Princeton Univ. F'20 COS 521: Advanced Algorithm Design

Lecture 3: Concentration Bounds

Lecturer: *Sahil Singla* Last updated: September 7, 2020

1 Preliminaries

Today's topic is deviation bounds: what is the probability that a random variable deviates from its mean by a lot? Recall that a random variable X is a mapping from a set of possible outcomes S to $\mathbb{R}$. We usually think of S as containing numerical quantaties and for now, that just means scalar numbers. Later in the course, we will consider situations where S contains vectors or matrices. The *expectation* or *mean* is denoted $\mathbb{E}[X]$ or sometimes as μ.

$$\mathbb{E}[X] \overset{\text{def}}{=} \sum_{s \in S} \Pr[X = s] \cdot s$$

To give an example, consider a random variable X that corresponds to the toss of a fair coin. X maps the possible outcomes $\{0, 1\}$ each to $1/2$, taking value 0 when the coin lands on tails, and 1 when the coin lands on heads. In this case, $\mathbb{E}[X] = 1/2$. In many settings we have a set of n random variables $X_1, X_2, X_3, \ldots, X_n$ defined on the same set of possible outcomes. For example, each X_i might correspond to the toss of a different random coin.

In addition to the expectation, the *variance* of a random variable is defined as:

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

We will often use μ to denote $\mathbb{E}[X]$ and σ^2 to denote $\mathrm{Var}[X]$.

Here are examples of facts that you might remember from discrete math or other undergrad classes. We won't prove them all in class, but it might be a good refresher to re-derive them yourself or in office hours.

- For any random variables, independent or not, $\mathbb{E}[\sum_i X_i] = \sum_i \mathbb{E}[X_i]$. This is called the **Linearity of Expectation**.
- If X_1, X_2 are independent random variables (formally, this means that for all a, b $\Pr[X_1 = a, X_2 = b] = \Pr[X_1 = a]\Pr[X_2 = b]$), then $\mathbb{E}[X_1 \cdot X_2] = \mathbb{E}[X_1] \cdot \mathbb{E}[X_2]$.
- When we say a set of random variables $X_1, \ldots X_n$ are *mutually independent*, we mean that for all $a_1, \ldots, a_n$, $\Pr[X_1 = a_1, X_2 = a_2, \ldots X_n = a_n] = \prod_i \Pr[X_i = a_i]$.
- We say that $X_1, \ldots, X_n$ are *pairwise independent* random variables if for all X_i, X_j, X_i and X_j are independent, but the set of all variables are not necessarily mutually independent.
- If $X_1, \ldots, X_n$ are pairwise independent, then $\mathrm{Var}\left[\sum_i X_i\right] = \sum_i \mathrm{Var}[X_i]$.

Exercise: Give an example of three random variables that are not *mutually independent*, but are *pairwise independent.*

1.1 Three progressively stronger tail bounds

As we saw in the past two lectures, and will see again and again in this class, one of our main goals when analyzing randomized algorithms will be to understand when random variables *behave as expected.* In other words, with what probability do they fall close to their expectation?

Any bound of this form is called a *tail bound* or *concentration inequality.* Today we will see three methods that give progressively stronger bounds, but under progressively stronger assumptions. They are Markov's inequality, Chebyshev's inequality, and the Chernoff bound.

2 Markov's Inequality

The first of a number of inequalities presented today, **Markov's inequality** says that any *non-negative* random variable X satisfies

$$\Pr\left(X \geq k\mathbb{E}[X]\right) \leq \frac{1}{k}.$$

Note that this is just another way to write the trivial observation that $\mathbb{E}[X] \geq k \cdot \Pr[X \geq k]$.

Can we give any meaningful upperbound on $\Pr[X < c \cdot \mathbb{E}[X]]$ where $c < 1$, in other words the probability that X is a lot less than its expectation? In general we cannot.

Exercise: For any $c < 1, \delta < 1$, find a distribution where $\Pr[X < c\mathbb{E}[X]] = 1 - \delta)$. In other words, X is very often far below it's expectation.

