An example that makes the bound less tight.

The proof of this inequality is technical and was not covered in class, and will not be included here. We can find the sample complexity for getting an error of at most $\epsilon$ with probability $1-\delta$ in a manner similar to what was done with Hoeffding's inequality. The probability of the error exceeding $\epsilon$, by Bernstein's inequality, is $2e^{-n\epsilon^2/(2\sigma^2+2M\epsilon/3)}$, which we bound with $\delta$.

$$\log 2/\delta = n\epsilon^2/(2\sigma^2 + 2M\epsilon/3) \tag{5}$$

$$n = \left(\frac{2\sigma^2}{\epsilon^2} + \frac{2M}{3\epsilon}\right)\log\frac{2}{\delta} \tag{6}$$

Since this inequality takes into account both the width of the interval over which the random variable is distributed and the variance of the random variable within the interval, it provides a tighter bound on the sample complexity than Hoeffding's inequality when $\sigma$ is small in comparison to $M$, as the complexity grows linearly with $M$ instead of quadratically. When $\sigma$ is comparable to $M$, both provide asymptotically similar bounds that are quadratic in $M$.

### 2.1 Comparison to CLT

By the Central Limit Theorem, $\overline{X}_n$ converges to $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ at large $n$, and we want to (roughly) compute $n$ such that $\Pr(\overline{X}_n > \mu+\epsilon) \approx \delta/2$. Using the approximation $\Pr(\mathcal{N}(0,1) > \epsilon') \approx e^{-\epsilon'^2/2} \Rightarrow \Pr(\mathcal{N}(\mu, \frac{\sigma^2}{n}) > \mu+\epsilon) \approx e^{-n\epsilon^2/2\sigma^2}$, we find that $n \approx \frac{2\sigma^2}{\epsilon^2}\log 2/\delta$. Comparing the last approximation to Bernstein's, we see that Bernstein's is larger.

We compare the sample complexity predicted by the Central Limit Theorem to what was actually obtained from Bernstein's inequality. There are three regimes.

**Comparing $\boldsymbol{\epsilon}$ to $\boldsymbol{\frac{\sigma^2}{M}}$ :**

- $\boldsymbol{\epsilon \gg \frac{\sigma^2}{M}}$: In this case $M/\epsilon \gg \sigma^2/\epsilon^2$, so the sample complexity is not (remotely) close to CLT.
- $\boldsymbol{\epsilon \approx \frac{\sigma^2}{M}}$ (up to constants): In this case $M/\epsilon \approx \sigma^2/\epsilon^2$, close to CLT but beaten by constant factor.

- $\boldsymbol{\epsilon \ll \frac{\sigma^2}{M}}$: Here $M/\epsilon \ll \sigma^2/\epsilon^2$, so Bernstein's sample complexity is comparable to CLT.

**Theorem 3.5.** *(Catoni 2012)*[1] *In general, the sample complexity of sample mean is* $n = \Omega(\frac{\sigma^2}{\epsilon^2 \delta})$.

This bound is predicted by Chebyshev's inequality (which would give a matching upper bound, up to constants). It is exponentially weaker compared to the predictions of the central limit theorem.

## 2.2 Summarizing what we know in one picture

Figure 2 shows what we know about the sample mean's distribution, in one picture. Specifically, we consider the distribution of $\frac{\sqrt{n}}{\sigma}\overline{X}_n$, which has variance 1.

This distribution has a central region, of width $O(1)$ around the true mean, which we know for sure is "Gaussian-like up to a constant". This is given by Chebyshev's inequality. More specifically, we can rephrase Chebyshev as $\Pr(|X - \mu| > \epsilon\sigma) \leq \frac{1}{\epsilon^2}$. In this context, we can say $\Pr(|X - \mu| > O(1)\sigma) \leq O(1)$.

Outside of this region, things can look arbitrarily bad. Even though the Central Limit Theorem tells us that, as $n \to \infty$, eventually the *entire* distribution must look Gaussian, there is no way to guarantee how fast that happens (in terms of $n$) that applies to all distributions. Put another way, in the previous paragraph, we know for sure that the central region of $O(1)$ width is Gaussian-like, and we know that the Gaussian-like region will grow with $n$ as $n \to \infty$ until it covers the entire number line (which is great!). However, the speed at which the Gaussian-like region grows may be arbitrarily slow, depending on what the underlying distribution is (not so great...).

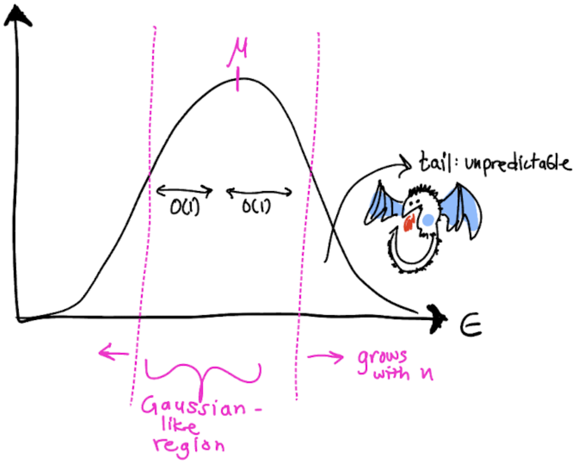

Figure 2: Concentration of the sample mean

# 3 Median-of-means algorithm

This is, for now, our last attempt to match the Central Limit Theorem. Again, the goal is to construct a new mean estimator that uses $O(\frac{\sigma^2}{\epsilon^2} \log \frac{1}{\delta})$ samples, to get to $\epsilon$ estimation

[1] Link

error with probability $1 - \delta$.

**Algorithm 2** Algorithm $\mathcal{A}'$

```
1. Repeat the following $O(\log \frac{1}{\delta})$ time: Take the sample mean of $\sigma^2/\epsilon^2$ samples ($\overline{X}_{\sigma^2/\epsilon^2}$).
2. Return the median of these outputs.
```

The analysis of median of means is straightforward: by Chebyshev's, each sample mean is within $\epsilon$ of the true mean with probability at least $2/3$. The median trick from last class boosts the success probability to $1 - \delta$.

Even though the sample complexity of median-of-means is optimal (up to constants), in practice, the estimator has terrible performance. In the last class we'll discuss how to do mean estimation well.

## 4 Takeaway

1. Don't use the sample mean blindly, the common claims about what it does are misleading.
2. Don't blindly use concentration inequalities. Despite using more "elementary" techniques (like Chebyshev's) than Hoeffding's and Bernstein's, we produced a better (in theory) algorithm with median-of-means.
3. Constant probability results can be useful sometimes, if applied correctly (for example, the last algorithm).
4. Don't use the last result in practice.