Math 626: High Dimensional Probability

Taught by Mark Rudelson
Lecture notes by Sayantan Khan

February 22, 2021

Contents

1 Introduction

High dimensional probability is the study of random variables taking values in $\mathbb{R}^n$ for large but fixed values of n. While this area has always been studied by classical probability theorists, it has also attracted attention from computer scientists, especially since the design and analysis of fast probabilistic algorithms requires tools and theorems from this field. A classical result that arises from pure mathematics, but has several real life applications is Dvoretzky's theorem, whose statement does not involve any probability at all, and yet the proof uses concentration inequalities.

Theorem 1.1 (Dvoretzky's theorem [Dvo61])**.** *Let X be an n-dimensional normed vector space. Then for any $\varepsilon > 0$, there exists a constant $c(\varepsilon) > 0$, and a linear subspace E of X such that, $\dim(E) \geq c(\varepsilon)\log(n)$, and for all $v \in E$, the following inequality relating the ambient norm and the Euclidean norm on E.*

$$(1-\varepsilon)\|v\|_2 \leq \frac{\|v\|_X}{M(X)} \leq (1+\varepsilon)\|x\|_2$$

Where $M(X)$ is a scaling factor that depends only on X, and $\|\cdot\|_X$ and $\|\cdot\|_2$ are the ambient and Euclidean norm on E.

[git] • Commit hash: f56b0c5 (2021-02-22)

Idea of proof. Pick a random subspace, and then show that with very high probability, the given inequality holds. We will prove the result in full detail later in the course. □

Remark 1.2. When the norm on X is the ℓ^1 norm, the lower bound on the dimension of E can be improved to be linear in $c(\varepsilon)n$. ◇

A computer science application of Dvoretzky's theorem Consider a subset S of ℓ_2^N, given by inequalities involving the norms of elements in ℓ_2^N. Suppose that we are required to optimize a linear function f on the set S. Since S is given by inequalities involve the ℓ^2-norm, it will be and intersection of interiors of ellipsoids, and consequently, optimizing f will be computationally expensive. But we can get around the computational expense by embedding ℓ_2^N into ℓ_1^M, where M is $O(N)$, by Remark 1.2. Since this embedding does not distort distances too much, we can replace S with a nearby polytope, given by inequalities involving the ℓ^1 norm instead. Optimizing a linear function on a polytope is computationally much easier, thanks to linear programming.

Empirical covariance estimation Let X be an $\mathbb{R}^n$-valued random variable, and let $\mathbb{E}(X) = 0$. The covariance of X is $\mathbb{E}(X^\top X)$, and will be denoted by Σ. Let $\{X_1, X_2, \ldots, X_m\}$ be i.i.d. samples of X. We define the sample covariance Σ_m in the following manner.

$$\Sigma_n = \frac{\sum_{i=1}^m X_i^\top X_i}{m}$$

As m tends to infinity, the sample covariance Σ_m will approach the true covariance, as we would expect the law of large numbers to predict. A harder, and more interesting question is to determine how many samples do we need to take to be within some threshold of the true covariance with high probability.

Just like in the scalar setting, one answers the question by proving appropriate concentration inequalities for matrix valued random variables. Here is the most general set up: Let Σ_m and Σ be matrices mapping some normed space X to some other normed space Y. We define the distance between Σ_m and Σ using the operator norm.

$$\begin{aligned}
\|\Sigma_m - \Sigma\|_{\text{op}} &= \max_{\|v\|_x \leq 1} \|(\Sigma_m - \Sigma)v\|_Y \\
&= \max_{\|v\|_x \leq 1} \max_{\|w\|_{Y^*} \leq 1} w\left((\Sigma_m - \Sigma)v\right) \\
&= \max_{\substack{(v,w) \in X \times Y^* \\ \|v\| \leq 1 \\ \|w\| \leq 1}} w((\Sigma_m - \Sigma)v)
\end{aligned}$$

We can consider $w((\Sigma_m - \Sigma)v)$ to be a family of scalar random variables parameterized by points in $X \times Y^*$, i.e. a random process V_u parameterized by $u \in X \times Y^*$. This leads to the following two questions.

(i) How to bound $\mathbb{E} \max V_u$.

(ii) How to bound $\mathbb{P}(|\max V_u - \mathbb{E} \max V_u| \geq t)$.

It turns out one can often answer (ii) without answering (i), which may seem surprising given that the most elementary concentration inequalities involve moment bounds (i.e. Markov's inequality). The starting point in answering (ii) is understanding *concentration of measure*.

2 Concentration of measure

Concentration of measure was originally observed Lévy, but first used by Milman in the early 1970s. Roughly speaking, concentration of measure is the following phenomenon: suppose $(X, d, \mathbb{P})$ is a metric space endowed with a probability measure and $f : X \to \mathbb{R}$ is a "nice" function. Then the value of f is essentially constant, i.e. there exists some constant $M(f)$ such that for small enough ε, the following probability bound holds.

$$\mathbb{P}(|f(x) - M(f)| < \varepsilon) \ll 1$$

Usually by nice, we will mean 1-Lipschitz, although similar results hold for a more general class of functions like convex or quasi-convex functions. We begin with the simplest version of measure concentration.

2.1 Concentration for linear functions

Let the metric space X in this setting be $\mathbb{R}^n$ for some fixed n, and let $\{X_1, \ldots, X_n\}$ be i.i.d scalar random variables. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a linear function, given by taking inner product with some vector $\mathbf{a}$. We define Y to be $\sum_{i=1}^n a_i X_i$. We will prove concentration inequalities for Y by imposing conditions on X_i: one such condition is requiring X_i to be *subgaussian*.

2.1.1 Subgaussian random variables

Definition 2.1 (Subgaussian decay)**.** A random variable X is said to be σ-subgaussian if there exists a constant $C > 0$, such that for all $t > 0$, the following inequality holds.

$$\mathbb{P}(|X| > t) \leq C \exp\left(-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2\right)$$

◇

Remark 2.2. The constant σ is often referred to as the variance proxy of the subgaussian random variable. ◇

Example 2.3. The following random variables are examples of subgaussian random variables.

(i) Normal random variables

(ii) Bounded random variables

◇

Lemma 2.4. *Let X be a random variable. Then the following are equivalent:*

(i) For all $t > 0$, $\mathbb{P}(|X| > t) \leq C \exp\left(-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2\right)$ (definition of subgaussian random variables).

(ii) There exists $a > 0$ such that $\mathbb{E}(\exp(aX^2)) < \infty$ (ψ_2 condition).

(iii) There exist C' and b such that for all $\lambda \in \mathbb{R}$, $\mathbb{E}(\exp(\lambda X)) \leq C' \exp(b\lambda^2)$ (Laplace transform condition).

(iv) There exists K such that for all $p \geq 1$, $\mathbb{E}(X^p)^{\frac{1}{p}} \leq K\sqrt{p}$ (moment condition).

Moreover, if $\mathbb{E}(X) = 0$, then the constant C' in the Laplace transform condition can be taken to be equal to 1.

Proof. $(i) \implies (ii)$ Using Fubini's theorem and a change of variables, we can express $\mathbb{E}(aX^2)$ as an integral involving tail bounds.

$$\begin{aligned}\mathbb{E}(aX^2) &= 1 + \int_0^\infty 2ate^{at^2} \cdot \mathbb{P}\left(|X| > t\right) \, \mathrm{d}t \\ &\leq 1 + \int_0^\infty 2Cat \exp\left(at^2 - \frac{1}{2}\left(\frac{t}{\sigma}\right)^2\right) \, \mathrm{d}t\end{aligned}$$

Clearly, picking a value of a smaller than $\frac{1}{2\sigma^2}$ will make the integral converge.

$(ii) \implies (iii)$ Since we want to estimate the expectation of $\exp(\lambda X)$, we multiply and divide by $\exp(aX^2)$ and complete the square.

$$\exp(\lambda X) = \exp\left(aX^2\right) \cdot \exp\left(\frac{\lambda^2}{4a}\right) \cdot \exp\left(-\left(\sqrt{a}X + \frac{\lambda}{2\sqrt{a}}\right)^2\right)$$

Note that the third term in the product is always less than 1. We now take the expectation of the right hand side.

$$\begin{aligned}\mathbb{E}(\exp(\lambda X)) &= \mathbb{E}\left(\exp\left(aX^2\right) \cdot \exp\left(\frac{\lambda^2}{4a}\right) \cdot \exp\left(-\left(\sqrt{a}X + \frac{\lambda}{2\sqrt{a}}\right)^2\right)\right) \\ &\leq \exp\left(\frac{\lambda^2}{4a}\right) \mathbb{E}\left(\exp(aX^2)\right)\end{aligned}$$

Setting $b = \frac{1}{4a}$ and $C' = \mathbb{E}(\exp(aX^2))$, we get the result.

$(iii) \implies (iv)$ We begin by getting a crude estimate for $\mathbb{E}(X^p)$ using the infinite series for $\exp(\lambda X)$.

$$\begin{aligned}\mathbb{E}(X^p) &\leq \frac{p!}{\lambda^p}\mathbb{E}(\exp(\lambda X)) \\ &= \frac{C'p! \exp(b\lambda^2)}{\lambda^p}\end{aligned}$$

Note that this inequality works for all values of λ, but to get the best inequality, we minimize the right hand side by varying λ over $\mathbb{R}$. The minimum is attained when $\lambda = \frac{\sqrt{p}}{2b}$: plugging that into the right hand side, and taking p^{th} roots gives us the following.

$$\mathbb{E}(X^p)^{\frac{1}{p}} \leq C'' \frac{(p!)^{\frac{1}{p}}}{\sqrt{p}}$$

Here, we have absorbed all the constants into C''. Using Stirling's approximation, the numerator is bounded above by p, giving us the inequality we want.