However, if we know an upperbound on X then we can make such a statement. If $X \leq z$ then for any $c < 1$ we have:

$$\Pr(X \leq c\mathbb{E}[X]) \leq \frac{z - \mathbb{E}[x]}{z - c\mathbb{E}[x]}.$$

Sometimes this is also called an averaging argument.

Exercise: Prove this using Markov's inequality, but on a different random variable.

Example 1. *Suppose you took a lot of exams, each scored from* 1 *to* 100*. If your average score was* 90 *then in at least half the exams you scored at least* 80*.*

Markov's inequality can sometimes be useful for making quick deductions about random variables. It also applies to any non-negative random variable. Because arbitrary non-negative random variables can behave wildly, we shouldn't hope for a stronger claim to hold without making some reference to properties of the random variable. We now move on to Chebyshev's inequality, which makes use of the variance.

3 Chebyshev's Inequality

The *variance of a random variable* X is one measure (there are others too) of how "spread out" it is around its mean. The variance is defined as $\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] -$

$\mathbb{E}[X]^2$, and we often denote it by σ^2. The square root of the variance, σ, is called the *standard deviation.*

A more powerful inequality, **Chebyshev's inequality**, says

$$\Pr[|X - \mathbb{E}[X]| \geq k\sigma] \leq \frac{1}{k^2},$$

Actually, Chebyshev's inequality is just a special case of Markov's inequality: by definition,

$$\mathbb{E}\left[|X - \mathbb{E}[X]|^2\right] = \sigma^2,$$

and so,

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

3.1 Example: Load balancing

Recall our setup from Lecture 2. Suppose we have n values, $a_1, \ldots, a_n$, from some universe $|U|$ and we want to hash these values to a table of size n. This is often call the "balls-into-bins" problem because we can think about hashing as randomly throwing balls into bins, and seeing how many balls each bin has. It's convenient to first analyze the case when the number of balls equals the number of bins, although this isn't always the setup.

In the first lecture we weren't able to obtain bounds on the maximum load of a particular bin – we just showed that on average the bins weren't too overloaded. This could be done using Markov's inequality.

It turns out that we *can* get a bound on the maximum load using Chebyshev's inequality. Let's just consider the the first bin and how many balls fall into it. Let $X_i = \mathbb{1}[\text{ball } i \text{ falls into bin } 1]$. Assume that we are using a pairwise-independent hash function, so:

$$\mathbb{E}[X_i] = \frac{1}{n}.$$

What's the variance of X_i?

$$\mathrm{Var}[X_i] = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 = \frac{1}{n} - \frac{1}{n^2} \leq \frac{1}{n}.$$

Now, let $X = \sum_{i=1}^n X_i$. X is the total number of balls that land in bin 1 and $\mathbb{E}[X] = 1$.

What's the variance of X? Since each X_i, X_j are *pairwise independent*,

$$\mathrm{Var}[X] = \sum_{i=1}^n \mathrm{Var}[X_i] \leq 1.$$

From Chebyshev's inequality, we therefore have that:

$$\Pr[|X - 1| \geq \sqrt{2n}] \leq \frac{1}{2n}.$$

So bin 1 has load $\leq \sqrt{2n} + 1$ with probability at least $(1 - \frac{1}{2n})$, and this exact same bound holds for all other bins. Thus, by a **union bound**, every bin has load $\leq \sqrt{2n} + 1$ with probability 1/2. That's not bad! For $n = 1,000,000$, we can say that the maximally loaded bin has $\lesssim 1400$ elements. Shortly, we will see how to get an even tighter bound than $O(\sqrt{n})$.

3.2 Another common use

We won't give a specific example in class, but it is helpful to mention that Chebyshev's inequality can often be used to analyze how well an average of many random variables concentrates around its expectation. In particular, suppose $Y_1, Y_2, \ldots, Y_t$ are i.i.d. (independent and identically distributed) random variables, meaning that they have the same distribution. Suppose each has variance σ_2. Then:

$$\mathrm{Var}\left(\frac{1}{t}\sum_i Y_i\right) = \frac{\sigma^2}{t}.$$

In other words, even if each Y_i does not concentrate close to its mean, taking an average quickly improves our variance and gives better concentration via Chebyshev's inequality.

