**Mathematical Toolkit** **Autumn 2023**

# Lecture 15: November 15, 2023

Lecturer: Madhur Tulsiani

## 1 Chernoff/Hoeffding Bounds

Let's recall the bounds we proved in the previous lecture for sums of independent Bernoulli random variables.

**Theorem 1.1** *Let $X_1, ..., X_n$, be $n$* independent *Bernoulli random variables, where $X_i$ takes value 1 with probability $p_i$. Let $Z = \sum_{i=1}^n X_i$ and let $\mu = \mathbb{E}[Z] = \sum_{i=1}^n p_i$. Then, we have for any $\delta > 0$,*

$$\begin{aligned} \mathbb{P}[Z \geq (1+\delta)\mu] &\leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} \\ \mathbb{P}[Z \leq (1-\delta)\mu] &\leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu}. \end{aligned}$$

*Moreover, when $\delta \in (0,1)$ both the above expressions can be bounded by $e^{-\delta^2\mu/3}$.*

### 1.1 Coin tosses once more

We will now compare the above bound with what we can get from Chebyshev's inequality. Let's assume that $X_1, ..., X_n$ are independent coin tosses, with $\mathbb{P}[X_i = 1] = \frac{1}{2}$. We want to get a bound on the value of $Z = \sum_{i=1}^n X_i$. Using Chebyshev's inequality, we get that

$$\mathbb{P}[|Z - \mu| \geq \delta\mu] \leq \frac{\mathsf{Var}[Z]}{\delta^2\mu^2}.$$

And since in this particular case we have that $\mathsf{Var}[Z] = n/4$ and $\mu = n/2$, we get that

$$\mathbb{P}[|Z - \mu| \geq \delta\mu] \leq \frac{1}{\delta^2 n}.$$

The above bound is only inversely polynomial in $n$, while the Chernoff-Hoeffding bound gives

$$\mathbb{P}[|Z - \mu| \geq \delta\mu] \leq 2 \cdot \exp\left(-\delta^2 n/24\right),$$

which is exponentially small in $n$. This fact will prove very useful when taking a union bound over a large collection of events, each of which may be bounded using a Chernoff-Hoeffding bound.

Let us also compare the bound we get for a deviation which is comparable to the standard deviation (square root of the variance) of the the random variable $Z$. Consider the probablility $\mathbb{P}\left[\left|Z - \frac{n}{2}\right| \geq k\sqrt{n}\right]$. By Chebyshev's inequality, this can be bounded as

$$\mathbb{P}\left[\left|Z - \frac{n}{2}\right| \geq k\sqrt{n}\right] \;=\; \mathbb{P}\left[|Z - \mu| \geq k\sqrt{n}\right] \;\leq\; \frac{\mathsf{Var}\left[Z\right]}{k^2 \cdot n} \;=\; \frac{1}{4k^2}\,.$$

On the other hand, using the above version of Chernoff-Hoeffding bounds with $\delta = 2k/\sqrt{n}$ gives

$$\mathbb{P}\left[\left|Z - \frac{n}{2}\right| \geq k\sqrt{n}\right] \;=\; \mathbb{P}\left[\left|Z - \frac{n}{2}\right| \geq \frac{2k}{\sqrt{n}} \cdot \frac{n}{2}\right] \;\leq\; 2\exp\left(-2k^2/3\right)\,.$$

Which gives a much stronger dependence on $k$ which is (up to a factor 2) the number of standard deviations we are far from the mean. In general, tail probabilities which decrease as $\exp(-\Omega(k^2))$ are referred to as "sub-gaussian" tails, and we will soon discuss Gaussian random variables which are the prototypical example of such behavior.

## 1.2 Union bounds

Consider the case where for $m$ sets $S_1, \ldots, S_m \subseteq [n]$, we define

$$Z_{S_i} \;=\; \sum_{j \in S_i} X_j\,.$$

While the variables $Z_{S_1}, \ldots, Z_{S_m}$ are *not* necessarily indepdent, each of these is a sum of few $X_j$ variables, which are indepdent. Thus, we can say that for any $S_i$,

$$\mathbb{P}\left[\left|Z_{S_i} - \frac{|S_i|}{2}\right| \;\geq\; t\right] \;\leq\; 2 \cdot \exp\left(-2t^2/\left(3\,|S_i|\right)\right) \;\leq\; 2 \cdot \exp\left(-2t^2/(3n)\right)\,,$$

where we choose $\delta = 2t/\,|S_i|$ so that $\delta\,|S_i|\,/2 = t$. Thus, by a union bound over all $i \in [m]$, we get that

$$\mathbb{P}\left[\exists i \in [m].\; \left|Z_{S_i} - \frac{|S_i|}{2}\right| \;\geq\; t\right] \;\leq\; 2m \cdot \exp\left(-2t^2/(3n)\right)\,.$$

Thus, when $t = \sqrt{3n \cdot \ln m}$, the probability of the above event is at most $2/m$. Check that it just using Chebyshev's inequality does not allow for such a strong bound on the probability of the above event.

