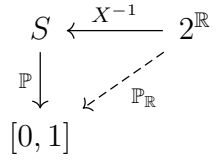
# Physics of Learning Theory
## Lecture 1
## Probability Review: Concentration bounds

January 29, 2025
Nilava Metya

## 1 Introduction

We will recall some probability theory and look at useful *deviation* or *concentration* bounds which are frequently used in analyzing algorithms (in learning theory). Recall that a (real-valued) random variable on a probability space $(\Omega, S, \mathbb{P})$ is nothing but a 'measurable function' $X : \Omega \to \mathbb{R}$. Here $\Omega$ is the universal or sample space where we think of events in, $S$ is a collection of events in $\Omega$ and $\mathbb{P} : S \to [0, 1]$ assigns probability to each event in $S$. The space of events $S$ is constrained to satisfy some obvious rules like $\Omega$ is an event, if $A$ is an event then so is $\Omega \setminus A$ and that a countable union of events is an event which makes it sensible to work with the concept of assigning probabilities to each event. We will often say that $\mathbb{P}[A]$ is the probability that event $A$ occurs. If $A = \{a\}$ is a singleton, we always write $\mathbb{P}[a]$ instead of $\mathbb{P}[\{a\}]$. The probability function $\mathbb{P}$ is also constrained to a couple of rules, namely, that the probability of the union of a mutually disjoint collection of events, which is an event, is the same as the sum of the probabilities of each of those events and that the probability that $\Omega$ occurs is $1$. Roughly a random variable is to be thought of as a way of assigning points of the sample space to real numbers which are *really real* and are more tangible to work with, while respecting the rules of $S$. Such a random variable induces a map $X^{-1} : 2^{\mathbb{R}} \to S$ by $X^{-1}(A) \coloneqq \{x \in \Omega \mid X(x) \in S\}$ for any $A \subseteq \mathbb{R}$, and hence induces a probability on $\mathbb{R}$ given by $\mathbb{P}_{\mathbb{R}}[A] = \mathbb{P}[X^{-1}(A)]$ where $A$ is any 'measurable' subset of $\mathbb{R}$. The random variable being a 'measurable function' precisely means that $X^{-1}(A)$ always lies in $S$.



## 1.1 Mean

The average or mean of a random variable $X$, often denoted as $\mathbb{E}[X]$, $\mu(X)$, or simply $\mu$ when the context is clear, is $\mathbb{E}[X] = \int_\Omega X \,\mathrm{d}\,\mathbb{P}$. For the discrete case, which we will mostly be interested in, this boils down to $\mathbb{E}[X] = \sum_{i\in\Omega} X(i)\mathbb{P}[i]$. Note that if $X$ is an indicator random variable for event $A$, that is, $X = 1$ if $A$ occurs and $0$ otherwise, then $\mathbb{E}[X] = \mathbb{P}[A]$.

*Example* 1. Consider tossing a fair coin. Here $\Omega = \{\mathrm{H}, \mathrm{T}\}$. The probability function is $\mathbb{P}[\varnothing] = 0, \mathbb{P}[\mathrm{H}] = \mathbb{P}[\mathrm{T}] = 0.5, \mathbb{P}[\{\mathrm{H}, \mathrm{T}\}] = 1$. A natural random variable to consider is $X(i) = \mathbf{1}_{\mathrm{H}} \coloneqq \begin{cases} 1 & \text{if } i = \mathrm{H} \\ 0 & \text{if } i = \mathrm{T} \end{cases}$. The corresponding probability induced on $\mathbb{R}$ is given by $\mathbb{P}_{\mathbb{R}}[A] = \begin{cases} 0 & \text{if } 0 \notin A, 1 \notin A \\ 0.5 & \text{if } 0 \in A, 1 \notin A \\ 0.5 & \text{if } 0 \notin A, 1 \in A \\ 1 & \text{if } 0 \in A, 1 \in A \end{cases}$. In this case, $\mathbb{E}[X] = 1 \cdot \mathbb{P}[\mathrm{H}] + 0 \cdot \mathbb{P}[\mathrm{T}] = 0.5$