4 Chernoff bounds

4.1 Motivation

How tight is Chebyshev's inequality? I suspect many of you have seen this picture before:

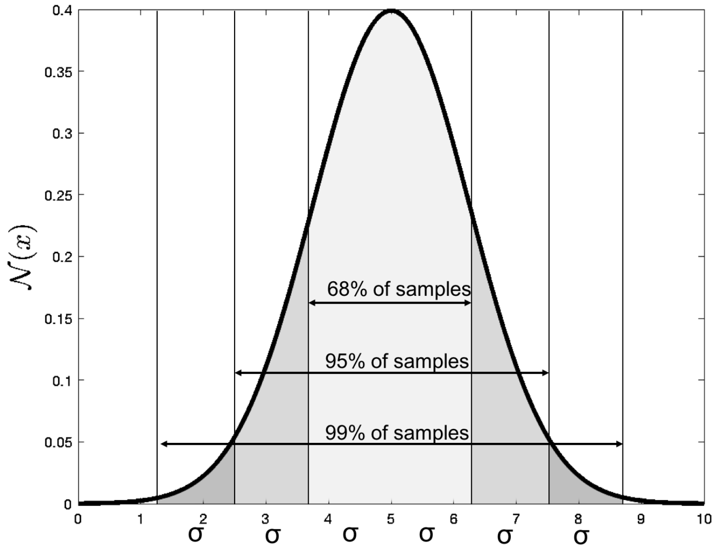


Figure 1: 68-95-99 rule for Gaussian bell-curve.

If X is distributed as a normal random variable, aka a Gaussian, aka a *Bell Curve*, and it has standard deviation σ (i.e. variance σ^2), then it is well known that:

$$\begin{aligned}
\Pr\left(|X - \mathbb{E}[X]| \geq 1\sigma\right) &\approx 32\% \\
\Pr\left(|X - \mathbb{E}[X]| \geq 2\sigma\right) &\approx 5\% \\
\Pr\left(|X - \mathbb{E}[X]| \geq 3\sigma\right) &\approx 1\% \\
\Pr\left(|X - \mathbb{E}[X]| \geq 4\sigma\right) &\approx .01\%
\end{aligned}$$

On the other hand, Chebyshev inequality would predict upper bounds of:

$$\begin{aligned}
&\Pr\left(|X - \mathbb{E}[X]| \geq 1\sigma\right) \leq 100\% \\
&\Pr\left(|X - \mathbb{E}[X]| \geq 2\sigma\right) \leq 25\% \\
&\Pr\left(|X - \mathbb{E}[X]| \geq 3\sigma\right) \leq 11\% \\
&\Pr\left(|X - \mathbb{E}[X]| \geq 4\sigma\right) \leq 6\%.
\end{aligned}$$

It appears that, at least for the common Gaussian distribution, we can obtain much stronger concentration bounds: the chance of landing outside a given number of standard deviations falls off very fast. This makes sense if we look at the probability density function, $\mathcal{N}$, of the Gaussian distribution:

$$\mathcal{N}(x) \sim e^{-x^2/2\sigma^2}.$$

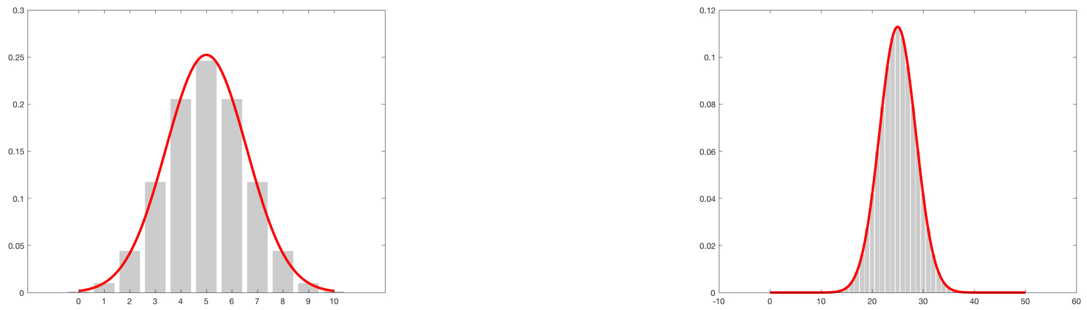
The distribution is falling off exponentially in x/σ.