$(iv) \implies (i)$ We rewrite the event $|X| > t$ in the following manner.

$$\begin{aligned}\mathbb{P}(|X| > t) &= \mathbb{P}(\exp(\lambda X^2) > \exp(\lambda t^2)) \\ &\leq \exp(-\lambda t^2)\mathbb{E}(\exp(\lambda X^2))\end{aligned}$$

Here, λ is a positive real number that we will specify later, and the inequality comes from Markov's inequality. Of course, we do not a priori know that $\mathbb{E}(\lambda X^2)$ is finite, but we will pick a λ small enough such that it is. Using Fubini's theorem, we can express $\mathbb{E}(\exp(\lambda X^2))$ in the following manner.

$$\mathbb{E}(\exp(\lambda X^2)) = 1 + \frac{\lambda \mathbb{E}(X^2)}{1!} + \frac{\lambda^2 \mathbb{E}(X^4)}{2!} + \cdots$$

Using the bound on the moments of X and Stirling's approximation, we get the following inequality.

$$\begin{aligned}\mathbb{E}(\exp(\lambda X^2)) &\leq \sum_{p=0}^{\infty} \frac{(2\lambda K^2 p)^p}{p!} \\ &\leq \sum_{p=0}^{\infty} \frac{(2\lambda e K^2 p)^p}{\sqrt{2\pi p} p^p}\end{aligned}$$

If we pick λ to be small enough that $2e\lambda K^2$ is much smaller than 1, then the infinite sum converges, and the expectation is finite. Setting $\frac{1}{2\sigma^2}$ to be equal to λ gives us the result.

We now show that the constant in the Laplace transform condition can be set to be 1 when $\mathbb{E}(X) = 0$. To do so, we recall the ψ_2 and the Laplace transform condition, i.e. there exist constants a, C' and b such that the following two inequalities hold for all $\lambda \in \mathbb{R}$.

$$\mathbb{E}(\exp(aX^2)) < \infty \tag{2.5}$$

$$\mathbb{E}(\exp(\lambda X)) \leq C \exp(b\lambda^2) \tag{2.6}$$

Suppose now that $\lambda^2 > 2a$. By the Laplace transform condition, we have the following inequality.

$$\begin{aligned}\mathbb{E}(\exp(2aX)) &\leq C' \exp(4ba^2) \\ &= \exp(4a^2 b')\end{aligned}$$

Where b' is $b + \frac{\log(C')}{4a^2}$. Since b' decreases as a increases, for any $\lambda^2 > 2a$, $\mathbb{E}(\exp(aX^2))$ will be less than $\exp(b'\lambda^2)$.

Now suppose that $\lambda^2 < 2a$. We begin by considering the special case where X is a symmetric random variable. By symmetry of X, we have the following inequality.

$$\begin{aligned}\exp(\lambda X) &= \frac{\exp(\lambda X) + \exp(-\lambda X)}{2} \\ &\leq \exp\left(\frac{\lambda^2 X^2}{2}\right)\end{aligned}$$

Taking expectations on both sides gives us the following.

$$\mathbb{E}(\exp(\lambda X)) \leq \mathbb{E}\left(\exp\left(\frac{\lambda^2}{2}X^2\right)\right)$$

Since $\lambda^2 < 2a$, $\frac{2a}{\lambda^2}$ is greater than 1, and we can use Hölder's inequality to bound the right hand term.

$$\mathbb{E}\left(\exp\left(\frac{\lambda^2}{2}X^2\right)\right) \leq \left(\mathbb{E}\left(\exp\left(aX^2\right)\right)\right)^{\frac{\lambda^2}{2a}}$$

Since $\mathbb{E}(\exp(aX^2))$ is a finite constant, the right hand side is $\exp(b''\lambda^2)$ for some constant b''.

Now suppose X is not symmetric. Let X' be an identical independent copy of X. Since $\mathbb{E}(X')$ is 0, we have the following equality.

$$\mathbb{E}(\exp(\lambda X)) = \mathbb{E}(\exp(\lambda(X - \mathbb{E}(X'))))$$

Since exp is a convex function, we can pull out the inner expectation, using Jensen's inequality.

$$\mathbb{E}(\exp(\lambda X)) \leq \exp(\lambda(X - X'))$$

Since $X - X'$ is symmetric, the result follows from the previous part, and the proof is complete. □

We now explain why care so much about the constant in the Laplace transform condition.

Theorem 2.7 (Hoeffding-Chernoff-Azuma inequality)**.** *Let $\{X_1, \ldots, X_n\}$ be i.i.d. subgaussian random variables with mean 0. Then for any $(a_1, \ldots, a_n) \in \mathbb{R}^n$ and any $t > 0$, the following probability bound holds.*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} a_i X_i\right| > t\right) \leq C \exp\left(-c\frac{t^2}{\|\mathbf{a}\|_2}\right)$$

Where C and c are some absolute constants.

Proof. Without loss of generality, we can assume $\|\mathbf{a}\|_2 = 1$. It will suffice to show that the sum of subgaussian random variables is subgaussian. We will show it verifying the Laplace transform condition. Let $\lambda \in \mathbb{R}$, and let $Y = \sum_{i=1}^{n} a_i X_i$. We compute $\mathbb{E}(\exp(\lambda Y))$.

$$\begin{aligned}
\mathbb{E}(\exp(\lambda Y)) &= \prod_{i=1}^{n} \mathbb{E}(\exp(\lambda a_i X_i)) \\
&\leq \prod_{i=1}^{n} \mathbb{E}(\exp(b\lambda^2 a_i^2)) \\
&= \mathbb{E}(\exp(b\lambda^2))
\end{aligned}$$

This proves the result. Note that having the Laplace transform coefficient equal to 1 helped, because if the coefficient C was greater than 1, then we would pick up a constant of C^n, which would be very large for large values of n. □

To see how this inequality is used in practice, consider the simplest possible example of a subgaussian random variable, a random variable that takes values 1 and -1 with probability $\frac{1}{2}$.

Let's recall some elementary facts about Fourier analysis before stating the example. Let $L^2([0,1])$ be the space of complex L^2 functions on $[0,1]$ and let $\{e_n\}_{n\in\mathbb{Z}}$ be the standard Fourier basis, i.e. $e_n(t) = \exp(2\pi int)$. Since $\{e_n\}$ forms an orthonormal basis, any function $f \in L^2([0,1])$ can be decomposed into its Fourier series.

$$f = \sum_{n\in\mathbb{Z}} \widehat{f}(n) e_n$$

The Fourier coefficient $\widehat{f}(n)$ is given by $\int_0^1 f(t)\overline{e_n(t)}\,\mathrm{d}t$. The map sending f to its Fourier coefficients is a linear isometry from $L^2([0,1])$ to $\ell^2(\mathbb{Z})$. For most sequences $\{\widehat{f}(n)\}$ in $\ell^2(\mathbb{Z})$, the associated function f will not be continuous, but we will show that under reasonably mild conditions on $\{\widehat{f}(n)\}$, f can be made to be continuous.

Let $\varepsilon_n \in \{-1,1\}$ for $n \in \mathbb{Z}$, and let $\varepsilon = \{\varepsilon_n\}_{n\in\mathbb{Z}}$. For any $f \in L^2([0,1])$ and any such *varepsilon*, define f_ε to be the following function.

$$f_\varepsilon = \sum_{n\in Z} \varepsilon_n \widehat{f}(n) e_n$$

We then have the following theorem.

Theorem 2.8. *Let f be a function in $L^2([0,1])$ whose Fourier coefficients satisfy the following inequality.*

$$\sum_{n\in\mathbb{Z}} (\log(|n|+1))^3 \left|\widehat{f}(n)\right|^2 < \infty$$

Let $\{\varepsilon_n\}$ be a sequence of i.i.d random variables taking values in 1 and -1 with probability $\frac{1}{2}$. Then f_ε is a continuous function with probability 1.

To prove the theorem, we will need several lemmas.

Lemma 2.9. *Let $N \in \mathbb{N}$. Then for any $\{a_n\}_{n=-N}^N$, the following probability bound holds.*

$$\mathbb{P}\left(\left\|\sum_{n=-N}^N \varepsilon_n a_n e_n\right\|_\infty > C\sqrt{\log(N)}\left(\sum_{n=-N}^N |a_n|^2\right)^{\frac{1}{2}}\right) \leq \frac{1}{N^2}$$

Here the constant C is absolute, i.e. independent of N.

Proof. The first step is to consider the maximum, not over all of $[0,1]$, but over the points $\left\{\frac{j}{N^2}\right\}_{j=0}^N$. It will suffice to bound the probability for any given point by $\frac{1}{N^4}$, and then use the union bound to get the desired inequality. Since ε_n is a subgaussian random variable with mean 0, by Hoeffding-Chernoff-Azuma inequality, we get the following bound for any $t \in [0,1]$.

$$\mathbb{P}\left(\left\|\sum_{n=-N}^N \varepsilon_n a_n e_n\right\|_\infty > C\sqrt{\log(n)}\left(\sum_{n=-N}^N |a_n|^2\right)^{\frac{1}{2}}\right) \leq C' \exp(-cC^2\log(n))$$

We can pick a C large enough so that the right hand side is bounded above by $\frac{1}{N^4}$, and that concludes the first step after we use the union bound. Note that in order to do this, we really needed something stronger than Chebyshev's inequality, since we needed an upper bound that can be made smaller than $\frac{1}{N^4}$.

The next step is to extend the argument to all of $[0,1]$. The key trick here will be to estimate the maximum value of trigonometric polynomials away from points of the form $\frac{j}{N^2}$ using the maximum we derived in step 1. Bernstein's inequality for trigonometric polynomial helps in this regard: given a trigonometric polynomial p of degree n, the following inequality relates the $\|\cdot\|_\infty$-norm of p' and p.

$$\|p'\|_\infty \leq n \|p\|_\infty$$

Let V be the maximum value of the trigonometric polynomial $p = \sum_{n=-N}^{N} \varepsilon_n a_n e_n$ achieves on points of the form $\frac{j}{N^2}$, and let W be the maximum value over all of $[0,1]$, and say it is achieved at some point t, and let s be the closest point of the form $\frac{j}{N^2}$. Then, by the mean value theorem, we get the following relation between V and W.

$$\begin{aligned} W &\leq V + \|p'\|_\infty |t-s| \\ &\leq V + \frac{N \|p\|_\infty}{N^2} \\ &= V + \frac{W}{N} \end{aligned}$$

This means for $N > 1$, $W \leq 2V$, and this proves the lemma. □

Proof of Theorem 2.8. For $M \in \mathbb{N}$, define the function $f_{M,\varepsilon}$ in the following manner.

$$f_{M,\varepsilon} = \sum_{2^M \leq |n| < 2^{M+1}} \varepsilon_n \widehat{f}(n) e_n$$

We now use Lemma 2.9 with $N = 2^{M+1}$, $a_n = 0$ for $|n| < 2^M$, and $a_n = \widehat{f}(n)$ for $2^M \leq n < 2^{M+1}$.

$$\mathbb{P}\left(\|f_{M,\varepsilon}\|_\infty > C\sqrt{M} \|f_{M,\varepsilon}\|_2 \right) < \frac{1}{2^{2(M+1)}}$$

By the Borel-Cantelli lemma, for almost every ε, $\|f_{M,\varepsilon}\|_\infty$ eventually becomes smaller than $C\sqrt{M} \|f_{M,\varepsilon}\|_2$.

