Overview

Submit your writeup, including any code, as a PDF via gradescope.[1] We recommend reading through the entire homework beforehand and carefully using functions for testing procedures, plotting, and running experiments. Taking the time to reuse code will help in the long run!

Data science is a collaborative activity. While you may talk with others about the homework, please write up your solutions individually. If you discuss the homework with your peers, please include their names on your submission. Please make sure any handwritten answers are legible, as we may deduct points otherwise.

Optimal Mean Estimation via Concentration Inequalities

Suppose we observe a sequence of i.i.d. random variables $X_1, \ldots, X_n$. Their distribution is unknown, and has unknown mean μ and known variance σ^2. In this question, we will investigate two different estimators for the mean μ: the sample mean, and the so-called "median of means" estimator. In particular, we will analyze them in terms of how many samples n are required to estimate μ to a given precision ϵ and for a confidence threshold δ.

We'll start with the sample mean for parts (a) - (c): in other words, we'll use $X_1, \ldots, X_n$ to compute an estimate $S_n = \frac{1}{n}\sum_i X_i$ for the mean μ. We want to see what sample size n guarantees that $\mathbb{P}(|\hat{\mu} - \mu| \geq \epsilon) \leq \delta$.

(a) (2 points) Let $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Use Chebyshev's inequality to show that $n = \lceil \frac{\sigma^2}{\delta\epsilon^2} \rceil$ samples are sufficient for $|S_n - \mu| \leq \epsilon$ with probability at least $1 - \delta$.

(b) (3 points) Now assume that each X_i is bounded between a and b. Use Hoeffding's inequality to compute the number of samples n sufficient for $|S_n - \mu| \leq \epsilon$ with probability at least $1 - \delta$. In particular, show that the dependence of n on δ is $O(\log(1/\delta))$.

(c) (2 points) Suppose we can't assume that each X_i is bounded. In that case, we can no longer apply Hoeffding's inequality, and can only use Chebyshev's inequality. This is a problem if we require a very high confidence level: in this part, you'll show why.

For this part only, assume that $\sigma^2 = 1$, $a = -1, b = 1$, and $\epsilon = 0.1$. Make a plot that shows, for particular values of δ, the number of samples n required based on your answers from parts (a) and (b). Your plot should show a range of ten δ values between 1/2 and 1/1000, using `np.geomspace(1/2,1/1000,10)`, and should be shown on a log-log scale. What do you observe?

[1] In Jupyter, you can download as PDF or print to save as PDF

To overcome this problem, we'll replace the sample mean with another estimator, and construct bounded random variables that will help us reason about the new estimator. To construct this estimator, we'll start by considering m groups of X_i, each with fixed size n_0. We'll compute the sample mean for each group, and call these sample means $S^{(1)}, \ldots, S^{(m)}$. Then, we'll use the median of all these group means as our estimate for the mean. The diagram below summarizes our approach.

$$\underbrace{\underbrace{X_1, X_2, \ldots X_{n_0}}_{\text{Sample mean } S^{(1)}} \quad \underbrace{X_{n_0+1}, X_{n_0+2}, \ldots, X_{2n_0}}_{\text{Sample mean } S^{(2)}} \quad \cdots \quad \underbrace{X_{n-n_0+1}, X_{n-n_0+2}, \ldots, X_n}_{\text{Sample mean } S^{(m)}}}_{\textbf{Median } S_n}$$

We do this because even though one such sample mean $S^{(i)}$ might be far from the true mean μ, we hope (and will show) that the median of all of them is more likely to be close to the true mean μ.

(d) (2 points) Fix a sample size $n_0 = \lceil \frac{4\sigma^2}{\epsilon^2} \rceil$. For each of the group means i, we define a binary random variable Z_i:

$$Z_i = \mathbb{1}(|S^{(i)} - \mu| \geq \epsilon).$$

In other words, Z_i is 0 if the corresponding group mean is close to the true mean μ (within ϵ), and 1 otherwise.

Show that $\mathbb{E}[Z_i] \leq 1/4$.

Hint: Z_i is a Bernoulli random variable.

(e) (2 points) We set $S_{\text{Med}} := \text{Median}(\{S^{(1)}, \ldots, S^{(m)}\})$. This is called the *median-of-means estimator*. Explain in words why having $|S_{Med} - \mu| \geq \epsilon$ implies that $\sum_{i=1}^{m} Z_i \geq \frac{m}{2}$.

Hint: If y is the median of m numbers, it means that $\lceil m/2 \rceil$ of the numbers are greater than or equal to y, and similarly $\lceil m/2 \rceil$ of the numbers are less than or equal to y.

(f) (2 points) By taking probabilities, part (e) implies

$$\mathbb{P}(|S_{\text{Med}} - \mu| \geq \epsilon) \leq \mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m} Z_i \geq \frac{1}{2}\right).$$

If we combine this fact with the result of (d), we can show that

$$\mathbb{P}(|S_{\text{Med}} - \mu| \geq \epsilon) \leq \mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m} \left(Z_i - \mathbb{E}[Z_i]\right) \geq \frac{1}{4}\right).$$

Now use Hoeffding's inequality to compute what number m is sufficient to ensure that $|S_{\text{Med}} - \mu| \leq \epsilon$ with probability at least $1 - \delta$. What is the final number of samples of X required?