*Example* 2. Consider tossing $n$ fair coins sequentially and independently. Here $\Omega = \{\mathrm{H}, \mathrm{T}\}^n$. So the singleton outcomes are tuples of H, T. The probability function is given by $\mathbb{P}[\boldsymbol{x}] = 2^{-n}$ for any element $x \in \Omega$ and then extending by countable additivity of $\mathbb{P}$. Consider $n$ random variables $X_1, \cdots, X_n$ where $X_i(\boldsymbol{x}) \coloneqq \begin{cases} 1 & \text{if } x_i = \mathrm{H} \\ 0 & \text{if } x_i = \mathrm{T} \end{cases}$. Each $X_i$ is the same random variable as the previous example after looking at the $i^{\text{th}}$ coordinate. A natural variable to consider is the total number of heads obtained in one round of tossing, that is $X = X_1 + \cdots + X_n$. The corresponding probability induced on $\mathbb{R}$ is given by $\mathbb{P}_{\mathbb{R}}[k] = \begin{cases} \binom{n}{k} 2^{-n} & \text{if } k \in \{0, \cdots, n\} \\ 0 & \text{otherwise} \end{cases}$ and extend by countable additivity. Here $\mathbb{E}[X] = \frac{n}{2}$.

One useful result used for calculating expectations of sums of random variables is that if $a, b \in \mathbb{R}$ and $X, Y$ are random variables then $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$. It's worthy to note that sums and scalings of random variables are random variables. This result does **not** depend on 'independence' of $X, Y$. Independence plays an important role for the average of products of random variables (which is a random variable). We say random variables $X_1, \cdots, X_n$ are (*mutually*) *independent* if $\mathbb{P}[\bigcap_{i=1}^n \{X_i \le a_i\}] = \prod_{i=1}^n \mathbb{P}[X_i \le a_i] \;\forall\; a_i \in \mathbb{R}$. This is a stronger notion than *pairwise* independence where we demand that only every pair of them are independent. Note that mutual independence implies pairwise independence. If $X, Y$ are independent then $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$.

For a random variable $X \ge 0$ and $a \in \mathbb{R}$ let $Y$ be the indicator random variable indicating whether $X \ge a$, that is, $Y$ is $1$ if $X \ge a$ and $0$ otherwise. Then clearly $X \ge aY$. Indeed if $X \ge a$ then $Y = 1$ so $X \ge aY$ and if $X < a$ then $Y = 0$ so that $X \ge 0 = aY$. Expectation preserves inequalities, so $\mathbb{E}[X] \ge a\mathbb{E}[Y] = a\mathbb{P}[X \ge a]$. This establishes

**Theorem 1** (Markov's inequality)
If $X$ is a non-negative random variable and $a \in \mathbb{R}$ then $\mathbb{P}[X \geq a] \leq \dfrac{\mathbb{E}[X]}{a}$.

## 1.2 Variance

Let's come to deviation now. One natural way to measure *deviation* is to look how on average much a random variable deviates either way from its mean (behavior). To look for deviation in either direction of $\mathbb{E}[X]$ we consider the random variable $(X - \mathbb{E}[X])^2$. Define the variance of a random variable $X$ as $\text{Var}[X] \coloneqq \mathbb{E}[(X - \mathbb{E}[X])^2]$. One useful result to compute variance is that if $X, Y$ are independent then $\text{Var}[aX + bY] = a^2\text{Var}[X] + b^2\text{Var}[Y]$. This extends to $n$ pairwise independent random variables. Another useful result is $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

Applying Theorem 1 to $(X - \mathbb{E}[X])^2 \geq 0$ gives

**Theorem 2** (Chebyshev's inequality)
If $X$ is a random variable and $a \in \mathbb{R}_{\geq 0}$ then $\mathbb{P}[|X - \mathbb{E}[X]| \geq a] \leq \dfrac{\text{Var}[X]}{a^2}$.

## 1.3 Higher moments

One might just ask why stop at the second power to measure deviation. What about the random variable $(X - \mathbb{E}[X])^k$ for $k \geq 2$? These are called higher centeral moments. Note that $\mathbb{E}\left[(X - \mathbb{E}[X])^k\right] = 0$ when $k$ is odd and the distribution of $X$ is symmetric about $\mathbb{E}[X]$. So it makes sense to consider the random variables $X_k \coloneqq |X - \mathbb{E}[X]|^k$ instead. If we have access to such numbers, we can use the same trick as the proof of Chebyshev's inequality and get $\mathbb{P}[|X - \mathbb{E}[X]| \geq a] \leq \dfrac{\mu_k}{a^k}$. Knowing all higher moments means that we know something known as the 'characteristic function' (not yet defined) of $X$ which uniquely determines $X$. But our aim was the study deviations using small information. Generally, higher moments are not known.

