Bonus Lecture: Concentration Inequalities and the Median-of-Means Estimator

Hsin-Yuan Huang (Robert)

Caltech

1 Introduction: The Challenge of Estimation

In science and engineering, we often want to determine a property of a system, say the expected value μ of some observable. We can't measure it infinitely many times, so we take a finite number of samples, $X_1, X_2, \ldots, X_N$, and use them to produce an estimate. The most natural estimator for the mean $\mu = \mathbb{E}[X]$ is the **sample mean**:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

While this is a great estimator in many well-behaved scenarios, it has a big weakness: it is very sensitive to outliers. A single corrupted measurement with a very large value can completely throw off the estimate. This is a common problem in quantum experiments where measurement errors can be non-Gaussian, or when dealing with distributions that have "heavy tails". This is exactly the situation arised from classical shadow tomography using unitary 3-designs. If we are unlucky, the random variables X_i can become exponentially large due to the (2^n+1) factor in the classical shadow snapshot $\hat{\rho}_i = (2^n+1)U_i^\dagger |b_i\rangle\langle b_i| U_i - \mathbb{I}$ of the unknown quantum state ρ.

How can we construct an estimator that is robust to outliers? This bonus lecture introduces the **median-of-means (MoM)** estimator, a simple but powerful alternative to the sample mean. To prove its effectiveness, we first need to build a toolkit of fundamental results from probability theory known as **concentration inequalities** that we have seen a few times in the class. These inequalities tell us how likely it is that a random variable deviates from its expected value.

2 The Basic Toolkit: Concentration Inequalities

We will build our theory from the ground up, starting with the simplest inequality.

2.1 Lemma 1: Markov's Inequality

This is the bedrock upon which other inequalities are built. It relates the probability of a non-negative random variable being large to its expectation.

Lemma 1 (Markov's Inequality)**.** *Let X be a non-negative random variable ($X \geq 0$) with a finite expectation. Then for any $a > 0$,*

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

Proof. The proof is elegant and straightforward. Let's consider the expectation of X:

$$\mathbb{E}[X] = \int_0^\infty x f(x) dx = \int_0^a x f(x) dx + \int_a^\infty x f(x) dx$$

where $f(x)$ is the probability density function. Since $x \geq 0$, the first integral is non-negative. In the second integral, $x \geq a$. Therefore:

$$\mathbb{E}[X] \geq \int_a^\infty x f(x) dx \geq \int_a^\infty a f(x) dx = a \int_a^\infty f(x) dx$$

The final integral is precisely the probability that $X \geq a$. So, we have:

$$\mathbb{E}[X] \geq a \cdot P(X \geq a)$$

Rearranging gives the desired result. □

2.2 Lemma 2: Chebyshev's Inequality

By applying Markov's inequality to the squared deviation from the mean, $(X - \mu)^2$, we get a much more useful bound that incorporates the variance. This is Chebyshev's inequality, our main workhorse for the first part of the analysis.

Lemma 2 (Chebyshev's Inequality)**.** *Let X be a random variable with finite mean $\mu = \mathbb{E}[X]$ and finite variance $\sigma^2 = Var(X) = \mathbb{E}[(X - \mu)^2]$. For any $\epsilon > 0$,*

$$P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

Proof. Define a new random variable $Y = (X - \mu)^2$. Note that Y is always non-negative, so we can apply Markov's inequality to it. The event $|X - \mu| \geq \epsilon$ is identical to the event $(X - \mu)^2 \geq \epsilon^2$, which is $Y \geq \epsilon^2$. Applying Markov's inequality with $a = \epsilon^2$:

$$P(Y \geq \epsilon^2) \leq \frac{\mathbb{E}[Y]}{\epsilon^2}$$

Substituting back the definitions of Y and the variance:

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

This completes the proof. Chebyshev's inequality tells us that if a random variable has small variance, it's unlikely to be found far from its mean. □

2.3 Lemma 3: Exponentially Decaying Bounds (Hoeffding's Inequality)

For sums of *independent* random variables, we can achieve much stronger, exponentially decaying bounds. A particularly useful version is Hoeffding's inequality, which applies to sums of bounded random variables.