Exercise: For Gaussian X with variance σ^2, show that $\Pr\left(|X - \mathbb{E}x| \geq c\sigma\right) \leq O(e^{-c^2/2})$.

Why are bounds for Gaussian random variables important in algorithm design?

The *Central Limit Theorem* says that the sum of n independent random variables (with bounded mean and variance) converges to the Gaussian distribution, even if those random variables themselves aren't Gaussian. For many random variables that appear in randomized algorithms, this convergence happens very quickly, meaning that we can analyze the sum by treating it as a Gaussian random variable.

A well known example is coin tossing. Let $X = \sum_{i=1}^{n} X_i$ be a random variable which is the sum of n random variables, $X_1, \ldots, X_n$, each being 1 with probability $1/2$ and 0 otherwise. X represents the number of heads that will appear when flipping n fair coins. It is possible to explicitly compute the distribution of X. As we see in Figure 2, this distribution quickly begins to look like a Gaussian distribution as n increases.



(a) Distribution of # of heads after 10 coin flips, compared to a Gaussian.

(b) Distribution of # of heads after 50 coin flips, compared to a Gaussian.

Figure 2: The distribution of the number of heads in a sequence of n coin tosses quickly converges to a Gaussian distribution, as predicted by the Central Limit Theorem.

This concentration to a Gaussian implies that we can get much better bounds on, e.g. coin tossing processes, than we would via Chebyshev's inequality. To do a back of the envelope calculation, if we flip n coins and all n coin tosses are fair (heads has probability $1/2$) then the Gaussian approximation has mean $n/2$ and variance $n/4$. Let X be the number of heads we see. We can bound $\Pr(|X - n/2| \geq k\sigma) \leq e^{-k^2/2}$. $\sigma = O(\sqrt{n})$, so if we want to be within ϵn of $n/2$, we need to set $k = \epsilon\sqrt{n}$.

How large do we need to set n to achieve this bound with probability $1/2$? We need $n = O(1/\epsilon^2)$. How about with probability $1/n^{10}$? We need $n = O(\log(n)/\epsilon^2)$. In other words, we pay very little to achieve much higher probability estimates. To give a real number example, if we flip 1000 coins, the chance of seeing at least 625 heads is less than 5.3×10^{-7}. These are pretty strong bounds!

4.2 Main Theorem

Of course, for finite n, the sum of n random variables is not necessarily *exactly* a Gaussian. That's where Chernoff bounds come in. They help us quantify this potentially very powerful Gaussian approximation. It turns out that the CLT converges pretty quickly for sums of *bounded* random variables, including binary variables like coins, that we can obtain tail bounds nearly identical to what we get for a true Gaussian. Any bound of this type we informally call a "Chernoff bound".

There are many forms of Chernoff bounds, often under various other names (Chernoff bound, Bernstein inequality, Hoeffding inequality, etc.). One particularly useful one applies to random variables bounded between $[-1, 1]$. To apply it to more general bounded variables, just scale them to $[-1, 1]$ first.

Theorem 1 (Quantitative version of CLT due to S. Bernstein)**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables and each $X_i \in [-1, 1]$. Let $\mu_i = \mathbb{E}[X_i]$ and $\sigma_i^2 = var[X_i]$. Then $X = \sum_i X_i$ satisfies*

$$\Pr[|X - \mu| > k\sigma] \leq 2\exp(-\frac{k^2}{4}),$$

where $\mu = \sum_i \mu_i$, variance $\sigma^2 = \sum_i \sigma_i^2$, and $k \leq \frac{1}{2}\sigma$.

This theorem is usually called the Bernstein inequality.