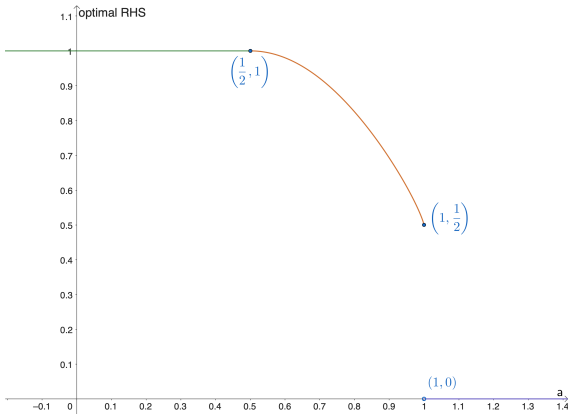
Here's a small trick to optimally apply Markov to a non-homogeneous function. Let's just take the 'best polynomial' ever known. It's non-homogeneous, positive, monotonic (but not *monotonous*) and has values at all points. We want to study the concentration of $e^{X-\mu}$. Take $f(x) = e^x \geq 0$. Chebyshev's inequality do this for $f = x^2$ but this was homegeneous so scaling the random variables had no effect on the inequalities obtained. Consider the random variable $Y_t = f(t(X - \mu))$ where $t$ is a real variable. Then applying Markov on $|Y_t| = Y_t$ gives $\mathbb{P}\left[Y_t = e^{t(X-\mu)} \geq e^{ta}\right] \leq \frac{\mathbb{E}[Y_t]}{e^{ta}} \ \forall \ t \geq 0, a \in \mathbb{R}$. This is equivalent to $\mathbb{P}[X - \mu \geq a] \leq \frac{\mathbb{E}\left[e^{t(X-\mu)}\right]}{e^{ta}}$. Since this is true for every $t \geq 0$, we conclude that $\mathbb{P}[X \geq a + \mu] \leq \inf\limits_{t \geq 0} \dfrac{\mathbb{E}\left[e^{t(X-\mu)}\right]}{e^{ta}}$. One issue with this argument is that $M_X(t) \coloneqq \mathbb{E}[\exp(tX)]$ may not always exist. Let's say they exist for $t \in [0, b]$ for some $b \geq 0$ (sanity check: $b = 0$

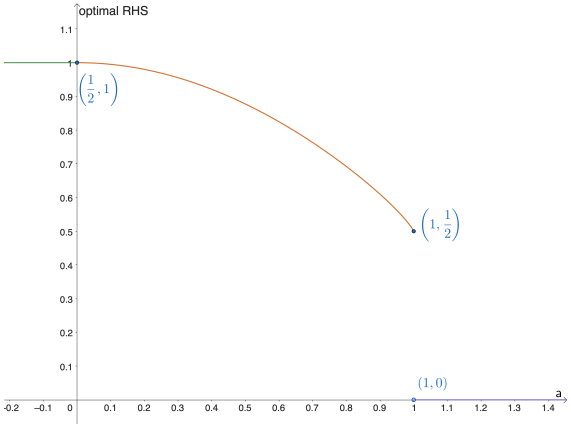
always works). Then we can modify our inequality to $\mathbb{P}[X \geq a+\mu] \leq \inf_{t\in[0,b]} \frac{\mathbb{E}\left[e^{tX}\right]}{e^{t(a+\mu)}}$. The *moment generating function* (mgf, in short) of a random variable $X$ is $M_X(t) = \mathbb{E}[\exp(tX)]$.