Lemma 3 (Hoeffding's Inequality)**.** *Let $Z_1, \ldots, Z_k$ be independent random variables such that $Z_j \in [0, 1]$ for all j. Let $\hat{\mu}_k = \frac{1}{k} \sum_{j=1}^k Z_j$ and $\mathbb{E}[\hat{\mu}_k] = \nu$. Then for any $\epsilon > 0$,*

$$P(\hat{\mu}_k - \nu \geq \epsilon) \leq \exp\left(-2k\epsilon^2\right).$$

The exponential dependence $\exp\left(-2k\epsilon^2\right)$ means the probability of large deviations becomes vanishingly small very quickly as k grows. We have seen this before and will use this powerful tool in the analysis of median-of-means estimator.

3 The Median-of-Means Estimator

We now have the tools to define and analyze our robust estimator. The core idea is to blunt the impact of outliers by taking a median of several independently estimated means.

3.1 The Algorithm

The median-of-means estimator is constructed as follows:

1. **Collect data:** Start with N i.i.d. samples $X_1, \ldots, X_N$ from a distribution with mean μ and variance σ^2.
2. **Partition:** Choose an integer k (the number of blocks). Partition the N samples into k disjoint blocks, each of size $m = N/k$. (We assume for simplicity that N is a multiple of k).
3. **Calculate block means:** For each block $j \in \{1, \ldots, k\}$, compute its local sample mean $\hat{\mu}_j$.
$$\hat{\mu}_j = \frac{1}{m} \sum_{i \in \text{Block } j} X_i$$
4. **Find the median:** The final estimate, $\hat{\mu}_{\text{MoM}}$, is the median of these k block means.
$$\hat{\mu}_{\text{MoM}} = \text{median}\{\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k\}$$

The intuition is that averaging within blocks reduces the variance of each $\hat{\mu}_j$. While some blocks might be "contaminated" by outliers leading to a bad $\hat{\mu}_j$, it is unlikely that a *majority* of them are. The median, being a robust statistic, will simply ignore these few bad estimates.

3.2 The Main Theorem and Sample Complexity

The power of the median-of-means estimator is captured by the following theorem. It provides a high-confidence guarantee that is exponentially stronger than what a direct application of Chebyshev's inequality would yield for the standard sample mean.

Theorem 1 (Median-of-Means Guarantee)**.** *Let $X_1, \ldots, X_N$ be i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$. For any desired failure probability $\delta \in (0,1)$, by setting the number of blocks to $k = 8\ln(1/\delta)$, the median-of-means estimator $\hat{\mu}_{MoM}$ satisfies:*

$$P\left(|\hat{\mu}_{MoM} - \mu| > \sigma\sqrt{\frac{32\ln(1/\delta)}{N}}\right) \leq \delta$$

This is a remarkable result. The number of samples needed scales with $\log(1/\delta)$, whereas a simpler analysis using only Chebyshev's would yield a much worse scaling of $1/\delta$. This logarithmic dependence means we can demand extremely high confidence (e.g., a one-in-a-billion failure rate, $\delta = 10^{-9}$) without needing an astronomical number of samples. For practical applications, it's often useful to rearrange the theorem to solve for the number of samples needed.

Corollary 1 (Sample Complexity for MoM)**.** *To estimate the mean μ to within an error ϵ with probability at least $1 - \delta$, it is sufficient to use N samples, where:*

$$N \geq \frac{32\sigma^2}{\epsilon^2} \ln\left(\frac{1}{\delta}\right)$$

Proof of Theorem. The proof is a simple two-stage argument. First, we use Chebyshev's inequality to show that any single block mean is unlikely to be a "bad" estimate. Second, we use Hoeffding's inequality to show that it is *exponentially* unlikely for a majority of the block means to be "bad".