Note that the above calculation used the following union bound

**Exercise 1.2** *Let $E_1, \ldots, E_k$ be events on the same outcome space $\Omega$. Then*

$$\mathbb{P}\left[E_1 \cup \cdots \cup E_k\right] \;\leq\; \sum_{i=1}^{k} \mathbb{P}\left[E_i\right] .$$

### 1.3 Dealing with $\pm 1$ random variables

A common variant of the above calculations also arises for the case of random variables which take values in the set $\{-1, 1\}$ instead of the set $\{0, 1\}$. Let $Y_1, \ldots, Y_n$ be independent random variables, which take values in the set $\{-1, 1\}$ with probability $1/2$ each (such random variables are called Rademacher random variables), and let $Z = \sum_{i=1}^{n} Y_i$. We can easily apply the results for Bernoulli random variables to this case by defining $X_i = (1 + Y_i)/2$. Note that the variables $X_1, \ldots, X_n$ are now independent Bernoulli random variables (with parameter $1/2$). Considering $Z' = \sum_{i=1}^{n} X_i$, we can write

$$Z' \;=\; \sum_{i=1}^{n} X_i \;=\; \sum_{i=1}^{n} \frac{1 + Y_i}{2} \;=\; \frac{n}{2} + \frac{Z}{2} .$$

We can thus analyze deviations from the mean (which is 0) for the variable $Z$ as

$$\mathbb{P}\left[|Z| \geq t\right] \;=\; \mathbb{P}\left[\left|Z' - \frac{n}{2}\right| \geq \frac{t}{2}\right] ,$$

where we can analyze the latter expression using the bounds developed above.

## 2 Balanced Allocations

We consider the following problem of allocating jobs to servers: We are given a set of $n$ servers $1, \ldots, n$ and $m$ jobs arrive one by one. We seek to assign each job to one of the servers so that the loads placed on all servers are as balanced as possible.

In developing simple, effective load balancing algorithms, randomization often proves to be a useful tool. We consider two approaches for this problem:

- **Random Choice**: one possible way to distribute the jobs is to simply place each job on a random server, chosen independently of the previous allocations.
- **Two Random Choices**: For each job, we choose two servers independently and uniformly at random and place the job on the server with less load (breaking ties arbitrarily).

We will show that using two random choices significantly reduces the maximum load on any server. For the entire analysis, we will work with the case when $m = n$. The analysis easily extends to an arbitrary $m$, but it easier to appreciate the bounds when $m = O(n)$ (and in particular when $m = n$).

It is convenient to think of the above in terms of the so called "balls and bins" model. Each job can be thought of a s ball and each server is a bin. We think of assigning job $j$ to a server $i$ as throwing the $j^{th}$ ball in bin $i$. The load of a server is the same as the number of balls in the corresponding bin.

## 2.1 Random choice

Suppose $Z_i =$ number of balls in bin $i$. We can write

$$Z_i = \sum_j X_{ij}, \quad \text{where} \quad X_{ij} = \begin{cases} 1 & \text{if ball } j \text{ is thrown in bin } i \\ 0 & \text{otherwise} \end{cases} .$$

Then, we have that each $Z_i$ is a sum of $m(= n)$ independent random variables with $\mathbb{E}[Z_i] = 1$. Let $t = \frac{3 \ln n}{\ln \ln n}$. By Chernoff/Hoeffding bounds, we have that for each $i$,

$$\mathbb{P}[Z_i \geq t] \leq \left(\frac{e}{t}\right)^t .$$

To bound the maximum load in across all bins, we use a union bound to say that

$$\mathbb{P}[\exists i \in [n].\, Z_i \geq t] \leq \sum_{i=1}^{n} \mathbb{P}[Z_i \geq t] \leq n \cdot \left(\frac{e}{t}\right)^t ,$$

which is at most $\frac{1}{n}$ for the above value of $K$. Hence, with probability at least $1 - \frac{1}{n}$, the maximum number of balls in a bin is at most $\frac{3 \ln n}{\ln \ln n}$.