Pick ε to be one of the instances where the above described situation does happen. We have that f is an infinite sum of continuous functions.

$$f = \varepsilon_0 \widehat{f}(0) e_0 + \sum_{M=0}^{\infty} f_{M,\varepsilon}$$

This will converge to a continuous function if the sequence of partial sums is uniformly Cauchy. To see that is indeed the case, pick K_1 and K_2 larger than the threshold M after

which $\|f_{M,\varepsilon}\| < C\sqrt{M}$.

$$\begin{aligned}
\|f_{K_2,\varepsilon} - f_{K_1,\varepsilon}\|_\infty &\leq \sum_{M=K_1}^{K_2} \|f_{M,\varepsilon}\|_\infty \\
&\leq \sum_{M=K_1}^{\infty} C\sqrt{M}\,\|f_{M,\varepsilon}\|_2 \\
&\leq C\left(\sum_{M=K_1}^{\infty}\left(\frac{1}{M}\right)^2\right)^{\frac{1}{2}}\left(\sum_{M=K_1}^{\infty} M^3\,\|f_{M,\varepsilon}\|_2^2\right)^{\frac{1}{2}} \\
&\leq \frac{C'}{K_1}\left(\sum_{n\in\mathbb{Z}}(C''\log(|n|+1))^3\widehat{f}(n)^2\right) \\
&\leq \frac{C'''}{K_1}
\end{aligned}$$

The upper bound goes to 0 as K_1 goes to ∞, which shows the sequence is uniformly Cauchy, and thus the limit is a continuous function. □

We now prove a moment bound for sums of subgaussian random variables.

Theorem 2.10 (Khintchine's inequality [Khi23])**.** *Let $\{X_1, \ldots, X_n\}$ be i.i.d subgaussian random variables with $\mathbb{E}(X_i) = 0$ and $\mathbb{E}(X_i^2) = 1$. Then for any $p \in [1, \infty)$, there exist constants A_p and B_p greater than 0 such that for any vector $\mathbf{a} \in \mathbb{R}^n$, the following moment bound holds.*

$$A_p\,\|\mathbf{a}\|_2 \leq \left(\mathbb{E}\left|\sum_{i=1}^n a_i X_i\right|^p\right)^{\frac{1}{p}} \leq B_p\,\|\mathbf{a}\|_2$$

Proof. Consider first the case where $p > 2$. Then, using Hölder's inequality for the convex function $x \mapsto x^{\frac{p}{2}}$, we get the following.

$$\left(\mathbb{E}\left|\sum_{i=1}^n a_i X_i\right|^2\right)^{\frac{1}{2}} \leq \left(\mathbb{E}\left|\sum_{i=1}^n a_i X_i\right|^p\right)^{\frac{1}{p}}$$

This shows that for $p > 2$, we can set $A_p = 1$. To get B_p, we use the moment condition on subgaussian random variables. Since the sum of subgaussian random variables is subgaussian, we have that $Y = \sum_{i=1}^n a_i X_i$ is subgaussian, and thus satisfies the moment bound. We have seen that the absolute constant K will only depend on $\|\mathbf{a}\|_2$, giving us the upper bound.

$$\mathbb{E}(|Y|^p)^{\frac{1}{p}} \leq K\sqrt{p}\,\|\mathbf{a}\|_2$$

Setting $B_p = K\sqrt{p}$ proves the result in this case.

For $p < 2$, it will suffice to prove it for $p = 1$, since the p^{th} moment of $|Y|$ is an increasing function of Y and bounded above by $\|\mathbf{a}\|_2$ by Hölder's inequality, which means we can set $B_p = 1$ (or the previous argument will also work, but $B_p = 1$ is better than $B_p = K\sqrt{p}$).

Thus we just need to show the lower bound for $p = 1$. In this case, the inequality follows from Cauchy-Schwartz.

$$\mathbb{E}(|Y|^2) \leq \sqrt{\mathbb{E}(|Y|)\mathbb{E}(|Y|^3)}$$

Using Khintchine's inequality for $p = 3$, we deal with $\mathbb{E}(|Y|^3) \leq B_3 \|\mathbf{a}\|_2$. Squaring both sides, we see that $A_1 = B_3^{-3}$ works, and the proof is complete. □

There is a far reaching generalization of Khintchine's inequality, due to Kahane.

Theorem 2.11 (Kahane inequality [Kah64])**.** *Let X be a normed vector space, and let $\{\varepsilon_1, \ldots, \varepsilon_n\}$ be i.i.d. Rademacher random variables. For any $p \in [1, \infty)$, there exists A_p and B_p greater than 0 such that for any $\{a_1, \ldots, a_n\}$ in X, the following holds.*

$$A_p \left(\mathbb{E} \left\| \sum_{j=1}^n \varepsilon_j a_j \right\|_X^2 \right)^{\frac{1}{2}} \leq \left(\mathbb{E} \left\| \sum_{j=1}^n \varepsilon_j a_j \right\|_X^p \right)^{\frac{1}{p}} \leq B_p \left(\mathbb{E} \left\| \sum_{j=1}^n \varepsilon_j a_j \right\|_X^2 \right)^{\frac{1}{2}}$$

The proof of this inequality requires more machinery than the previous result, so we'll defer the proof until we have developed the required tools.

Recall that we got strong tail decay for bounded random variables using Hoeffding-Chernoff-Azuma inequality, since they're subgaussian, but it turns out, we can do much better than that using boundedness.

Theorem 2.12 (Bennett's inequality)**.** *Let $\{X_1, \ldots, X_n\}$ be i.i.d random variables satisfying the following properties.*

(i) $\|X_j\|_\infty \leq 1$.

(ii) $\mathbb{E}(X_j) = 0$ *and* $\mathbb{E}(X_j^2) = \delta$.

Then for any $\mathbf{a} \in \mathbb{R}^n$, we have the following tail decay estimate.

$$\mathbb{P}\left(\left| \sum_{j=1}^n a_j X_j \right| > t \right) \leq \begin{cases} 2\exp\left(-\frac{t^2}{2e\delta\|\mathbf{a}\|_2^2} \right) & \text{for } t \leq t_* \\ 2\exp\left(-\frac{t}{4\|\mathbf{a}\|_\infty} \cdot \log\left(\frac{t\|\mathbf{a}\|_\infty}{\delta\|\mathbf{a}\|_2^2} \right) \right) & \text{for } t > t_* \end{cases}$$

Here $t_* = e\delta\|\mathbf{a}\|_2^2$.

Before we start to prove the theorem, let's show why the described tail bound is a very natural upper bound to consider. Consider $\mathbf{a} = (1, 1, \ldots, 1)$. Then $\sum a_j X_j$ is approximately $\mathcal{N}(0, \delta\|\mathbf{a}\|_2^2) = \mathcal{N}(0, \delta n)$. The tail should then behave something like $\exp\left(-\frac{t^2}{2\delta n}\right)$, which is precisely the first case in the upper bound. If δ is small, i.e. δn is bounded above by some constant λ, then the central limit theorem asymptotic does not apply, but rather the Poisson limit theorem asymptotic applies.

$$\begin{aligned} \mathbb{P}\left(\sum_{j=1}^n a_j X_j > t \right) &\sim \sum_{j=t}^\infty e^{-\lambda} \frac{\lambda^j}{j!} \\ &\sim e^{-\lambda} \frac{\lambda^j}{j!} \\ &\sim \exp\left(-t \log\left(\frac{t}{e\lambda} \right) \right) \end{aligned}$$

Contrast this with the second case of the tail in Bennett's inequality.

Proof of Theorem 2.12. Without loss of generality, assume that $\|\mathbf{a}\|_\infty = 1$. Set $Y = \sum_{j=1}^n a_j X_j$, and let $\lambda > 0$. We then estimate the Laplace transform of Y.

$$\mathbb{E}\left(\exp(\lambda Y)\right) = \prod_{j=1}^n \mathbb{E}\left(\exp\left(\lambda a_j X_j\right)\right)$$

We use an elementary inequality to estimate the Laplace transform.

$$e^x \leq 1 + x + \frac{x^2}{2} e^{|x|}$$

Thus, we have the following.

$$\begin{aligned}\mathbb{E}\left(\exp\left(\lambda a_j X_j\right)\right) &\leq \mathbb{E}\left(1 + \lambda a_j X_j + \frac{\lambda^2 a_j^2 X_j^2}{2} \exp\left(|\lambda a_j X_j|\right)\right) \\ &\leq 1 + 0 + \frac{\lambda^2 a_j^2 \delta}{2} e^\lambda\end{aligned}$$

Putting all of it back together, we have the following inequality.

$$\begin{aligned}\mathbb{E}\left(\exp\left(\lambda Y\right)\right) &\leq \prod_{j=1}^n \left(1 + \frac{\lambda^2 a_j^2 \delta e^\lambda}{2}\right) \\ &\leq \prod_{j=1}^n \exp\left(\frac{\lambda^2 a_j^2 \delta e^\lambda}{2}\right) \\ &= \exp\left(\frac{\delta \lambda^2 e^\lambda}{2} \|\mathbf{a}\|_2^2\right)\end{aligned}$$

Now that we have the Laplace transform estimate, we can use it to estimate tail probabilities.

$$\begin{aligned}\mathbb{P}(Y > t) &= \mathbb{P}(\exp(\lambda Y) > e^{\lambda t}) \\ &\leq \frac{\exp\left(\frac{\delta \lambda^2 e^\lambda}{2} \|\mathbf{a}\|_2^2\right)}{e^{\lambda t}} \\ &= \exp\left(-\lambda t + \frac{\delta \lambda^2 e^\lambda}{2} \|\mathbf{a}\|_2^2\right)\end{aligned}$$

The next step is to optimize this inequality as we vary λ. We begin by optimizing in the region $\lambda \leq 1$. In this case, the optimum λ is $\frac{t}{e\delta \|\mathbf{a}_2^2\|}$. Plugging in this value of λ gives the first case of the upper bound, and this is valid for $t \leq e\delta \|\mathbf{a}\|_2^2 = t_*$.