*Example* 3 (Bernouli). Say $X$ takes values $0, 1$ with probability $\frac{1}{2}$ each. Then $M_X(t) = \mathbb{E}[\exp(tX)] = \frac{1}{2}\exp t + \frac{1}{2}$ always exists. Our above inequality takes the form $\mathbb{P}[X \geq a] \leq \frac{1}{2}\inf_{t\geq 0} \frac{\exp t + 1}{e^{ta}}$. If $a \leq \frac{1}{2}$ then the RHS is $1$ at $t = 0$. If $\frac{1}{2} < a < 1$ then the RHS is $\frac{1}{2(1-a)^{1-a}a^a}$ at $t = \ln(a) - \ln(1-a)$. Taking $a \to 1^-$ gives that if $a = 1$, the RHS is $\frac{1}{2}$ attained at $''t = +\infty''$ (can also be checked directly by plugging in $a = 1$ directly). If $a > 1$ the RHS is $0$ again at $''t = +\infty''$.

*Example* 4 (Rademacher). Say $X$ takes values $\pm 1$ with probability $\frac{1}{2}$ each. Such a random variable is called a *Rademacher random variable*. Then $M_X(t) = \mathbb{E}[\exp(tX)] = \frac{1}{2}(e^t + e^{-t})$ always exists. Our above inequality takes the form $\mathbb{P}[X \geq a] \leq \frac{1}{2}\inf_{t\geq 0} \frac{e^t + e^{-t}}{e^{ta}}$. The RHS looks like

$$\begin{cases} 1 & \text{at } t^* = 0 \text{ if } a \leq 0 \\ \sqrt{\frac{1}{(1-a)^{1-a}(1+a)^{1+a}}} & \text{at } t^* = \frac{1}{2}\ln\left(\frac{1+a}{1-a}\right) \text{ if } a \in (0,1]. \\ 0 & \text{at } t^* = \infty \text{ if } a > 1 \end{cases}$$




## 1.4 Sub-Gaussian random variables

Now let's apply it to our favorite distribution – the Gaussian. Recall that the Gaussian distribution $Z$ with mean $\mu$ and variance $\sigma^2$ has the density $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$. The moment generating function of this Gaussian is $M(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$ and exists $\forall\, t \in \mathbb{R}$. Substituting this into our 'moment-based Markov inequality' gives $\mathbb{P}[Z \geq \mu + a] \leq \inf_{t\geq 0} \exp\left(\frac{\sigma^2 t^2}{2} - at\right) = \exp\left(-\frac{a^2}{2\sigma^2}\right)$. This means $\mathbb{P}[|Z-\mu| \geq a] \leq 2\exp\left(-\frac{a^2}{2\sigma^2}\right)$ for any $a \geq 0$.

This calculation let's us study the deviation of a large class of random variables, if this class is defined properly. If we revisit the calculation done for the Gaussian, we see that the only necessary property of any random variable $X$ that can get the same bound is the existence of some $\sigma^2$ such that we can get a similar function as an upper bound on the mgf of $X$. More precisely, we demand that there exist a real number $\sigma > 0$ such that $\mathbb{E}\left[\exp(t(X - \mathbb{E}[X]))\right] \leq \exp\left(\frac{t^2\sigma^2}{2}\right) \ \forall\ t \in \mathbb{R}$. Alternately, instead of using the proof and calculation details, one might suggest to study those class of random variables whose deviations are bounded by those of the Gaussian. They turn out to be the same.

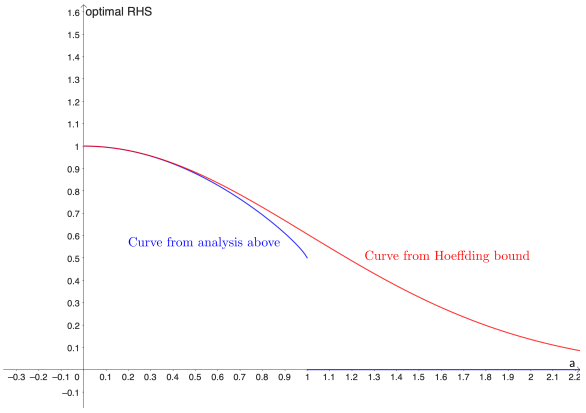
**Definition 3**
A random variable $X$ with mean $\mu$ is said to be *sub-Gaussian* if there is a constant $c > 0$ and a Gaussian $Z \sim \mathcal{N}(0, \tau^2)$ such that $\mathbb{P}\left[|X - \mu| \geq a\right] \leq c\,\mathbb{P}\left[|Z| \geq a\right]\ \forall\ a \geq 0$.

