Algorithms II October 28, 2025

Lecture 12: Sampling and Concentration Inequalities

Notes by Ola Svensson[1]

These notes are based on the lecture notes of Lecture 2 in Shayan Oveis Gharan's course "CSE 521: Design and Analysis of Algorithms I" available here:

http://courses.cs.washington.edu/courses/cse521/17wi/

- In this lecture we analyze the task of polling by learning important concentration inequalities: Markov's inequality, Chebychev's inequality, and Chernoff's inequality.

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 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: e.g. to estimate the percentage of the students (you) that have started preparing for the midterm.

We start this lecture by a simple example: Suppose that the average salary in Switzerland is 6000 CHF/month. What fraction of the working population that receives at most 8000 CHF/month? It turns out that always at least 1/4 of the workers receive at most 8000 CHF/month. In the worst case, $\frac{1}{4}$ of the workers receive 0 CHF/month, and $\frac{3}{4}$ get 8000. We can justify this claim using Markov's inequality.

1 Markov's Inequality

Theorem 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 salary example, X denotes the average salary, $\mathbb{E}\left[X\right] = 6000$ and $k = 4/3$. The inequality says at most 3/4 of the workers receive more than 8000 or at least 1/4 receive less than 8000.

Proof The proof is a simple one line argument,

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

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

[1] **Disclaimer:** These notes were written as notes for the lecturer. They have not been peer-reviewed and may contain inconsistent notation, typos, and omit citations of relevant works.

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

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

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 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. Chebyshev's Inequality lets you say more if you know the distribution's variance.

Definition 2 (Variance) *The variance of a random variable X is defined as*

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

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

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

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, $\text{Var}(X) \geq 0$. So, by above identity we always have

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

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

Theorem 3 (Chebyshev's Inequality) *For any random variable X,*

$$\mathbb{P}[|X - \mathbb{E}X| > \epsilon] \leq \frac{\text{Var}(X)}{\epsilon^2}$$

or equivalently

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

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

The second inequality in theorem can be read that any random variable is within 3 standard deviations of the expectation with probability $8/9 \geq 88\%$. It turns out that Chebyshev's inequality is just Markov's inequality applied to the variance R.V., $Y = (X - \mathbb{E}[X])^2$.

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

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

In other words,

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

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

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

as desired ■

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 $1 \leq i, j \leq n$

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

Lemma 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,

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

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,

$$\mathbb{P}\left[|X-\mu|\geq\epsilon\right]\leq\frac{\mathrm{Var}(D)}{n\epsilon^2}\tag{1}$$

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, 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,

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

Therefore, by (1)

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

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 that 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.

We now move on to stronger concentration bounds, a.k.a., Chernoff bounds that can applied when the random variables are mutually independent. We will 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 $1-\delta$.

3 Chernoff bounds

Chernoff bounds is a family of strong concentration inequalities named after the mathematician Herman Chernoff[2]. We start with some intuition before stating one of these inequalities formally. After that we return to our polling example. Finally, we give the sketch of a proof of the inequality.

3.1 Law of Large Numbers

The Law of Large Numbers (LLN) is a theorem which states that the average of the results obtained from a large number of independent trials of an experiment tends towards the expected value. Central limit theorems state that for an infintie sequence of random independent variables $X_1, X_2, \ldots$ with mean μ and unit variance.

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right)\to\mathcal{N}(0,1).\tag{2}$$

as n goes to infinity. In this course, we are interested in quantitative forms of this convergence. We will study this in the form of *strong concentration bounds*, a.k.a., Chernoff bounds.

[2] See *http://math.mit.edu/˜goemans/18310S15/chernoff-notes.pdf that also form a basis of these notes.* for more comprehensive notes of these bounds.

Recall that Chebyshev's inequality implies that for any random variable X,

$$\mathbb{P}\left[|X - \mathbb{E}\left[X\right]| \geq k\sigma\right] \leq \frac{1}{k^2} \tag{3}$$

Strong concentration bounds imply that if X is an average of independent random variables with standard deviation σ that satisfy certain other nice properties, then

$$\mathbb{P}\left[|X - \mathbb{E}X| \geq k\sigma\right] \leq e^{-\Omega(k^2)}$$

In other words, they give exponentially improved bounds compared to Chebyshev's inequality. Note that to get this strong bound we want X to be an average of mutually independent random variables; so unlike Chebyshev's inequality pairwise independent is not enough.

3.2 Formal statements

There are many different forms of Chernoff bounds, each tuned to slightly different assumptions. We will start with the statement of the bound for the simple case of a sum of independent Bernoulli trials, i.e., the case in which each random variable only takes the values 0 or 1. For example, the polling application is such an example.

Theorem 5 (Chernoff Bounds) *Let $X = \sum_{i=1}^n X_i$, where $X_i = 1$ with probability p_i and $X_i = 0$ with probability $1 - p_i$, and all X_i are independent. Let $\mu = \mathbb{E}\left[X\right] = \sum_{i=1}^n p_i$. Then*

(i) ***Upper Tail:*** $\mathbb{P}\left[X \geq (1+\delta)\mu)\right] \leq e^{-\frac{\delta^2}{2+\delta}\mu}$ *for all $\delta > 0$;*

(ii) ***Lower Tail:*** $\mathbb{P}\left[X \leq (1-\delta)\mu)\right] \leq e^{-\frac{\delta^2}{2}\mu}$ *for all $\delta > 0$.*

Notice that the lower and upper tail take slightly different forms. Curiously, this is necessary and boils down to the use of different approximation of the logarithmic function.