In the region $\lambda > 1$, we use the inequality $\lambda \leq e^\lambda$.

$$\begin{aligned}\mathbb{P}(Y > t) &\leq \exp\left(-\lambda t + \frac{\delta \lambda^2 e^\lambda}{2} \|\mathbf{a}\|_2^2\right) \\ &\leq \exp\left(-\lambda t + \frac{\delta \lambda e^{2\lambda}}{2} \|\mathbf{a}\|_2^2\right)\end{aligned}$$

Choose λ such that $\lambda \|\mathbf{a}\|_2^2 e^{2\lambda} = t$. Plugging that value in, we get the second case.

Finally, doing this for $-t$ gives similar bounds, and we combine the two using union bound to get the claimed result. This completes the proof. □

The theorems in this section illustrate how strong the condition of being subgaussian is, especially when taking sums of i.i.d copies of subgaussians. In the next section, we will investigate another similar tail decay condition.

2.1.2 Subexponential random variables

Definition 2.13 (Subexponential random variable)**.** A random variable X is said to be k-subexponential if for all $t > 0$, the following holds.

$$\mathbb{P}\left(|X| > t\right) \leq 2\exp\left(-\frac{t}{k}\right)$$

◇

Just like in the case of subgaussian random variables, we have a number of equivalent definitions of subexponential random variables.

Lemma 2.14. *Let X be a random variable. Then the following conditions are equivalent.*

(i) X is k-subexponential.

(ii) There exists a $b > 0$ such that $\mathbb{E}\left(\exp(b|X|)\right) \leq 2$ (ψ_1 condition).

(iii) For all $p \geq 1$, $\mathbb{E}\left(|X|^p\right)^{\frac{1}{p}} \leq Cp$ (moment condition).

Moreover, if $\mathbb{E}(X) = 0$, there exists $\lambda_0 > 0$ such that for all $|\lambda| < \lambda_0$, $\mathbb{E}\left(\exp(\lambda X)\right) \leq \exp\left(\widetilde{C}\lambda^2\right)$.

Proof. The proofs of the three equivalences are similar in spirit to the versions for subgaussians, so we will skip the proof, and just prove the moreover part.

Writing $\mathbb{E}(\exp(\lambda X))$ as an infinite sum of expectations, we get the following chain of inequalities for a small enough value of λ.

$$\begin{aligned}
\mathbb{E}(\exp\left(\lambda X\right)) &= 1 + \mathbb{E}(\lambda X) + \sum_{j=2}^{\infty} \frac{\lambda^j \mathbb{E}(X^j)}{j!} \\
&\leq 1 + 0 + \sum_{j=2}^{\infty} \frac{(\lambda C j)^j}{j!} \\
&\leq 1 + \sum_{j=2}^{\infty} (\lambda C e)^j \\
&= 1 + \frac{(\lambda C e)^2}{1 - \lambda C e}
\end{aligned}$$

We get the first inequality from the moment condition, the second from Stirling's approximation, and the third follows from geometric convergence for $\lambda Ce < 1$. Note now that if $\lambda Ce < \frac{1}{2}$, we get the following inequalities.

$$\begin{aligned} 1 + \frac{(\lambda Ce)^2}{1 - \lambda Ce} &\leq 1 + 2C^2e^2\lambda^2 \\ &\leq \exp\left(2C^2e^2\lambda^2\right) \end{aligned}$$

This proves the result. □

We can now prove a strong tail bound for sums of subexponential random variables like we did in Hoeffding-Chernoff-Azuma inequality.

Theorem 2.15 (Bernstein's inequality)**.** *Let $\{X_1, \ldots, X_n\}$ be i.i.d subexponential random variables with mean 0, and let $\mathbf{a}$ be a vector in $\mathbb{R}^n$. Then the following holds.*

$$\mathbb{P}\left(\left|\sum_{j=1}^{n} a_j X_j\right| > t\right) \leq 2\exp\left(-c\left(\frac{t^2}{\|\mathbf{a}\|_2^2} \wedge \frac{t}{\|\mathbf{a}\|_\infty}\right)\right)$$

Here $a \wedge b$ denotes the minimum of a and b.

Remark 2.16. The tail of a sum of i.i.d random variables behaves very much like the situation described above. When t is small, the tail behave like a subgaussian, and when t is large, the tail behaves like a subexponential random variable. ◇

Sketch of proof of Theorem 2.15. Like with all the other sum tail bounds, the proof of this theorem is via bounding the Laplace transform of the random variable. For small λ, we have the upper bound to be $\exp(\widetilde{C}\lambda^2)$ and then we optimize over λ, and for large t, we use Markov's inequality. □

2.1.3 Applications of subgaussian and subexpontial random variables

Before we list some of the applications, we make a remark on why the conditions for subexpontial and subgaussian random variables were called ψ_1 and ψ_2 conditions respectively. Let α be a number greater than 0. Define a function ψ_α in the following manner.

$$\psi_\alpha(x) \coloneqq \exp(x^\alpha) - 1$$

Using this function, we can define norms on random variables.

$$\|X\|_{\psi_\alpha} \coloneqq \inf\left(K > 0 \mid \mathbb{E}\left(\psi_\alpha\left(\frac{|X|}{K}\right)\right) \leq 1\right)$$

With this norm, the space of subexponential and subgaussian random variables form Banach spaces with respect to $\|\cdot\|_{\psi_1}$ and $\|\cdot\|_{\psi_2}$ respectively. These Banach spaces are often known as Orlicz spaces.

Our first application of Hoeffding-Chernoff-Azuma and Bernstein's inequality will be the Johnson-Lindenstrauss lemma*.

*While this lemma was only a small, and rather easy, part of a hard technical paper of Johnson and Lindenstrauss, the lemma is (supposedly) one of the most cited lemmas in computer science.

Theorem 2.17 (Johnson-Lindenstrauss lemma [JL84])**.** *Let $F \subset \mathbb{R}^N$ be a finite set. Then for any $\varepsilon > 0$, there exists a linear mapping $\varphi : F \to \mathbb{R}^n$ with $n \leq \frac{C}{\varepsilon^2} \log(\#F)$ such that the mapping does not distort distances too much, i.e. the following inequalities hold for all x and y in F.*

$$(1-\varepsilon)\|x-y\|_2 \leq \|\varphi(x)-\varphi(y)\|_2 \leq (1+\varepsilon)\|x-y\|_2$$

This is useful for computer scientists because it allow dimension reduction. Often, one has finitely many vectors in a very high dimensional space, and one only cares about their metric structure. This theorem allows one to reduce the ambient dimension significantly while not distorting the metric structure too much, and furthermore, the new ambient dimension is $O(\log \#F)$, whereas creating a metric graph would involve $O(\#F^2)$ computations.

While the ℓ^2 norm is natural for geometry, in computer science applications, the ℓ^1 norm is preferred, because of its relation to linear programming. So a natural follow up question is whether one can perform similar dimension reduction in ℓ^1 instead. This was open for a long time, but recently shown to be impossible (see [BC05]) in a very strong way. It was shown that in order to have at most ε distortion in the ℓ^1 distance, the ambient dimension would be at least $C\#F$, i.e. linear in the size of the dataset, rather than logarithmic. The original proof is this fact was quite long and non-trivial, but Lee and Naor soon gave a simpler proof (see [LN04]) that relied on some highly non-trivial functional analysis. Johnson and Naor also characterized Banach spaces that allow strong dimension reduction, and it turns out those spaces are quite similar to Hilbert spaces (see [JN08]).

Proof of Theorem 2.17. Define a set $V \subset \mathbb{R}^n$ in the following manner.

$$V = \left\{ \frac{x-y}{\|x-y\|_2} \mid \{x,y\} \subset F \text{ and } x \neq y \right\}$$

We will work with this set V instead. V is contained in S^N, and has cardinality $\frac{\#F^2-\#F}{2}$.

Let G be an $n \times N$ matrix with i.i.d subgaussian entries g_{ij} satisfying the following two properties.

$$\mathbb{E}(g_{ij}) = 0$$
$$\mathbb{E}(g_{ij}^2) = 1$$

Fix a point $v \in V$ and let $n = \frac{\theta}{\varepsilon^2}\log(\#F)$, where θ a parameter we'll pick later. We will prove that the following inequality holds with high probability.

$$\left| \|Gv\|_2 - \sqrt{n} \right| \leq \varepsilon$$

If the probability is high enough, we can do this for all elements of V simultaneously using the union bound.

Let $i \in \{1, \ldots, n\}$. Then we define Y_i to be the i^{th} entry of Gv.

$$Y_i = \sum_{j=1}^{N} g_{ij} v_i$$

We then compute the first and second moments of Y_i.

$$\begin{aligned}\mathbb{E}(Y_i) &= 0\\ \mathbb{E}(Y_i^2) &= \sum_{j=1}^{n} v_i^2 \mathbb{E}(g_{ij}^2)\\ &= 1\end{aligned}$$

The random variable Y_i is a linear combination of subgaussian random variable, and thus is also subgaussian. Also, as $i \neq i'$, Y_i and $Y_{i'}$ are i.i.d.