Alternately, a random variable $X$ with mean $\mu$ is said to be *sub-Gaussian* if there exists $\sigma > 0$ such that $\mathbb{E}\left[\exp(t(X - \mathbb{E}[X]))\right] \leq \exp\left(\frac{t^2\sigma^2}{2}\right) \ \forall\ t \in \mathbb{R}$. This $\sigma^2$ is said to be the sub-Gaussian parameter and acts as a proxy for variance.

This is quite a nice class because sub-Gaussianity is preserved under linear combinations. In particular, if $X_1, X_2$ are independent sub-Gaussians with parameters $\sigma_1^2, \sigma_2^2$ respectively then $X_1 + X_2$ is also a sub-Gaussian with parameter $\sigma_1^2 + \sigma_2^2$. In other words, the variance proxies add up just like the Gaussian. Using this property we immediately get

**Theorem 4** (Hoeffding)
If $\{X_i\}_{i=1}^m$ are independent sub-Gaussians with means $\{\mu_i\}_{i=1}^m$ and variance proxies $\{\sigma^2\}_{i=1}^m$ respectively. Then $\mathbb{P}\left[\sum_{i=1}^m (X_i - \mu_i) \geq t\right] \leq \exp\left\{-\frac{t^2}{2\sum_i \sigma_i^2}\right\}$ for all $t \geq 0$.

At our current discussion stage, sub-Gaussians seem quite useless. But, a lot of the 'good' random variables we see are actually sub-Gaussian. In fact if $X$ is a bounded random variable taking values in $[a, b]$ then $X$ is sub-Gaussian with parameter $\left(\frac{b-a}{2}\right)^2$. Here's a comparison of the bounds obtained with fine analysis as in Example 4 vs what the Hoeffding bound gives us. Notice the smoothness difference.




**Corollary 5**
Let $X_1, \cdots, X_n$ be independent bounded random variables such that $X_i \in [a_i, b_i]$ (almost surely) and sample mean $\overline{X}$. Then $\mathbb{P}\left[\overline{X} - \mathbb{E}\left[\overline{X}\right] \geq t\right] \leq \exp\left\{-\frac{2n^2t^2}{\sum_i (b_i - a_i)^2}\right\}$ for all $t \geq 0$.

## 1.5 Variations and applications of the Hoeffding bound

The first variation is obtained by replacing the sum $\sum_{i=1}^{k} X_i$ with its sample mean $\frac{1}{k}\sum_{i=1}^{k} X_i$.

**Corollary 6**
Let $X_1, \cdots, X_n$ be independent bounded random variables such that $X_i \in [a_i, b_i]$ (almost surely) and sample mean $\overline{X}$. Then $\mathbb{P}\left[\overline{X} - \mathbb{E}\left[\overline{X}\right] \geq t\right] \leq \exp\left\{-\frac{2n^2t^2}{\sum_i (b_i - a_i)^2}\right\}$ for all $t \geq 0$.

One can change parameters from $t$ to $\varepsilon := t + \mathbb{E}\left[\overline{X}\right]$ to get $\mathbb{P}\left[\overline{X} \geq \varepsilon\right] \leq \exp\left\{-\frac{2n^2\left(\varepsilon - \mathbb{E}\left[\overline{X}\right]\right)^2}{\sum_i (b_i - a_i)^2}\right\}$ for all $\varepsilon \geq \mathbb{E}\left[\overline{X}\right]$.

Using the above with all $X_i$'s (hence $a_i, b_i$'s) negated gives a lower tail bound, that is, $\mathbb{P}\left[\overline{X} \leq \varepsilon\right] \leq \exp\left\{-\frac{2n^2\left(\varepsilon - \mathbb{E}\left[\overline{X}\right]\right)^2}{\sum_i (b_i - a_i)^2}\right\}$ for all $\varepsilon \leq \mathbb{E}\left[\overline{X}\right]$.

