CSE 521: Design and Analysis of Algorithms I **Fall 2018**

Lecture 2: Concentration Bounds

Lecturer: Shayan Oveis Gharan *10-03-2018* *Scribe:*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

Suppose there is an unknown distribution, D and we want to estimate the mean. A possible suggestion is to draw independent samples

$$x_1, x_2, \ldots, x_n$$

from D and return the empirical average,

$$\frac{1}{n}\sum_{i=1}^{n} x_i.$$

Laws of large number say that as n goes to infinity the empirical average converges to the mean. The question we want to address in this lecture is "how large should n be" in order to get a an ϵ-additive approximation of the true expectation? As a real world application, we can use this idea to estimate the people opinion in polling by asking only a few of the voters randomly.

We start this lecture by a simple example: Suppose that the average GPA in CSE 521 is 3.0 / 4.0. What fraction of the students have received at least a 3.5? It turns out in the worst case 1/7-th fraction have received 0.0 and the rest, i.e., 6/7-th fraction have received 3.5. In other words, the worst case is when everybody who has received below 3.5 indeed got 0 and all of those who got more than 3.5 indeed receive nothing more than 3.5. We can justify this claim using Markov's inequality.

2.1 Markov's Inequality

Theorem 2.1 (Markov's Inequality)**.** *Let $X \geq 0$ be a random variable. Then for all k,*

$$\mathbb{P}\left[X \geq k \cdot \mathbb{E}\left[X\right]\right] \leq \frac{1}{k}$$

equivalently:

$$\mathbb{P}\left[X \geq k\right] \leq \frac{\mathbb{E}\left[x\right]}{k}.$$

So, in our class average GPA example, X denotes the GPA of a random student, $\mathbb{E}\left[X\right] = 3$ and $k = 7/6$. The inequality says at most 6/7 fraction of the students received at least 3.5 or at least 1/7 receive less than 3.5.

Proof. The proof is a simple one line argument,

$$\mathbb{E}\left[X\right] = \sum_i \mathbb{P}\left[X = i\right] \geq \sum_{i \geq k} i \cdot \mathbb{P}\left[X = i\right] \geq \sum_{i \geq k} k \cdot \mathbb{P}\left[X = i\right] = k \cdot \mathbb{P}\left[X \geq k\right]$$

So, $\mathbb{P}\left[X \geq k\right] \leq \mathbb{E}\left[X\right]/k$ as desired. $\square$

Observe that in the above proof is tight, i.e., all inequalities are equalities, if the distribution of X has only two points mass,

$$X = \begin{cases} 0 & \text{w.p. } 1 - 1/k \\ k + \epsilon & \text{w.p. } 1/k \end{cases}.$$

In other words, this example shows that if $\mathbb{E}[X]$ is the only information that we have about X, then Markov's inequality is the best bound we can prove on deviations from the expectation of X.

2.1.1 Applications of Markov's Inequality: Fixed points of permutations

Let $[n] := \{1, \ldots, n\}$. A permutation, $\sigma : [n] \to^{1-1}_{\text{onto}} [n]$, is a bijection between $[n]$ and $[n]$. Suppose we a choose a uniformly random permutation σ. What is the probability that for two i, j, $\sigma_i = i$ and $\sigma_j = j$, i.e., that the permutation has two fixed points?

Let $X_i = \mathbb{I}[\sigma_i = i]$. Let $X = \sum X_i$. Note that X is exactly equal to the number of fixed points of σ. So we want to upper bound $\mathbb{P}[X \geq 2]$. We are going to use Markov's inequality, but first we need to calculate $\mathbb{E}[X]$.

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}\left[\sum X_i\right] \\ &= \sum \mathbb{E}[X_i] \qquad \text{(by linearity of expection, not proven here)} \\ &= \sum_i \mathbb{P}[X_i = 1] \qquad \text{(expectation of an indicator)} \\ &= \sum_i \frac{1}{n} \\ &= 1 \end{aligned}$$

So by Markov Inequality,

$$\mathbb{P}[X \geq 2] \leq \frac{1}{2}.$$

2.2 Chebyshev's Inequality

Markov's Inequality is the best bound you can have if all you know is the expectation. In its worst case, the probability is very spread out. The Chebyshev Inequality lets you say more if you know the distribution's variance.

Definition 2.2 (Variance)**.** *The variance of a random variable X is defined as*

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}X)^2\right]$$

Let us prove an identity on $\mathrm{Var}(X)$.

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}X)^2\right] \\ &= \mathbb{E}\left[X^2 - 2X\mathbb{E}[x] + (\mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[X^2\right] + (\mathbb{E}[X])^2 - 2(\mathbb{E}[X])^2 \\ &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \end{aligned}$$

where we used linearity of expectation. Note that for any number X, $(X - \mathbb{E}X)^2 \geq 0$. Therefore, for any random variable X, $\mathrm{Var}(X) \geq 0$. So, by above identity we always have

$$\mathbb{E}\left[X^2\right] \geq \mathbb{E}\left[X\right]^2,$$

i.e., the 2nd moment is at least the 1st moment squared.

Theorem 2.3 (Chebyshev's Inequality)**.** *For any random variable X,*

$$\mathbb{P}\left[|X - \mathbb{E}X| > \epsilon\right] < \frac{\mathrm{Var}(X)}{\epsilon^2}$$

or equivalently

$$\mathbb{P}\left[|X - \mathbb{E}\left[X\right]| > k\sigma\right] \leq \frac{1}{k^2}$$

where $\sigma = \sqrt{\mathrm{Var}(X)}$ is the standard deviation of X.