Let us now get estimates on $\|Gv\|_2^2 - n$.

$$\begin{aligned}\|Gv\|_2^2 - n &= \sum_{i=1}^{n} Y_i^2 - n\\ &= \sum_{i=1}^{n} (Y_i^2 - 1)\end{aligned}$$

Let $Z_i = Y_i^2 - 1$. Then $\mathbb{E}(Z_i) = 0$, and Z_i are subexponential. This is a consequence of the fact that squares of subgaussian random variables are subexponential, which can be checked using the tail bound, or the moment condition.

We use Bernstein's inequality to control the tail of the sum of Z_i.

$$\begin{aligned}\mathbb{P}\left(\left|\|Gv\|_2^2 - n\right| > t\right) &= \mathbb{P}\left(\left|\sum_{i=1}^{n} Z_i\right| > t\right)\\ &\leq 2\exp\left(-c\left(\frac{t^2}{n} \wedge t\right)\right)\end{aligned}$$

Set $t = \varepsilon n$, where $\varepsilon < 1$. In this regime, $t^2 < t$, which means the tail bound is $2\exp\left(-c\frac{t^2}{n}\right)$.

We now use the union bound to bound the probability that some $v \in V$ violates this condition.

$$\begin{aligned}\mathbb{P}\left(\exists\ v \in V \mid \left|\|Gv\|_2^2 - n\right| > \varepsilon n\right) &\leq 2\exp\left(-c\varepsilon^2 n\right) \cdot \#F\\ &= 2\exp\left(-c\varepsilon^2 \frac{\theta}{\varepsilon^2}\log(\#F) + \log(\#F)\right)\\ &= \#F^{2(1-c\theta)}\end{aligned}$$

A large enough θ makes the upper bound much smaller than 1.

Thus, with very high probability, $\left|\|Gv\|_2^2 - n\right| \leq \varepsilon n$. We claim that this matrix $\varphi \coloneqq \frac{G}{\sqrt{n}}$ does the dimension reduction with distortion bounded by ε. This can be seen by expanding out the definition of v.

$$\begin{aligned}\varepsilon n &\geq \left|\|Gv\|_2^2 - n\right|\\ &= \left|\frac{\|Gx - Gy\|_2^2}{\|x - y\|_2^2} - n\right|\end{aligned}$$

Dividing both sides by n, we get the inequality we wanted.

$$\varepsilon \geq \left| \frac{\|\varphi x - \varphi y\|_2^2}{\|x-y\|_2^2} - 1 \right|$$

This proves the result. □

2.2 Concentration for quadratic forms

Let $\{X_i\}$ be i.i.d copes of a random variable, and let A be a symmetric matrix associated to a quadratic form. We define a new random variable Y by evaluating the quadratic form A on $\mathbf{X} = (X_1, \ldots, X_n)$.

$$Y \coloneqq \langle \mathbf{X}, A\mathbf{X} \rangle$$

Consider now the singular value decomposition of A, i.e. a decomposition as UDV, where U and V lie in $O(n)$ and D is diagonal. The diagonal entries of D can be arranged to be non-decreasing, i.e. $s_1(A) \geq s_2(A) \geq \cdots \geq s_n(A) \geq 0$. The diagonal entries are called the singular values of A. The j^{th} singular value of A is the square root of the j^{th} largest eigenvalue of $AA^\top$.

We also consider the Frobenius norm of the matrix A.

$$\|A\|_F \coloneqq \left(\sum_{i,j=1}^{n} a_{ij} \right)^{\frac{1}{2}}$$

This is the inner-product norm for the following inner product on matrices.

$$\langle A, B \rangle = \operatorname{tr}(AB^\top)$$

One can express the Frobenius norm of A using the singular values.

$$\|A\|_F^2 = \sum_{i=1}^{n} s_i(A)^2$$

Geometrically, the singular values are the lengths of the axes of the ellipsoid that is the image under A of the standard sphere in $\mathbb{R}^n$. Note that one can also express the operator norm in terms of the singular values: it's the sup norm on the singular values, just like the Frobenius norm is the ℓ^2-norm on the singular values.

We can now state a concentration inequality for quadratic forms.

Theorem 2.18 (Hanson-Wright inequality ([HW71] and [Wri73]))**.** *Let A be any $n \times n$ matrix, and let $\{X_1, \ldots, X_n\}$ be i.i.d subgaussian random variables with $\mathbb{E}(X_i) = 0$. Let $\mathbf{X} = (X_1, \ldots, X_n) \in \mathbb{R}^n$. Then for any $t > 0$:*

$$\mathbb{P}\left(|\langle \mathbf{X}, A\mathbf{X} \rangle - \mathbb{E}\left(\langle \mathbf{X}, A\mathbf{X} \rangle \right)| > t \right) \leq 2 \exp\left(-c \left(\frac{t^2}{\|A\|_F^2} \wedge \frac{t}{\|A\|_{\text{op}}} \right) \right)$$

Remark 2.19. Note that we can only expect a tail like we had in Bernstein's inequality, since entries of $\langle \mathbf{X}, A\mathbf{X} \rangle$ be squares of subgaussians, i.e. subexponential. ◇

Remark 2.20. The original papers of Hanson and Wright proved a slightly weaker version of the above theorem. The version stated is from 2013, and appeared in [RV13]. ◇

Proof of Theorem 2.18. We first decompose A as the sum of a diagonal matrix A_{diag}, and a matrix $\widetilde{A}$ with all diagonal entries equal to 0.

$$A = A_{\text{diag}} + \widetilde{A}$$

Since we are trying to bound the probability that $\langle \mathbf{X}, A\mathbf{X}\rangle$ deviates by more than t from its mean, it will suffice to bound the probability that $\langle \mathbf{X}, A_{\text{diag}}\mathbf{X}\rangle$ deviates by more than $\frac{t}{2}$ and the probability that $\langle \mathbf{X}, \widetilde{A}\mathbf{X}\rangle$ deviates by more than $\frac{t}{2}$.

$$\begin{aligned}\mathbb{P}\left(\left|\langle \mathbf{X}, A\mathbf{X}\rangle - \mathbb{E}\left(\langle \mathbf{X}, A\mathbf{X}\rangle\right)\right| > t\right) \leq \mathbb{P}&\left(\left|\langle \mathbf{X}, A_{\text{diag}}\mathbf{X}\rangle - \mathbb{E}\left(\langle \mathbf{X}, A_{\text{diag}}\mathbf{X}\rangle\right)\right| > \frac{t}{2}\right)\\ &+ \mathbb{P}\left(\left|\langle \mathbf{X}, \widetilde{A}\mathbf{X}\rangle - \mathbb{E}\left(\langle \mathbf{X}, \widetilde{A}\mathbf{X}\rangle\right)\right| > \frac{t}{2}\right)\end{aligned}$$

Note that since $\widetilde{A}$ has zeroes on the diagonal, and the X_is are independent with mean 0, that means $\mathbb{E}\left(\langle \mathbf{X}, \widetilde{A}\mathbf{X}\rangle\right) = 0$.

Observe that the first term is a tail bound on a sum of subexponential random variables.

$$\langle \mathbf{X}, A_{\text{diag}}\mathbf{X}\rangle - \mathbb{E}\left(\langle \mathbf{X}, A_{\text{diag}}\mathbf{X}\rangle\right) = \sum_{i=1}^{n} a_{ii}\left(X_i^2 - \mathbb{E}(X_i^2)\right)$$

Since X_i are subgaussian random variables, $X_i^2 - \mathbb{E}(X_i^2)$ are subexponential random variables with mean 0, which means we can apply Bernstein's inequality.

$$\mathbb{P}\left(\left|\langle \mathbf{X}, A_{\text{diag}}\mathbf{X}\rangle - \mathbb{E}\left(\langle \mathbf{X}, A_{\text{diag}}\mathbf{X}\rangle\right)\right| > \frac{t}{2}\right) \leq 2\exp\left(-c\left(\frac{t^2}{\sum_{i=1}^n a_{ii}^2} \wedge \frac{t}{\max(a_{ii})}\right)\right)$$

Note that we can bound the denominators using the appropriate matrix norm.

$$\begin{aligned}\sum_{i=1}^{n} a_{ii}^2 &\leq \|A\|_F^2\\ \max(a_{ii}) &\leq \|A\|\end{aligned}$$

Note that the upper bound for the diagonal term is of the form stated in the theorem, which means that all we need to do prove a similar or better bound for the second term, which involves $\widetilde{A}$. That boils down to finding a probability bound for the following event. For convenience of notation, we will now denote $\widetilde{A}$ by just A, where it is understood that A is a matrix with all diagonal entries equal to 0. We will also denote $\langle \mathbf{X}, A\mathbf{X}\rangle$ by Y.

$$\mathbb{P}\left(|Y| > t\right) \leq 2\exp\left(-c\left(\frac{t^2}{\|A\|_F^2} \wedge \frac{t}{\|A\|_{\text{op}}}\right)\right)$$

Recall that when we proved Bernstein's inequality, we did so bounding the Laplace transform of Y, i.e. getting upper bounds for $\mathbb{E}\left(\exp\left(\lambda Y\right)\right)$. Since Y in our case is a quadratic function

in independent random variables X_i, estimating the Laplace transform is a little tricky, and we need to use *decoupling*.