## 2.2 The power of two random choices

It is a somehwat surprising result (which can still be proved using Chernoff bounds) that two random choices can reduce the maximum load to $O(\ln \ln n)$. The proof technique is due to Azar et al. [ABKU94, ABKU99] and various applications were explored by Mitzenmacher in his thesis [Mit96]. We will not discuss the proof of this result, but you are encouraged to look up the analysis from the notes in 2016 (or from the book by Mitzenmacher and Upfal).

# 3 Probability over (uncountably) infinite probability spaces

Extending the idea of defining a probability *for each outcome* becomes problematic, when we try to extend it to uncountably infinite spaces. For example, let $\Omega = [0,1]$. Let $\nu : [0,1] \to [0,1]$ be a function, which we want to think of as a probability distribution. Define the set

$$S_n \;=\; \left\{ x \in [0,1] \mid \nu(x) \geq \frac{1}{n} \right\}.$$

Since we want the total probability to add up to 1, we must have $|S_n| \leq n$. Also,

$$\text{Supp}(\nu) \;=\; \{x \in [0,1] \mid \nu(x) > 0\} \;\subseteq\; \cup_{n=1}^{\infty} S_n \,.$$

Since $\cup_{n=1}^{\infty} S_n$ is a countable set, $\nu(x) > 0$ only for countably many points $x$. Hence, it is problematic to think of the probability of the outcome $x$, for each $x \in [0,1]$. This can be resolved by only talking of probabilities of *events* for an allowed set of events obeying some nice properties. Such a set is known as a $\sigma$-algebra or a $\sigma$-field.

**Definition 3.1** *Let $2^\Omega$ denote the set of all subsets of $\Omega$. A set $\mathcal{F} \subseteq 2^\Omega$ is called a* $\sigma$-field *(or $\sigma$-algebra) if*

1. $\emptyset \in \mathcal{F}$.
2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$ *(where $A^c = \Omega \setminus A$).*
3. *For a (countable) sequence $A_1, A_2, \ldots$ such that each $A_i \in \mathcal{F}$, we have $\cup_i A_i \in \mathcal{F}$.*

We then think of the sets in $\mathcal{F}$ as the allowed events. We can now define probabilities as follows.

**Definition 3.2** *Given a $\sigma$-field $\mathcal{F} \subseteq 2^\Omega$, a function $\nu : \mathcal{F} \to [0,1]$ is known as a* probability measure *if*

1. $\nu(\emptyset) = 0$.
2. $\nu(E^c) = 1 - \nu(E)$ *for all $E \in \mathcal{F}$.*
3. *For a (countable) sequence of disjoint sets $E_1, E_2, \ldots$ such that all $E_i \in \mathcal{F}$, we have*

$$\nu\left(\cup_i E_i\right) \;=\; \sum_i \nu(E_i)\,.$$

Note that the above definition do not say anything about unions of an uncountably infinite collection of sets. We can of course define probability measures on $\mathcal{F} = 2^\Omega$ and hence define $\nu(x)$ for all $x \in \Omega$. However, as we saw above, such measures will only have $\nu(x) > 0$ countably many $x$. Consider the following example.

**Example 3.3** *Let $\Omega = [0,1]$ and $\mathcal{F} = 2^\Omega$. Let $T = \left\{0, \frac{1}{3}, \frac{2}{3}, 1\right\}$. For each $S \in \mathcal{F}$, define*

$$\nu(S) \;=\; \frac{|S \cap T|}{4}.$$

In the above example, $\nu(x) > 0$ only for the points in a finite set $T$, which is very restrictive. We would like to formalize intuitive notions like the "uniform distribution" on the space $\Omega = [0,1]$: a probability measure that satisfies $\nu([a,b]) = b - a$ for $a, b \in [0,1]$ or more generally, for any event $E$ and a circular shift $E \oplus x$ for $x \in [0,1]$, we want $\nu(E) = \nu(E + x)$. It is a non-trivial result that such a probability measure indeed exists. This probability measure is known as the *Lebesgue* measure and is defined over a $\sigma$-algebra known as the *Borel* $\sigma$-algebra. The Borel $\sigma$-algebra does not contain all subsets of $[0,1]$ but does contain all intervals $[a,b]$. In fact, one can use the axiom of choice to show that we *cannot* include all subsets. The reason is that countable unions of very "thin" disjoint sets can reconstruct a "thick" set.