Combining we have

**Corollary 7**
Let $X_1, \cdots, X_n$ be independent bounded random variables such that $X_i \in [a_i, b_i]$ (almost surely) and sample mean $\overline{X}$ and $\mu = \mathbb{E}\left[\overline{X}\right]$. Then

$$\mathbb{P}\left[\overline{X} \geq \varepsilon\right] \leq \exp\left\{-\frac{2n^2\left(\varepsilon - \mu\right)^2}{\sum_i (b_i - a_i)^2}\right\} \ \forall\, \varepsilon \geq \mu$$

$$\mathbb{P}\left[\overline{X} \leq \varepsilon\right] \leq \exp\left\{-\frac{2n^2\left(\varepsilon - \mu\right)^2}{\sum_i (b_i - a_i)^2}\right\} \ \forall\, \varepsilon \leq \mu.$$

That is, we have a symmetric tail bound on either side of $\mu$.

Recall that the Hoeffding bound gives the same bounds for a Bernouli random variable as a random variable taking values in $[0, 1]$. Somehow this extra information about Bernouli random variables can be incorporated to get the stronger Chernoff bound.

**Theorem 8** (Chernoff Bound)
Let $X_1, \cdots, X_n$ be independent $\{0, 1\}$ values random variables such that $p_i = \mathbb{E}[X_i]$, with $X = \sum_i X_i$ and $\mu = \mathbb{E}[X] = \sum_i p_i$. Then $\mathbb{P}[X \geq (1+\varepsilon)\mu] \leq \exp\left\{-\frac{\varepsilon^2\mu}{2+\varepsilon}\right\}$ for $\varepsilon > 0$ and $\mathbb{P}[X \leq (1-\varepsilon)\mu] \leq \exp\left\{-\frac{\varepsilon^2\mu}{2}\right\}$ for $\varepsilon \in (0, 1)$.

To understand why the Chernoff bound is slightly stronger, let's fix a probability parameter $\delta \in (0, 1)$ (to be thought of as the failure probability). Say $X_1, \cdots, X_n$ are

$\{0,1\}$ valued random variables with $p = \mathbb{E}[X_i]$ for each $i$. Then using Corollary 6 with $t = \sqrt{\frac{-\ln\delta}{2n}}$ gives $\mathbb{P}\left[\overline{X} \geq p + \sqrt{\frac{-\ln\delta}{2n}}\right] \leq \delta$ and using Theorem 8 with $\varepsilon = \sqrt{\frac{-3\ln\delta}{pn}}$ gives $\mathbb{P}\left[\overline{X} \geq p + \sqrt{\frac{-6p\ln\delta}{2n}}\right] \leq \delta$ as long as $p > \frac{3\ln(1/\delta)}{n}$. Note that this scenario happens only when $\delta$ is exponentially small (in terms of $n$). If $p$ is constant, the Chernoff bound gives no useful information for the rate. However, in certain scenarios the iid Bernouli parameters $p \equiv p_n$ depend on the number of samples and $p_n \to 0$ so Chernoff speaks louder.

Now we look at some examples where we apply the Hoeffding (or Chernoff bound) to analyze algorithms.

*Example* 5 (Boosting in two sided errors). Suppose we designed a randomized algorithm $f$ to answer a $0/1$ question and on any given input $x$, it answers correctly with probability $\frac{2}{3}$. How can we use $f$ to correctly predict its actual answer of input $x$ with very high confidence. Of course, we may or may not get the correct answer if we run $f$ once on $x$. Intuitively, if we run $f$ on $x$ $3000$ times, we expect to get about $2000$ correct answers and $1000$ wrong answers. Of course, then with high confidence we predict that the answer which is reported most number of times (that is, more than half the times) is the correct one. Intuitively, this makes sense. But, how do we quantify this confidence? We want to answer the question that how many times should we run $f$ on $x$ so that we succeed with probability $1 - \frac{1}{n}$.
Let's run the algorithm $n$ times on $x$ and let the outputs be $X_1, \cdots, X_k \in \{0,1\}$. Suppose the actual answer of $x$ on the actual question was $a \in \{0,1\}$ ($a$ is not random, but $X_i$'s are). Our reported answer is $Y = \mathbf{1}_{\overline{X} \geq \frac{1}{2}}$. This is also a random variable and we will show that the probability of $Y$ not being $a$ is very small. Note that $X_i$ are all Bernouli$((1+a)/3)$, so Hoeffding bound is good enough. Corollary 7 gives a the same tail bound on $\mathbb{P}\left[\overline{X} \geq \frac{1}{2}\right]$ and $\mathbb{P}\left[\overline{X} \leq \frac{1}{2}\right]$ (corresponding to the 'bad' event $\{Y \neq a\}$ for $a = 0, 1$ respectively). Thus $\mathbb{P}[Y \neq a] \leq \exp\{-2k((1-2a)/6)^2\} = \exp\{-k/18\}$ irrespective of whether $a$ is $0$ or $1$. Hence $k \geq 18\ln n$ trials gives us a confidence of $\geq 1 - \frac{1}{n}$.

