PRINCETON COS 521: ADVANCED ALGORITHM DESIGN

Basic Large Deviation Bounds

TODAY we recall elementary probability and some useful large deviation bounds with some applications. We will use this frequently in this course.

Recall that a (real-valued) random variable $X : S \to \mathbb{R}$ is function that maps a sample space S (i.e., outcomes of a random "experiment") into real numbers $\mathbb{R}$. The *expectation* or *mean* (think average) is denoted $\mathbb{E}[X]$ or sometimes as μ:

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

For example, when you toss a fair coin, the sample space has two outcomes: "heads" or "tails." A natural random variable in this context is the "indicator" of "heads". This random variable maps "heads" into 1 and "tails" into 0. In this case, note that $\mathbb{E}[X] = 1/2$.

As another example, consider the random "experiment" where you toss n fair coins. There is a natural set of n random variables $X_1, X_2, X_3, \ldots, X_n$ one for each indicator of a given coin landing on heads.

We are often interested in understanding how a random variable deviates from its mean. There are various quantitative measures that we will use to assess this deviation. Perhaps the most basic one is *variance* defined by:

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

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

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 call 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 $X_1, \dots X_n$ are *mutually independent*, we mean that for all $a_1, \dots, a_n$, $\Pr[X_1 = a_1, X_2 = a_2, \dots X_n = a_n] = \prod_i \Pr[X_i = a_i]$.
- We say that $X_1, \dots, X_n$ 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 $X_1, \dots, X_n$ are pairwise independent, then $\mathbf{Var}\left[\sum_i X_i\right] = \sum_i \mathbf{Var}[X_i]$.

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

Three progressively stronger tail bounds

As discussed above, we are often interested in understanding the probability that a random variable deviates from its mean. A practical scale to measure such deviation is the number of standard deviations (this, e.g., allows you to compare deviations of random variables that may take values at different “scales"). We will typically be interested in *large deviations* – the chance that a random variable takes a value that is $\gg \sigma$ away from its mean.

Inequalities that bound such probabilities are often called *tail bounds* or *concentration inequalities*.

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 $\mathbb{E}[X] \geq k \cdot \Pr[X \geq k]$.

Equivalently,

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

Can we give any meaningful upper bound 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 often far below its expectation.

However, if we know an upper bound 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 by applying Markov's inequality on a different random variable.

Here's an application:

Example 1. *Suppose you took many exams, each scoring 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.

Chebyshev's Inequality

The *variance of a random variable* X is one measure (among many others) of how "spread out" it is around its mean. The variance is defined as $\mathbf{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, and we often denote it by σ^2. The square root of the variance, σ, is called the *standard deviation*.

Here's **Chebyshev's inequality**:

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

Chebyshev's inequality is obtained by simply applying Markov's inequality to the non-negative random variable $(X - \mathbb{E}X)^2$.

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

and so,

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

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:

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

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.

Chernoff bounds

We will now see a helpful inequality with a much stronger bound on the large deviation probability. In general, such inequalities follow the following maxim:
Random variables that are simple functions of sufficiently independent random variables must be tightly concentrated around the mean.

We will see a very special case of this principle in action. Our simple function would simply be the sum (or average) of mutually independent random variables. But the general principle (and in fact even the proof techniques to an extent) apply to more complicated scenarios where the constituent random variables are only k-wise independent for some large k instead of mutually independent and the "simple" function is a polynomial in the constituent random variables instead of just being the sum. Later on in this class, we will see a generalization of an inequality to *matrix-valued random variables* that will also be extremely useful. But first, let's discuss the tail bounds in relation to Chebyshev's inequality a bit further.

Discussion

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

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

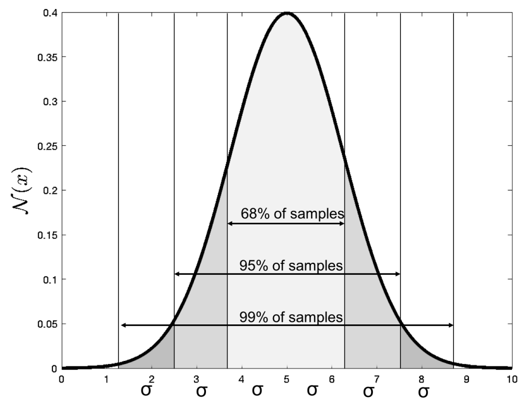


Figure 1: 68-95-99 rule for Gaussian bell-curve.

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 off 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:

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

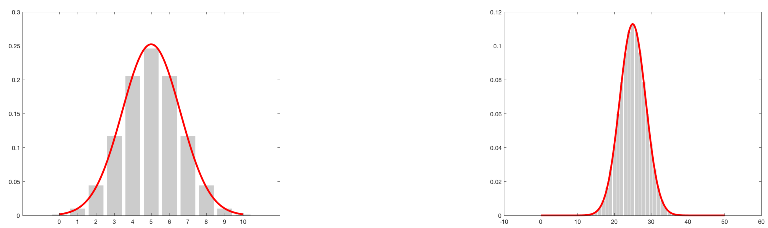
The distribution is falling off exponentially in x/σ.

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!

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 *in the large deviation/tail setting*.

There are many forms of such inequalities, 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 (Bernstein's Inequality)**.** *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*

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

where $\mu = \sum_i \mu_i$ *and* $\sigma^2 = \sum_i \sigma_i^2$. *Also,* $k \le \frac{1}{2}\sigma$.

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 $\Pr(|X - n/2| \ge \epsilon n) \le \delta$ 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 $t > \Omega(\frac{1}{\epsilon^2}\log 1/\delta)$.

Proof

Instead of proving Theorem 1, we prove a version for binary valued variables that 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*

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

Exercise: Find an example setting of parameters p_i where the above theorem gives a stronger bound on the large deviation probability compared to Bernstein's inequality.
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, $\Pr[X \le (1-\epsilon)\mu]$. 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,

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

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}:

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

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 $e^{\mu[\epsilon - \log(1+\epsilon)(1+\epsilon)]} \leq e^{-.38\mu\epsilon} \leq e^{-\mu\epsilon^2/3\epsilon} \leq e^{-\mu\epsilon^2/(3+3\epsilon)}$. □

Bibliography