4.3 Simple Application: Coins and statistical polling

Suppose we flip n fair coins again. Let X be the number of heads we see. We can use the above theorem to formally bound $\Pr(|X - n/2| \geq \epsilon n) \leq \delta$ as long as $n = O(\log(1/\delta)/\epsilon^2)$. In other words, if we want to test whether or not a coin is within ϵ of fair (i.e. it is heads and tails, each with probability $> 1/2 - \epsilon$), then we can do so by averaging $O(\log(1/\delta)/\epsilon^2)$, and our test will only fail with probability δ.

Exercise: Show that Chebyshev's inequality would predict that the same fairness test requires $O(\frac{1}{\epsilon^2\delta^2})$ – i.e. it gives an exponentially worse dependence on δ!

More generally, opinion polls and statistical sampling rely on tail bounds. Suppose there are n arbitrary numbers in $[0, 1]$. If we pick t of them randomly with replacement then the sample mean is within an additive ϵ of the true mean with probability at least $1-\delta$ if $t > \Omega(\frac{1}{\epsilon^2}\log 1/\delta)$.

4.4 Proof

Instead of proving Theorem 1, we prove a simpler theorem for binary valued variables which showcases the basic idea. We'll give a complete proof of this bound, which will be enough to prove a pretty powerful hashing application.

Theorem 2. *Let $X_1, X_2, \ldots, X_n$ be independent $0/1$-valued random variables and let $p_i = \mathbb{E}[X_i]$, where $0 < p_i < 1$. Then the sum $X = \sum_{i=1}^n X_i$, which has mean $\mu = \sum_{i=1}^n p_i$, satisfies*

$$\Pr[X \geq (1+\epsilon)\mu] \leq e^{\frac{-\epsilon^2\mu}{3+3\epsilon}}.$$

Remark: It's actually possible to prove a slightly tighter bound where the right hand side is $e^{\frac{-\epsilon^2\mu}{2+\epsilon}}$. Additionally, there is an analogous inequality that bounds the probability of deviation *below* the mean, $\Pr[X \leq (1-\epsilon)\mu]$. For that bound, the right hand side becomes $e^{\frac{-\epsilon^2\mu}{2}}$ On homeworks, you're free to use any versions of Chernoff bounds that you find in other course notes (or Wikipedia). There are many variants.

Proof. Surprisingly, this inequality also is proved using the Markov inequality, albeit applied to a different random variable.

We introduce a positive dummy variable t that we will set to some non-negative value later. We observe that

$$\mathbb{E}[e^{tX}] = \mathbb{E}[e^{t\sum_i X_i}] = \mathbb{E}[\prod_i e^{tX_i}] = \prod_i \mathbb{E}[e^{tX_i}], \tag{1}$$

where the last equality holds because the X_i random variables are *mutually independent.* Now, because each X_i is $0/1$, we have that:

$$\mathbb{E}[e^{tX_i}] = (1-p_i) + p_i e^t.$$

Therefore,

$$\begin{aligned}\prod_i \mathbb{E}[e^{tX_i}] = \prod_i [1 + p_i(e^t - 1)] \leq \prod_i e^{p_i(e^t-1)} \\ = e^{\sum_i p_i(e^t-1)} = e^{\mu(e^t-1)}.\end{aligned} \tag{2}$$

In the step with an inequality, we used that $1 + x \leq e^x$. (This holds for all x – it's a surprisingly useful inequality to remember.) Finally, apply Markov's inequality to the random variable e^{tX}:

$$\Pr[X \geq (1+\epsilon)\mu] = \Pr[e^{tX} \geq e^{t(1+\epsilon)\mu}] \leq \frac{\mathbb{E}[e^{tX}]}{e^{t(1+\epsilon)\mu}} \leq \frac{e^{(e^t-1)\mu}}{e^{t(1+\epsilon)\mu}},$$

using lines (1) and (2) and the fact that t is positive. Since the statement holds for *any* t, we can obtain a bound by setting t to any positive value we wish. If we set $t = \log(1+\epsilon)$, we get:

$$\Pr[X \geq (1+\epsilon)\mu] \leq e^{\mu[\epsilon - \log(1+\epsilon)(1+\epsilon)]}.$$