Introduce new random variables $\{\delta_1, \ldots, \delta_n\}$ which are independent from each other, as well as all X_i, and are distributed like Bernoulli $\left(\frac{1}{2}\right)$. We then have the following.

$$\mathbb{E}\left(\exp(\lambda Y)\right) = \mathbb{E}\left(\exp\left(4\lambda \sum_{i,j=1}^{n} \mathbb{E}_\delta\left(\delta_i(1-\delta_j)\right) a_{ij} X_i X_j\right)\right)$$

Here, $\mathbb{E}_\delta$ represents taking expectation over the sample space of the δ_i, and $\mathbb{E}$ represents taking the expectation over the sample space of X_i. Note that the equality holds because for $i = j$, $\delta_i(1-\delta_j) = 0$. We now use Jensen's inequality to pull out the $\mathbb{E}_\delta$.

$$\mathbb{E}\left(\exp(\lambda Y)\right) \leq \mathbb{E}\mathbb{E}_\delta\left(\exp\left(4\lambda \sum_{i,j=1}^{n} \left(\delta_i(1-\delta_j)\right) a_{ij} X_i X_j\right)\right)$$

Let I be the random subset of $\{1, \ldots, n\}$ where $i \in I$ iff $\delta_i = 1$. We can condition the above expectation on the value of I to simplify things.

$$\mathbb{E}\left(\exp(\lambda Y)\right) \leq \mathbb{E}_\delta \mathbb{E}\left(\exp\left(4\lambda \sum_{i\in I} \sum_{j\notin I} a_{ij} X_i X_j\right) \,\middle|\, I\right)$$

We can simplify the inner integral using the fact that I is fixed.

$$\mathbb{E}\left(\exp\left(4\lambda \sum_{i\in I} \sum_{j\notin I} a_{ij} X_i X_j\right) \,\middle|\, I\right) = \prod_{j\notin I} \mathbb{E}\exp\left(\left[4\lambda \sum_{i\in I} a_{ij} X_i\right] \cdot X_j\right)$$

We now use the fact that each X_j is subgaussian with mean 0 to bound the right hand side in the following manner.

$$\prod_{j\notin I} \mathbb{E}\exp\left(\left[4\lambda \sum_{i\in I} a_{ij} X_i\right] \cdot X_j\right) \leq \prod_{j\notin I} \mathbb{E}\left(\exp\left[16C\lambda^2 \left(\sum_{i\in I} a_{ij} X_i\right)^2\right]\right) \tag{2.21}$$

We now use a trick to turn the expectation of the quadratic form as in (2.21) to expectation of a bilinear form. Note that for a standard normal random variable g, one has the following explicit formula for the Laplace transform.

$$\mathbb{E}\left(\theta g\right) = \exp\left(\frac{\theta^2}{2}\right)$$

Let $\{g_1, \ldots, g_n\}$ be i.i.d standard normal random variables that are independent of X_i and δ_i. We get that the right hand side of (2.21) is the following.

$$\prod_{j\notin I} \mathbb{E}\left(\exp\left[16C\lambda^2 \left(\sum_{i\in I} a_{ij} X_i\right)^2\right]\right) = \mathbb{E}\left(\exp\left(C'\lambda \sum_{j\notin I} \sum_{i\in I} a_{ij} X_i g_j\right)\right)$$

Repeating this whole process again, with X_i rather than X_j, we end up with the following upper bound.

$$\mathbb{E}\left(\exp\left(4\lambda\sum_{i\in I}\sum_{j\notin I}a_{ij}X_iX_j\right)\,\middle|\, I\right)\leq\mathbb{E}\left(\exp\left(C''\lambda^2\sum_{i\in I}\left(\sum_{j\notin I}a_{ij}g_j\right)^2\right)\,\middle|\, I\right)$$

Note that the upper bound is independent of X_i, and only depends on the Bernoulli random variables δ_i and the normal random variables g_i.

We will now use that fact that $\mathbf{g}=(g_1,\ldots,g_n)$ is a rotation invariant random vector in $\mathbb{R}^n$ to estimate the upper bound. Let P_I be the projection to the subspace spanned by $\{e_i\}_{i\in I}$, and let $B_I=P_IA(\mathrm{Id}-P_I)$. Then the innermost double sum can be expressed as a norm.

$$\sum_{i\in I}\left(\sum_{j\notin I}a_{ij}g_j\right)^2=\|B_I\mathbf{g}\|_2^2$$

This means we need to estimate the following conditional expectation.

$$\mathbb{E}\left(C''\lambda^2\|B_I\mathbf{g}\|_2^2\,\middle|\, I\right)$$

Let $U_ID_IV_I$ be the singular value decomposition of B_I. Since $\|\cdot\|_2$ is invariant under $O(n)$ action, and the distribution of $\mathbf{g}$ is also $O(n)$ invariant, $B_I\mathbf{g}$ has the same distribution as $D_I\mathbf{g}$. This means we really just need to understand the distribution of the singular values of B_I.

$$\begin{aligned}\mathbb{E}\left(\exp\left(C''\lambda^2\|B_I\mathbf{g}\|_2^2\right)\,\middle|\, I\right)&=\mathbb{E}\left(\exp\left(C''\lambda^2\|D_I\mathbf{g}\|_2^2\right)\,\middle|\, I\right)\\&=\mathbb{E}\left(\exp\left(C''\lambda^2\sum_{j=1}^n s_j^2(B_I)g_j^2\right)\,\middle|\, I\right)\\&=\prod_{j=1}^n\mathbb{E}\left(\exp\left(C''\lambda^2s_j^2(B_I)g_j^2\right)\,\middle|\, I\right)\end{aligned}$$

One can explicitly compute $\mathbb{E}(\exp(\theta g^2))$ for a normal random variable g for $\theta\in\left[0,\frac{1}{2}\right]$.

$$\mathbb{E}(\exp\left(\theta g^2\right))=\frac{1}{\sqrt{1-2\theta}}$$

Hence,

$$\mathbb{E}\left(C''\lambda^2\|B_I\mathbf{g}\|_2^2\,\middle|\, I\right)=\prod_{j=1}^n\left(1-2C''\lambda^2s_j^2(B_I)\right)^{-\frac{1}{2}}$$

For $x\in\left[0,\frac{1}{2}\right]$, $\frac{1}{\sqrt{1-x}}\leq e^x$. This lets us bound the right hand side in the following manner.

$$\prod_{j=1}^n\left(1-2C''\lambda^2s_j^2(B_I)\right)^{-\frac{1}{2}}\leq\prod_{j=1}^n\exp\left(2C''\lambda^2s_j^2(B_I)\right)$$

Note however that to use the simplification, we required $2C''\lambda^2 s_j^2(B_I) \leq \frac{1}{2}$. If we pick a λ smaller than $\frac{c'''}{s_1(B_I)}$, then the inequalities are valid. That ends up giving the following bound.

$$\mathbb{E}\left(\exp\left(C\lambda \left\|B_I\mathbf{g}\right\|_2^2\right)\right) \leq \exp\left(\widetilde{C}\lambda^2 \left\|B_I\right\|_F^2\right)$$

This is valid when $\lambda \leq \frac{c'''}{s_1(B_I)} = \frac{c'''}{\|B_I\|}$. Finally, we can get rid of the dependence on I using the following inequalities involving matrix norms.

$$\begin{aligned}\left\|B_I\right\|_F &= \left\|P_I A(\mathrm{Id} - P_I)\right\|_F \\ &\leq \left\|A\right\|_F \\ \left\|B_I\right\| &= \left\|P_I A(\mathrm{Id} - P_I)\right\| \\ &\leq \left\|A\right\|\end{aligned}$$

We can now conclude our estimates. We get the following inequality for all λ less than $\frac{c'''}{\|A\|}$.

$$\begin{aligned}\mathbb{E}\left(\exp(\lambda Y)\right) &\leq \mathbb{E}_\delta \mathbb{E}\left(\exp\left(C\lambda \left\|B_I\mathbf{g}\right\|_2^2\right)\right) \\ &\leq \exp\left(\widetilde{C}\lambda^2 \left\|B_I\right\|_F^2\right) \\ &\leq \exp\left(\widetilde{C}\lambda^2 \left\|A\right\|_F^2\right)\end{aligned}$$

We are able to integrate easily over the δ sample space because the right hand side does not depend on I at all. The rest of the proof follows exactly like the end of Bernstein's inequality, where we optimized over λ to get the best bound via Markov's inequality. □

We now show a rather unexpected application of the Hanson-Wright inequality.

Theorem 2.22. *Let A be an $n \times n$ matrix, and let $\mathbf{X} \in \mathbb{R}^n$ be a random vector with i.i.d subgaussian coordinates with mean 0 and variance 1. Then for any $\tau > 0$, the following holds.*

$$\mathbb{P}\left(\left|\left\|A\mathbf{X}\right\|_2 - \left\|A\right\|_F\right| > \tau\right) \leq 2\exp\left(-c\frac{\tau^2}{\left\|A\right\|^2}\right)$$

Remark 2.23. This result is surprising for two reasons: first, we will prove this using the Hanson-Wright inequality, but the tail bound in this inequality is better than the tail bound in Hanson-Wright. Second, all of our earlier concentration inequalities were concentration inequalities about the mean, but in this case $\mathbb{E}\left(\left\|A\mathbf{X}\right\|_2\right) \neq \left\|A\right\|_F$. ◇

Proof. Let $Y = \left\|A\mathbf{X}\right\|_2^2 = \langle \mathbf{X}, A^\top A\mathbf{X}\rangle$. This expression as the quadratic form $A^\top A$ will let us apply the Hanson-Wright inequality. Using the Hanson-Wright inequality, we obtain the following.

$$\mathbb{P}\left(\left|\langle \mathbf{X}, A^\top A\mathbf{X}\rangle - \mathbb{E}\left(\langle \mathbf{X}, A^\top A\mathbf{X}\rangle\right)\right| > t\right) \leq 2\exp\left(-c\left[\frac{t^2}{\left\|A^\top A\right\|_F^2} \wedge \frac{t}{\left\|A^\top A\right\|}\right]\right)$$

We can explicitly evaluate $\mathbb{E}\left(\langle \mathbf{X}, A^\top A\mathbf{X}\rangle\right)$: since the entries of $\mathbf{X}$ are i.i.d with mean 0 and variance 1, the expectation simplifies to $\mathrm{tr}(A^\top A) = \|A\|_F^2$. We also have these inequalities involving products of matrices.

$$\|AB\|_F \leq \|A\| \cdot \|B\|_F$$
$$\|AB\| \leq \|A\| \cdot \|B\|$$

Using these two inequalities, we can simplify the inequality we got from the application of Hanson-Wright inequality.

$$\mathbb{P}\left(\left|\|A\mathbf{X}\|_2^2 - \|A\|_F^2\right| > t\right) \leq 2\exp\left(-c\left[\frac{t^2}{\|A\|^2 \cdot \|A\|_F^2} \wedge \frac{t}{\|A\|^2}\right]\right)$$

Let $t = \varepsilon \cdot \|A\|_F^2$. Then the above inequality can be rewritten as follows.

$$\mathbb{P}\left(\left|\frac{\|A\mathbf{X}\|_2^2}{\|A\|_F^2} - 1\right| > \varepsilon\right) \leq 2\exp\left(-c\frac{\|A\|_F^2}{\|A\|^2}\left[\varepsilon^2 \wedge \varepsilon\right]\right)$$

Rewriting the inequality in this manner makes it clearer which of the terms in the upper bound is smaller. The analysis splits into two cases.