*Example* 6 (Johnson-Lindenstrauss lemma [JL84]). Say we are a dimension $d$, a probability parameter $\delta \in (0, 1/2)$, fault tolerance $\varepsilon \in (0,1)$, a positive integer $m > \frac{-\ln\delta}{\varepsilon^2}$ and any vector $\boldsymbol{x} \in \mathbb{R}^d$. We pick a matrix $M \in \mathbb{R}^{m\times d}$ whose entries are independent $\mathcal{N}(0,1)$'s and consider $\Pi = \frac{1}{\sqrt{m}}M$. Then $\mathbb{P}\left[(1-\varepsilon)\|\boldsymbol{x}\|_2 \leq \|\Pi\boldsymbol{x}\|_2 \leq (1+\varepsilon)\|\boldsymbol{x}\|_2\right] \geq 1 - \delta$. This is known as the famous Johnson-Lindenstrauss dimensionality reduction. In fact, if we are given $n$ points $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \in \mathbb{R}^d$, by a union bound argument (because the above was for one $\boldsymbol{x}$) we can show that for $m = \mathcal{O}(\ln(n/\delta)/\varepsilon^2)$, all the distances $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$ are preserved under a random such $\Pi$ with probability $1 - \delta$. For instance, with probability $0.99$, we can reduce the dimension to $m = \mathcal{O}(\ln n/\varepsilon^2)$, upto $\varepsilon$ error. In other words, the existence of such a $\Pi$ has positive probability for small enough $\delta$. Since the $\Pi$'s were combinatorial (i.e., chosen from a finite set), we conclude that such a dimension-reducing $\Pi$ always exists.

We end with a variation of the Hoeffding bound and its relation to the geometry of (convex) bodies in $\mathbb{R}^n$. Suppose we have mean zero random variables $X_1, \cdots, X_n$ with $X_i \in \{-a_i, a_i\}$ satisfying $\sum_i a_i^2 = 1$. Then $\mathbb{E}\left[\sum_i X = 0\right]$. (The two-sided tail version of) Hoeffding bound gives $\mathbb{P}\left[|\sum_i X_i| \geq t\right] \leq 2\exp\left\{-\frac{t^2}{4\sum_i a_i^2}\right\} = 2\exp\left\{-t^2/4\right\}$. In other words, if $\boldsymbol{a}$ is a fixed vector with $\|\boldsymbol{a}\|_2 = 1$ and $\boldsymbol{Y} = (Y_1, \cdots, Y_n) \in \{\pm 1\}^n$ is chosen uniformly at random (so each $Y_i$ is an independent Rademacher), then $\mathbb{P}\left[|\langle \boldsymbol{a}, \boldsymbol{Y}\rangle| \geq t\right] \leq 2\exp\left\{-t^2/4\right\}$. A geometric interpretation is as follows. We call $C_n \coloneqq \{\pm 1\}^n$ as the boolean cube. $|\langle \boldsymbol{a}, \boldsymbol{Y}\rangle|$ measures the distance of $\boldsymbol{Y}$ from the hyperplane through $\mathbf{0}$ perpendicular to $\boldsymbol{a}$. The above bound just says that at least $1 - 2e^{-t^2/4}$ fraction of the volume of $C_n$ lies within distance $t$ from this hyperplane. This is the starting point of the realm of the vast area called *isoperimetric inequalities*.

# References


[JL84] William B. Johnson and Joram Lindenstrauss. "Extensions of Lipschitz mappings into Hilbert space". In: *Contemporary mathematics* 26 (1984), pp. 189–206. URL: https://api.semanticscholar.org/CorpusID:117819162.

[Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.