To see that this bound simplifies to give Theorem 2, we need a quick case argument. Looking at the Taylor series of $\log(1+\epsilon)$, we have:

$$\log(1+\epsilon) = \epsilon - \frac{\epsilon^2}{2} + \frac{\epsilon^3}{3} - \frac{\epsilon^4}{4} + \cdots$$

and

$$\log(1+\epsilon)(1+\epsilon) = \epsilon + \frac{\epsilon^2}{2} - \frac{\epsilon^3}{6} + \frac{\epsilon^4}{20} - \cdots$$

For $\epsilon \in [0,1]$, we thus have $\log(1+\epsilon)(1+\epsilon) \geq \epsilon + \epsilon^2/3$. It follows that $e^{\mu[\epsilon - \log(1+\epsilon)(1+\epsilon)]} \leq e^{-\mu\epsilon^2/3} \leq e^{-\mu\epsilon^2/(3+3\epsilon)}$. On the other hand, when $\epsilon > 1$, $\log(1+\epsilon)(1+\epsilon) \geq 1.38\epsilon$. It follow that $e^{\mu[\epsilon - \log(1+\epsilon)(1+\epsilon)]} \leq e^{-.38\mu\epsilon} \leq e^{-\mu\epsilon^2/3\epsilon} \leq e^{-\mu\epsilon^2/(3+3\epsilon)}$. □

5 Load balancing revisited

With our Chernoff bound pf Theorem 2 in place, let's revisit our "balls-in-bins" analysis. Using a Chebyshev bound, we were able to bound the max load of n bins after inserting n balls by $O(\sqrt{n})$. The Chernoff bound will do exponentially better.

Again, we will analyze things one bin at a time. Let $X_i = \mathbb{1}[\text{ball } i \text{ falls into bin } 1]$ and let $X = \sum_{i=1}^{n} X_i$. $\mu = \mathbb{E}X = 1$. To apply Chernoff we will assume fully random hash functions[1]. Since $\mu = 1$, from Theorem 2, we have that:

$$\Pr[X \geq 1 + 6\log n] \leq e^{-6\log n/3} \leq \frac{1}{n^2}.$$

So bin 1 gets $\leq 1 + 6\log n$ balls with probability at least $(1 - 1/n^2)$. By a union bound, we conclude that *all bins* have $\leq 1 + 6\log n$ with probability $1 - 1/n$.

This bound of $O(\log n)$ on the maximum load of any bin improves exponentially on our bound of $O(\sqrt{n})$ from Chebyshev. Moreover, it holds with much higher probability. In fact, we could have succeeded with probability $(1 - 1/n^c)$ for any constant c if we just increase the constant factor on $6\log n$ a bit.

5.1 Power of Two Choices

The above $O(\log n)$ bound is very good, but it turns out that a simple alternative hashing scheme can do *even better*. Consider the method you use at the supermarket checkout:

[1] There was a question about this in class. It's actually possible to prove Chernoff bounds using $O(\log n)$-wise independence, which is much better than full independence, but not as simple as the 2-wise independence we assume for our Chebyshev bound. See recent work on improving over $O(\log n)$ independence in [1] or even more recent work considering "power of two choices" like methods [2].

instead of going to a random checkout counter you try to go to the counter with the shortest line. In the hashing setting this is computationally too expensive: one has to check all n queues. A much simpler version is the following: when the ball comes in, pick 2 random bins, and place the ball in the one that has fewer balls. It turns out that this modified rule ensures that the maximal load drops to $O(\log\log n)$, which is a huge improvement. This called the *power of two choices* and was first proven in the conference version of [3].

How about 3 choices? 4? d? Surprisingly there's not much to be gained after 2. The bound only improves to $O(\log\log n/\log d)$ for d choices.

References

[1] L. Elisa Celis, Omer Reingold, Gil Segev, and Udi Wieder. Balls and bins: Smaller hash families and faster evaluation. SIAM J. Comput., 42(3):1030–1050, 2013.

[2] Xue Chen. Derandomized Balanced Allocation. Preprint, 2017. https://arxiv.org/abs/1702.03375

[3] Yossi Azar, Andrei Z. Broder, Anna R. Karlin, and Eli Upfal. Balanced Allocations. SIAM J. Comput. 29, 1, 1999.