Step 1: Bounding the Unreliability of a Single Block. Let's analyze a single block mean, $\hat{\mu}_j$, computed from $m = N/k$ samples. Its expectation is $\mathbb{E}[\hat{\mu}_j] = \mu$ and its variance is $\text{Var}(\hat{\mu}_j) = \sigma^2/m$. We'll call a block **"bad"** if its mean deviates from the true mean μ by more than some threshold ϵ'. Using Chebyshev's inequality, we can bound the probability of this event:

$$P(|\hat{\mu}_j - \mu| > \epsilon') \leq \frac{\text{Var}(\hat{\mu}_j)}{(\epsilon')^2} = \frac{\sigma^2}{m(\epsilon')^2}$$

The core strategy is to choose this threshold ϵ' such that the probability of any block being "bad" is a small, constant value. Let's fix this probability to be 1/4.

$$\frac{\sigma^2}{m(\epsilon')^2} = \frac{1}{4} \implies \epsilon' = \frac{2\sigma}{\sqrt{m}}$$

So, for this specific choice of error threshold ϵ', we have guaranteed that $P(\text{block } j \text{ is "bad"}) \leq 1/4$.

Step 2: Aggregating Blocks with Hoeffding's Inequality. Now, we consider the collection of k block means. The final estimate $\hat{\mu}_{\text{MoM}}$ fails to be within ϵ' of μ only if the median of $\{\hat{\mu}_1, \ldots, \hat{\mu}_k\}$ is itself a "bad" estimate. This can only happen if **at least half** of the block means are "bad". Let's formalize this. Define k independent indicator random variables: $Z_j = 1$ if block j is bad (i.e., $|\hat{\mu}_j - \mu| > \epsilon'$), and $Z_j = 0$ otherwise. From Step 1, we know the probability $p_j = P(Z_j = 1) \leq 1/4$.

Let $\hat{p} = \frac{1}{k}\sum_{j=1}^{k} Z_j$ be the fraction of bad blocks. The MoM estimator fails only if $\hat{p} \geq 1/2$. Let's bound this probability. The expected fraction of bad blocks is $\nu = \mathbb{E}[\hat{p}] \leq 1/4$. The event we want to bound is $P(\hat{p} \geq 1/2)$. This can be rewritten as:

$$P(\hat{p} - \nu \geq 1/2 - \nu) \leq P(\hat{p} - \nu \geq 1/2 - 1/4) = P(\hat{p} - \nu \geq 1/4)$$

This is precisely the form required by Hoeffding's inequality (Lemma 3), where the deviation is 1/4. Applying the inequality gives:

$$P(\hat{p} \geq 1/2) \leq \exp\left(-2k(1/4)^2\right) = \exp\left(-\frac{k}{8}\right)$$

We want this total failure probability to be at most δ. Setting $\delta = e^{-k/8}$ and solving for k gives us the required number of blocks:

$$k = 8\ln(1/\delta)$$

Step 3: Assembling the Final Result. We now combine the results from the two steps. The final error ϵ reported in the theorem is simply the threshold ϵ' we defined in Step 1.

1. From Step 1: $\epsilon = \epsilon' = 2\sigma/\sqrt{m}$
2. From Step 2: $k = 8\ln(1/\delta)$

Using the relation $m = N/k$, we substitute k into the expression for ϵ:

$$\epsilon = \frac{2\sigma}{\sqrt{N/k}} = \frac{2\sigma\sqrt{k}}{\sqrt{N}} = \frac{2\sigma\sqrt{8\ln(1/\delta)}}{\sqrt{N}} = \sigma\sqrt{\frac{4 \cdot 8\ln(1/\delta)}{N}} = \sigma\sqrt{\frac{32\ln(1/\delta)}{N}}$$

This is precisely the expression for the error in the theorem statement. □