The second inequality in theorem can be read that any random variable is within 3 standard deviation of the expectation with probability 90%. It turns out that Chebyshev's inequality is just Markov's inequality applied to the variance R.V., $Y = (X - \mathbb{E}\left[X\right])^2$.

Proof. Let $Y := (X - \mathbb{E}X)^2$ be a nonnegative random variable. So, by Markov's inequality,

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

In other words,

$$\mathbb{P}\left[|X - \mathbb{E}\left[X\right]|^2 \geq \epsilon^2\right] \geq \frac{\mathrm{Var}(X)}{\epsilon^2}.$$

Taking square root of the both sides of the inequality gives,

$$\mathbb{P}\left[|X - \mathbb{E}\left[X\right]| \geq \epsilon\right] \geq \frac{\mathrm{Var}(X)}{\epsilon^2}$$

as desired □

2.2.1 Polling

In this section we use Chebyshev's inequality to answer the question that we raised at the beginning of this lecture. Suppose there is an unknown distribution D with mean μ and we want to estimate μ using independent samples of D,

$$X_1, X_2, \ldots, X_n$$

First, observe that by linearity of expectation,

$$\mathbb{E}\left[\frac{1}{n}\sum_i X_i\right] = \mu.$$

So, we want to use Chebyshev's inequality to upper bound,

$$\mathbb{P}\left[\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right]$$

To use Chebyshev's inequality, first we need to calculate the variance. Let $X = \frac{X_1+\cdots+X_n}{n}$ be the empirical average. We use the following lemma to bound the variance of X.

We say a set of random variables $X_1, X_2, \ldots, X_n$ are *pairwise independent* if for all $1 \le i, j \le n$

$$\mathbb{E}[X_i X_j] = \mathbb{E}[X_i]\,\mathbb{E}[X_j].$$

Lemma 2.4. *For any set of pairwise independent random variables $X_1, \ldots, X_n$*

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}\,X_1 + \cdots + \mathrm{Var}\,X_n$$

Proof. We can write,

$$\begin{aligned}
\mathrm{Var}(X_1 + \cdots + X_n) &= \mathbb{E}\left[(X_1 + \cdots + X_n)^2\right] - (\mathbb{E}X_1 + \mathbb{E}X_2 + \cdots + \mathbb{E}X_n)^2 \\
&= \mathbb{E}\left[\sum_{i,j} X_i X_j\right] - \sum_{i,j} \mathbb{E}[X_i]\,\mathbb{E}[X_j] \\
&= \sum \mathbb{E}\left[X_i^2\right] - (\mathbb{E}[X_i])^2 \\
&= \sum_{i=1}^{n} \mathrm{Var}(X_i).
\end{aligned}$$

In the second to last equality we used pairwise independence. □

Let's go back to the polling example; recall $X_1, \ldots, X_n$ are independent samples of D, so they are pairwise independent, and by the above lemma,

$$\mathrm{Var}(X) = \mathrm{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n^2}\mathrm{Var}(X_1 + \cdots + X_n) = \frac{1}{n^2}(\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)) = \frac{\mathrm{Var}(D)}{n}$$

Therefore, by Chebyshev's inequality,

$$\mathbb{P}\left[|X - \mu| \ge \epsilon\right] \le \frac{\mathrm{Var}(D)}{n\epsilon^2} \tag{2.1}$$

Now, let/'s continue on the polling example, suppose for all i,

$$X_i = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{otherwise} \end{cases},$$

i.e., p fraction of the population would vote yes on the election, and we want to estimate p within ϵ additive error. So, it all we need to do is to upper bound the variance of X_i, First, we calculate the second moment, for all i,

$$\mathbb{E}\left[X_i^2\right] = 1^2 \cdot p + 0^2 \cdot (1-p) = p.$$

Therefore,

$$\mathrm{Var}(X_i) = \mathbb{E}\left[X_i^2\right] - \mathbb{E}[X_i]^2 = p - p^2 = p(1-p) \le \frac{1}{4}.$$

Therefore, by (2.1)

$$\mathbb{P}\left[\left|\frac{\sum_i X_i}{n} - p\right| \ge \epsilon\right] \le \frac{1}{4n\epsilon^2}$$

Suppose we choose $10,000$ individuals from the population randomly and we calculate the empirical mean; by above inequality with probability $15/16$ our estimate is within 2% of the true mean. Note that the importance of this inequality is the the size of the sample is independent of the size of the population. In general if we want to obtain an ϵ-additive error with probability $1-\delta$ we need $O(1/\delta\epsilon^2)$ many samples.

Note that the above analysis can easily be extended to the case where X_i's are not necessarily Bernoulli. In particular, suppose D is distributed on an interval $[a,b]$ where D can take any real number in this interval. It fallows that the variance of D is at most $(b-a)^2$. This is because the different of any two numbers in the support of D is at most $b-a$. Therefore, following the same analysis if we have n samples $X_1,\dots,X_n$ of such a D then

$$\mathbb{P}\left[\left|\frac{\sum_i X_i}{n}-\mu\right|\geq\epsilon\right]\leq\frac{(b-a)^2}{n\epsilon^2}.$$

where μ is the mean of D. So, to get an ϵ-additive error with probability at least $1-\delta$ it is enough to have $n\geq\frac{(b-a)^2}{\epsilon^2\delta}$ many samples.

Next lecture we will see a stronger concentration bounds, a.k.a., Chernoff bounds. We see that for the same polling example it is enough to use $O(\frac{1}{\epsilon^2}\log\frac{1}{\delta})$ samples to obtain an ϵ-additive approximation of the mean with probability $1-\delta$.