Case 1 ($\varepsilon \leq 1$)**:** In this case, $\varepsilon^2 \leq \varepsilon$, so the tail bound simplifies in the following manner.

$$\mathbb{P}\left(\left|\frac{\|A\mathbf{X}\|_2^2}{\|A\|_F^2} - 1\right| > \varepsilon\right) \leq 2\exp\left(-c\frac{\|A\|_F^2}{\|A\|^2}\varepsilon^2\right)$$

Observe now that for positive values of x, $|x^2 - 1| \geq |x - 1|$. This means that we can replace the $\frac{\|A\mathbf{X}\|_2^2}{\|A\|_F^2}$ in left hand side of the above inequality by its square root. We do that, and set $\tau = \varepsilon \|A\|_F$.

$$\mathbb{P}\left(|\,\|A\mathbf{X}\|_2 - \|A\|_F\,| > \tau\right) \leq 2\exp\left(-c\frac{\tau^2}{\|A\|^2}\right)$$

This proves the inequality for all $\tau \in [0, \|A\|_F]$.

Case 2 ($\varepsilon > 1$)**:** In this case $\varepsilon \leq \varepsilon^2$, so the tail bound simplifies in the following manner.

$$\mathbb{P}\left(\left|\frac{\|A\mathbf{X}\|_2^2}{\|A\|_F^2} - 1\right| > \varepsilon\right) \leq 2\exp\left(-c\frac{\|A\|_F^2}{\|A\|^2}\varepsilon\right)$$

Note again that for positive values of x, $|x^2 - 1| \geq |x - 1|^2$. We use this fact to make a substitution on the left hand side, and let $\tau = \sqrt{\varepsilon}\,\|A\|_F$.

$$\mathbb{P}\left(|\,\|A\mathbf{X}\|_2 - \|A\|_F\,| > \tau\right) \leq 2\exp\left(-c\frac{\|A\|_F^2}{\|A\|^2}\varepsilon\right)$$

This proves in the inequality for $\tau \in (\|A\|_F, \infty)$.

□

2.3 Concentration for matrix-valued random variables

So far, we have looked at concentration inequalities for functions acting on vector-valued random variables, more specifically, linear and quadratic functions on vectors of random variables. In this section, we will look at concentration inequalities for matrix valued random variables. We will start by proving the matrix Bernstein inequality.

Theorem 2.24 (Matrix Bernstein Inequality)**.** *Let $\{X_1, \dots, X_N\}$ be independent $n \times n$ symmetric random matrices satisfying the following conditions for some constant K independent of j.*

$$\begin{aligned} \mathbb{E}(X_j) &= 0 \\ \|X_j\|_\infty &\leq K \end{aligned}$$

Let σ^2 denote the following quantity.

$$\sigma^2 \coloneqq \left\| \sum_{j=1}^{N} \mathbb{E}(X_j^2) \right\|$$

Then for any $t > 0$, we have the following tail bound on the sum of X_j.

$$\mathbb{P}\left(\left\| \sum_{j=1}^{N} X_j \right\| > t \right) \leq 2n \exp\left(-c \left(\frac{t^2}{\sigma^2} \wedge \frac{t}{K} \right) \right)$$

Remark 2.25. Note that this inequality looks very similar to the original Bernstein inequality: the only difference is that the coefficient multiplied with exp in this case is $2n$, rather than 2. This means that the bound is not dimension independent, and that does cause issues in practice. However, the bound is optimal. ◇

To prove Theorem 2.24, we will need the following linear algebraic result of Lieb.

Theorem 2.26 (Lieb's concavity theorem [Lie73])**.** *Let H be an $n \times n$ symmetric matrix. Consider the function f defined on the set of $n \times n$ symmetric positive-definite matrices.*

$$f(X) \coloneqq \operatorname{tr}(\exp(H + \log(X)))$$

The function f is concave, i.e. for any $t \in [0, 1]$, and any X and Y in the space of symmetric positive-definite matrices, the following inequality holds.

$$f(tX + (1-t)Y) \geq tf(X) + (1-t)f(Y)$$

Using Lieb's concavity theorem and Jensen's inequality, we get the following corollary.

Corollary 2.27. *Let H be an $n \times n$ symmetric matrix, and let Y be a random symmetric matrix. Then the following inequality holds.*

$$\mathbb{E}(\operatorname{tr}(\exp(H + Y))) \leq \operatorname{tr}(\exp(H + \log(\mathbb{E}(\exp(Y))))) \tag{2.28}$$

Using this corollary, we can prove the matrix Bernstein inequality.

Proof of Theorem 2.24. We begin by giving an easy upper bound for the norm of a symmetric matrix X.

$$\|X\| \leq \lambda_{\max}(X) + \lambda_{\max}(-X)$$

Here, $\lambda_{\max}(X)$ denotes the largest eigenvalue of X. Expressed in terms of tail bounds, we get the following.

$$\mathbb{P}\left(\left\|\sum_{j=1}^{N} X_j\right\| > t\right) \leq \mathbb{P}\left(\lambda_{\max}\left(\sum_{j=1}^{N} X_j\right) > \frac{t}{2}\right) + \mathbb{P}\left(\lambda_{\max}\left(\sum_{j=1}^{N} -X_j\right) > \frac{t}{2}\right)$$

Note that it will suffice to get an upper bound for one of the terms on the right hand side, and multiply that by 2.

Just like in the proof of Bernstein's inequality, it will help to bound the Laplace transformation of the random variables involved. To do so, we need to chain several simplifying equalities and inequalities. The first equality we get from the fact that the largest eigenvalue of the exponential of a symmetric matrix is the exponential of the largest eigenvalue of the original matrix.

$$\mathbb{E}\left(\exp\left(\lambda\lambda_{\max}\left(\sum_{j=1}^{N} X_j\right)\right)\right) = \mathbb{E}\left(\lambda_{\max}\exp\left(\lambda\left(\sum_{j=1}^{N} X_j\right)\right)\right) \tag{2.29}$$

Here, λ is an arbitrary positive number. Now note that exp of a symmetric matrix has only positive eigenvalues. This means that the largest eigenvalue is smaller than the trace.

$$\mathbb{E}\left(\lambda_{\max}\exp\left(\lambda\left(\sum_{j=1}^{N} X_j\right)\right)\right) \leq \mathbb{E}\left(\operatorname{tr}\exp\left(\lambda\left(\sum_{j=1}^{N-1} X_j + X_N\right)\right)\right) \tag{2.30}$$

We now condition on the values of $\{X_1, \ldots, X_{N-1}\}$, which lets us treat the right hand side of (2.30) using (2.28).

$$\mathbb{E}\left(\operatorname{tr}\exp\left(\lambda\left(\sum_{j=1}^{N-1} X_j + X_N\right)\right)\right) \leq \mathbb{E}\left(\operatorname{tr}\exp\left(\lambda\left(\sum_{j=1}^{N-1} X_j\right) + \log\mathbb{E}\exp\left(\lambda X_N\right)\right)\right)$$

We repeat this process $N-2$ additional times, conditioning on all but the last X_i to get the following inequality, combining (2.29), (2.30), and (2.28).

$$\mathbb{E}\left(\exp\left(\lambda\lambda_{\max}\left(\sum_{j=1}^{N} X_j\right)\right)\right) \leq \operatorname{tr}\left(\exp\left[\sum_{j=1}^{N}\log\mathbb{E}\exp\left(\lambda X_j\right)\right]\right) \tag{2.31}$$

We now deal with $\mathbb{E}\exp(\lambda X_j)$. Note that for $t \in [0,1]$, we have the following elementary inequality.

$$\exp(t) \leq 1 + t + t^2$$

We can extend this to symmetric matrices, where the inequality $A \leq B$ is understood to mean that $B - A$ is positive definite.

$$\begin{aligned}\mathbb{E}\exp(\lambda X_j) &\leq I + \mathbb{E}(\lambda X_j) + \mathbb{E}\left(\lambda^2 X_j^2\right) \\ &= I + \lambda^2 \mathbb{E}(X_j^2)\end{aligned}$$

Note that the above inequality holds when λ is small enough such that the largest eigenvalue of λX_j is less than 1. Now we use the inequality $1 + x \leq \exp(x)$ to get an upper bound of $\exp\left(\lambda^2 \mathbb{E}(X_j^2)\right)$. We combine this with (2.31) to get the following.

$$\mathbb{E}\left(\exp\left(\lambda \lambda_{\max}\left(\sum_{j=1}^{N} X_j\right)\right)\right) \leq \operatorname{tr}\left(\exp\left[\sum_{j=1}^{N} \lambda^2 \mathbb{E}(X_j^2)\right]\right) \tag{2.32}$$

Now we again use the fact that the exponential of a symmetric matrix is positive definite to bound the trace above by n times the largest eigenvalue.

$$\mathbb{E}\left(\exp\left(\lambda \lambda_{\max}\left(\sum_{j=1}^{N} X_j\right)\right)\right) \leq n\lambda_{\max}\left(\exp\left[\sum_{j=1}^{N} \lambda^2 \mathbb{E}(X_j^2)\right]\right) \tag{2.33}$$

Finally, we use the fact that $\lambda_{\max} X$ is less than or equal to $\|X\|$ to get the upper bound in terms of matrix norm.

$$\begin{aligned}\mathbb{E}\left(\exp\left(\lambda \lambda_{\max}\left(\sum_{j=1}^{N} X_j\right)\right)\right) &\leq n\left(\exp\left[\lambda^2 \left\|\sum_{j=1}^{N} \mathbb{E}(X_j^2)\right\|\right]\right) &(2.34)\\ &\leq n\left(\exp\left[\lambda^2 \sigma^2\right]\right) &(2.35)\end{aligned}$$

This gives a bound on the Laplace transformation of the random variable we were trying to tail bound. We now use Markov's inequality to get the tail bound.