Before returning to our polling example, let us mention this very useful Chernoff Bound (often called Hoeffding's Inequality).

Theorem 6 *Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $a \leq X_i \leq b$ for all i. Let $X = \sum_{i=1}^n X_i$ and set $\mu = \mathbb{E}\left[X\right]$. Then*

(i) ***Upper Tail:*** $\mathbb{P}\left[X \geq (1+\delta)\mu)\right] \leq e^{-\frac{2\delta^2\mu^2}{n(b-a)^2}}$ *for all $\delta > 0$;*

(ii) ***Lower Tail:*** $\mathbb{P}\left[X \leq (1-\delta)\mu)\right] \leq e^{-\frac{\delta^2\mu^2}{n(b-a)^2}}$ *for all $\delta > 0$.*

Notice, that the above tail bounds apply to bounded random variables, regardless of their distribution!

3.3 Application: Polling

Let us continue our polling example: Consider a set of n Bernoulli random variables $X_1, X_2, \ldots X_n$ where for all i, $X_i = 1$ w.p. p and $X_i = 0$ w.p. $1 - p$. By Theorem 5,

$$\begin{aligned}
\mathbb{P}\left[\left|\frac{\sum X_i}{n} - p\right| \geq \epsilon\right] &= \mathbb{P}\left[\left|\sum X_i - pn\right| \geq n\epsilon\right] \\
&\leq \mathbb{P}\left[\sum X_i \geq (1+\epsilon/p)pn\right] + \mathbb{P}\left[\sum X_i \leq (1-\epsilon/p)pn\right] \\
&\leq e^{-\frac{\epsilon^2 n}{3}} + e^{-\frac{\epsilon^2 n}{2}}
\end{aligned}$$

So, if we want to estimate the probability p within an additive error ϵ with probability $1-\delta$ it is enough to let

$$n = 3\frac{\ln(2/\delta)}{\epsilon^2}.$$

To give you a point of comparison, recall that we showed that using Chebyshev inequality, to estimate p with additive error of ϵ with probability at $1-\delta$ we need about $1/(\delta\epsilon^2)$. So, for example, if we want $1-2^{-100}$ probability of success the Chernoff bound implies we only need about $100/\epsilon^2$ many samples, whereas Chebyshev's inequality says we want $2^{100}/\epsilon^2$ many samples. You can see that Chernoff bounds implies a significantly smaller number of samples.

Upshot: The failure probability decreases exponentially with respect to the number of samples whereas the confidence interval ϵ only decreases proportional to the square-root of the number of samples.

3.4 Proof sketch of Theorem 5

We show how to prove the upper tail bound. The proof for the lower tail is analogous. Perhaps surprisingly, we again resort to Markov's inequality (similar to the proof of Chebychev's inequality, however, here the calculations become a little more involved). For any $s > 0$,

$$\begin{aligned}\mathbb{P}\left[X \geq a\right] &= \mathbb{P}\left[e^{sX} \geq e^{sa}\right] \\ &\leq \frac{\mathbb{E}\left[e^{sX}\right]}{e^{sa}} \qquad \text{(by Markov's inequality).}\end{aligned}$$

We now analyze the numerator. By the independence of the random variables $X_1, \dots, X_n$

$$\mathbb{E}\left[e^{sX}\right] = \mathbb{E}\left[e^{s(X_1+X_2+\cdots+X_n)}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{sX_i}\right]$$

Now using that X_i is a Bernoulli random variable that takes value 1 with probability p_i:

$$\begin{aligned}\prod_{i=1}^{n} \mathbb{E}\left[e^{sX_i}\right] &= \prod_{i=1}^{n} (p_i \cdot e^s + (1-p_i)\cdot 1) \\ &= \prod_{i=1}^{n} 1 + p_i(e^s - 1) \\ &\leq \prod_{i=1}^{n} e^{p_i(e^s-1)} \qquad \text{(using that } 1+y \leq e^y \text{ with } y = p(e^s-1)\text{)} \\ &= e^{\mu(e^s-1)}\end{aligned}$$

Now setting $a = (1+\delta)\mu$ and $s = \ln(1+\delta)$ we get[3]

$$\mathbb{P}\left[X \geq a\right] \leq \frac{e^{\mu\delta}}{(1+\delta)^{\mu(1+\delta)}} = \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu},$$

which can be simplified to be at most $e^{-\frac{\delta^2}{2+\delta}\mu}$.

[3] Our choice of s is motivated as follows: we are trying to make our upper bound for the tail probability to be as small as possible. To do this, we can minimize our expression for the upper bound as a function of s.

The lower tail follows by similar calculations: For any $s > 0$,

$$\begin{aligned}\mathbb{P}\left[X \leq a\right] &= \mathbb{P}\left[e^{-sX} \geq e^{-sa}\right] \\ &\leq \frac{\mathbb{E}\left[e^{-sX}\right]}{e^{-sa}} \qquad \text{(by Markov's inequality).}\end{aligned}$$

Now same simplifications follow and at the end we choose $a = (1-\delta)\mu$ and $s = -\ln(1-\delta)$.