**Proposition 3.4** *Let $\Omega = [0,1]$. A measure satisfying the requirement that $\nu(E) = \nu(E + x)$ for all $E \in \mathcal{F}$ cannot be defined over the $\sigma$-algebra $\mathcal{F} = 2^\Omega$.*

**Proof:** For the sake of contradiction, assume that such a measure exists. Let $\mathcal{B}$ be the set of numbers in $[0,1)$ with a finite binary expansion, and define the equivalence relation between points $x, y \in [0,1]$:

$$x \sim y \qquad \text{if } \exists b \in \mathcal{B} \text{ such that } x = y \oplus b.$$

Thus $x$ and $y$ are equivalent if we can change only finitely many of the binary expansion of one, to get the other. Let $[x]$ denote one such equivalence class. Note that since there are countably many elements in $\mathcal{B}$, $[x]$ is also countable. In particular, $[0] = \mathcal{B}$. Because an equivalence defines a partition, it follows that there must be uncountably many distinct $[x]$'s that are furthermore disjoint. Now, by the axiom of choice, construct a set $V$ that selects only one element from each such distinct $[x]$. $V$ thus has uncountably many elements, but in some sense, is "thin". Consider all the circular shifts of $V$ of the form $V \oplus b$ for $b \in \mathcal{B}$. These are disjoint, since we never recreate the same element within the equivalence class of a given point $x$ (why?) nor jump from the equivalence class of $x$ to that of another. Furthermore as $b$ varies, each $x$ recreates its entire equivalence class, and it follows that:

$$\bigcup_{b \in \mathcal{B}} V \oplus b = [0,1].$$

So now we ask, what can $\nu(V)$ be? It certainly cannot be positive, since otherwise $\nu([0,1]) = \sum_{b \in \mathcal{B}} \nu(V \oplus b) = \sum_b \nu(V) = \infty$. But it cannot be zero either, since otherwise $\mathbb{P}([0,1]) = \sum_b \nu(V) = 0$. This is a contradiction. ∎

What went wrong? This is a very involved debate, but essentially the issue is an interaction between countable additivity and our ability to have created $V$ in the first place. The attitude of probability theory can be interpreted as either denying that such sets exists, or accepting that they do exist, but refusing to define the probability measure over them. The latter turns out to be much more productive, because the notion of restricting the probability measure to only given subsets has many versatile uses, including a generalization of the notion of conditioning.

## 3.1 Random variables over uncountably infinite probability spaces

To define a random variable, we need to define a $\sigma$-algebra on the range of the random variable. A random variables is then defined as a *measurable* function from the probability space to the range: functions where the pre-image of every subset in the range $\sigma$-algebra is an event in $\mathcal{F}$.

An important case is when the range is $[0, 1]$ or $\mathbb{R}$. In this case we say that we have a *real-valued* random variable, and we use the Borel $\sigma$-algebra unless otherwise noted. For countable probability spaces, we wrote the expectation of a real-valued random variable as a sum. For uncountable spaces, the expectation is an integral with respect to the measure.

$$\mathbb{E}[X] \;=\; \int_{\Omega} X(\omega) d\nu \,.$$

The definition of the integral with respect to a measure requires some amount of care, though we will not be able to discuss this in much detail. Let $\nu$ be any probability measure over the space $\mathbb{R}$ equipped with the Borel $\sigma$-algebra. Define the function $F$ as

$$F(x) \;:=\; \nu((-\infty, x]) \,,$$

which is well defined since the interval $(-\infty, x]$ is in the Borel $\sigma$-algebra. This can be used to define a random variable $X$ such that $\mathbb{P}[X \leq x] = F(x)$. The function $F$ is known as the distribution function or the cummulative density function of $X$.

When the function $F$ has the form

$$F(x) \;=\; \int_{-\infty}^{x} f(z) dz \,,$$

then $f$ is called the density function of the random variable $X$. In this case, one typically refers to $X$ as a "continuous" random variable. To calculate the above expectation for continuous random variables, we can use usual (Lebesgue) integration:

$$\mathbb{E}[X] \;=\; \int_{\mathbb{R}} x f(x) dx \,.$$

(The notion of density can be extended to between any two measures, via the Radon-Nikodym theorem. In that context, the density $f$ of a continuous random variable is referred to as the Radon-Nikodym derivative with respect to the Lebesgue measure. In the earlier example with the measure concentrated on the finite set $T$, the probability of each point is the Radon-Nikodym derivative with respect to the counting measure of $T$: $\nu_T = \sum_{t \in T} \delta_t$.)

## References


[ABKU94] Yossi Azar, Andrei Z Broder, Anna R Karlin, and Eli Upfal, *Balanced allocations*, Proceedings of the twenty-sixth annual ACM symposium on Theory of computing, ACM, 1994, pp. 593–602. 4

[ABKU99] _____, *Balanced allocations*, SIAM journal on computing **29** (1999), no. 1, 180–200. 4

[Mit96] Michael David Mitzenmacher, *The power of two random choices in randomized load balancing*, Ph.D. thesis, PhD thesis, Graduate Division of the University of California at Berkley, 1996. 4