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{j=1}^{N} X_j\right) > \tau\right) \leq n \exp\left(\lambda^2 \sigma^2 - \lambda \tau\right)$$

Note that this bound comes from (2.34), which is only valid when $\lambda \max \|X_j\| \leq 1$. We optimize over the values of λ now. This calculation is identical to the one we did for the original Bernstein inequality, and the tail bound follows. □

We now discuss a few corollaries of the matrix Bernstein inequality. The first corollary pertains to matrices that need not be symmetric, or even square.

Corollary 2.36. *Let $\{X_1, \ldots, X_N\}$ be $m \times n$ matrices, satisfying the following conditions for some constant K.*

$$\begin{aligned}\mathbb{E}(X_j) &= 0 \\ \|X_j\|_\infty &\leq K\end{aligned}$$

Let $\sigma^2 = \left\|\sum_{j=1}^N \mathbb{E}\left(X_j X_j^\top\right)\right\| + \left\|\sum_{j=1}^N \mathbb{E}\left(X_j^\top X_j\right)\right\|$. *Then we have the following tail bound.*

$$\mathbb{P}\left(\left\|\sum_{j=1}^N X_j\right\| > t\right) \leq 2(n+m)\exp\left(-c\left(\frac{t^2}{\sigma^2} \wedge \frac{t}{K}\right)\right)$$

Proof. The proof follows from Theorem 2.24 after symmetrizing the rectangular matrices. Consider the $(n+m)\times(n+m)$ matrix X_j^s, constructed in the following manner.

$$X_j^s = \begin{pmatrix} 0 & X_j \\ X_j^\top & 0 \end{pmatrix}$$

Applying Theorem 2.24 to $\{X_1^s, \ldots, X_N^s\}$ gives us the proof. □

Another easy corollary is getting expectation bounds from tail bounds.

Corollary 2.37. *Let* $\{X_1, \ldots, X_N\}$ *be as in Theorem 2.24. Then we have the following expectation bound.*

$$\mathbb{E}\left\|\sum_{j=1}^N X_j\right\| \leq C\left(\sigma\sqrt{\log(n)} + K\log(n)\right)$$

We now return to the problem of covariance estimation that we mentioned in the introduction.

Empirical covariance estimation Let X be an $\mathbb{R}^n$-valued random variable with mean 0. The population covariance Σ is defined to be $\mathbb{E}XX^\top$. Let $\{X_1, \ldots, X_N\}$ be i.i.d copies of X. We define sample covariance Σ_N to be the sample average of the covariance of each X_j.

$$\Sigma_N \coloneqq \frac{\sum_{j=1}^N X_j X_j^\top}{N}$$

By the law of large numbers, Σ_N converges to the population covariance. However, it's more useful to know how fast Σ_N converges to Σ.

Theorem 2.38. *Let* X *and* $\{X_1, \ldots, X_N\}$ *be random vectors as described previously. Suppose that the following inequality holds almost surely for some constant* S*.*

$$\|X\|_2^2 \leq S\mathbb{E}\|X\|_2^2$$

Let $\varepsilon \in (0,1)$*, and set* N *to be the following.*

$$N = C\frac{Sn\log(n)}{\varepsilon^2}$$

Here, C *is some fixed constant, and we round* N *to the nearest integer. Then we have the following concentration inequality for sample covariance.*

$$\mathbb{E}\|\Sigma_N - \Sigma\| \leq \varepsilon \cdot \|\Sigma\|$$

Proof. Define Y_j to be $X_j X_j^\top - \Sigma$. The expectation of Y_j is clearly 0. We can express the left hand side of the inequality we are trying to prove in terms of Y_j.

$$\begin{aligned}\mathbb{E}\left\|\Sigma_N - \Sigma\right\| &= \mathbb{E}\left\|\frac{\sum_{j=1}^N X_j X_j^\top}{N} - \Sigma\right\| \\ &= \frac{1}{N}\mathbb{E}\left\|\sum_{j=1}^N Y_j\right\|\end{aligned}$$

By Corollary 2.37, we can bound the above expectation.

$$\mathbb{E}\left\|\sum_{j=1}^N Y_j\right\| \leq C\left(\sigma\sqrt{\log(n)} + K\log(n)\right) \tag{2.39}$$

We need to compute what σ and K are in this context. Recall that K was an almost sure upper bound on $\|Y_j\|$.

$$\begin{aligned}\|Y_j\| &= \left\|X_j X_j^\top - \Sigma\right\| \\ &\leq \|X_j\|^2 + \|\Sigma\|\end{aligned}$$

We now use the hypothesis we had.

$$\begin{aligned}\|X_j\|^2 &\leq S\mathbb{E}\|X\|_2^2 \\ &= S\mathbb{E}\left(\operatorname{tr}(X_j X_j^\top)\right) \\ &= S\operatorname{tr}(\Sigma) \\ &\leq Sn\|\Sigma\|\end{aligned}$$

Thus, K in this context in $Sn\|\Sigma\|$.

To compute σ, we do something similar.

$$\begin{aligned}\mathbb{E}Y_j^2 &= \mathbb{E}\left(X_j X_j^\top - \Sigma\right)^2 \\ &\leq \mathbb{E}\left(X_j X_j^\top\right)^2 \\ &\leq \|X_j\|_2^2 \cdot \mathbb{E}\left(X_j X_j^\top\right) \\ &\leq Sn\|\Sigma\|\,\Sigma\end{aligned}$$

The quantity σ^2 then works out to be less than $NSn\|\Sigma\|^2$. Plugging these values into (2.39) gives us the result we want. □

3 Martingales

3.1 Azuma's martingale inequality

We begin by recalling the definition of a martingale.

Definition 3.1 (Martingale)**.** Let $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots \subset \mathcal{F}_n$ be a filtration of sigma algebras of the sample space Ω, with $\mathcal{F}_0 = \{\emptyset, \Omega\}$. A sequence of random variables $\{M_1, M_2, \ldots\}$ is said to be a martingale adapted with respect to the filtration $\{cF_i\}$ if the following holds for all i and j, where $j > i$.

$$\mathbb{E}\left(M_j \mid \ \mathcal{F}_i\right) = M_i$$

$\diamond$

Using the notion of a martingale, we can state a generalization of a special case of Hoeffding's inequality.

Theorem 3.2 (Azuma's martingale inequality)**.** *Let $\{M_j\}$ be a martingale. Assume that for all $j \geq 1$, there exist constants d_j such that the following holds.*

$$\|M_j - M_{j-1}\|_\infty \leq d_j$$

Then for any $t > 0$, we have the following concentration inequality for M_n.

$$\mathbb{P}\left(|M_n - \mathbb{E}M_n| > t\right) \leq 2\exp\left(-\frac{t^2}{\sum_{j=1}^n d_j^2}\right)$$

To see how this is a generalization of a special case of Hoeffding's inequality, let $\{B_i\}$ be i.i.d copies of a random variable with zero mean whose absolute value is bounded by 1 almost surely. If we define M_n by $\sum_{j=1}^n d_j B_j$, then it's easy to verify that $\{M_n\}$ is a martingale satisfying the above hypotheses, and the tail bound in this case follows from Hoeffding's inequality.

The reason why the martingale variant of the inequality requires more work is that it is not the case that all martingales are sums of independent random variables. However, it turns out that a lot of results that hold for sums of independent random variables can also be made to work for martingales.

Proof of Theorem 3.2. We prove this concentration inequality again by estimating the Laplace transform of M_n. Let $\lambda > 0$, and define f in the following manner.

$$f(\lambda) = \mathbb{E}(\exp(\lambda M_n))$$

To estimate the Laplace transform, we split up M_n in an artificial manner.

$$\mathbb{E}(\exp(\lambda M_n)) = \mathbb{E}(\exp(\lambda M_{n-1} + \lambda(M_n - M_{n+1})))$$

Observe that since M_{n-1} and $M_n - M_{n-1}$ are not necessarily independent, we cannot turn this into a product of expectations. But we do know that the expectation of $M_n - M_{n-1}$ when conditioned on M_{n-1} is 0.

$$\mathbb{E}(\exp(\lambda M_{n-1} + \lambda(M_n - M_{n+1}))) = \mathbb{E}\left(\exp\left(\lambda M_{n-1}\right)\right) \cdot \mathbb{E}\left(\exp(\lambda(M_n - M_{n-1})) \mid \ M_{n-1}\right)$$

We deal with the second term in the product first. To do, so consider the following elementary inequality, which we will state without proof. For all $x \in \mathbb{R}$, the following holds.

$$e^x \leq x + e^{x^2}$$

We use this to bound the second term of the product as follows.

$$\begin{aligned}\mathbb{E}\left(\exp(\lambda(M_n - M_{n-1})) \mid\ M_{n-1}\right) \leq\ & \mathbb{E}\left(\lambda(M_n - M_{n-1}) \mid\ M_{n-1}\right) \\ & + \mathbb{E}\left(\exp\left(\lambda^2(M_n - M_{n-1})^2\right) \mid\ M_{n-1}\right)\end{aligned}$$

Observe that the first term vanishes, and the second term is bounded above by $\exp(\lambda^2 d_j^2)$.

To deal with the $\mathbb{E}\left(\exp\left(\lambda M_{n-1}\right)\right)$ term, we do the same thing to get an upper bound of $\exp(\lambda d_{j-1}^2)$. We repeat this process inductively, to get the following bound.

$$\begin{aligned}\mathbb{E}\left(\exp\left(\lambda M_n\right)\right) &\leq \exp\left(\lambda^2\left(\sum_{j=1}^{n} d_j^2\right)\right) \cdot \mathbb{E}\left(\lambda M_0\right) \\ &= \exp\left(\lambda^2\left(\sum_{j=1}^{n} d_j^2\right)\right) \cdot \mathbb{E}\left(\lambda M_n\right)\end{aligned}$$

Using Markov's inequality, and optimizing over λ, the inequality follows. □

An application of Azuma's martingale inequality is the *average bounded differences inequality.*

Corollary 3.3 (Average bounded differences inequality)**.** *Let $\{X_1, \ldots, X_n\}$ be independent real valued random variables. Let f be a function from $\mathbb{R}^n$ to $\mathbb{R}$ such that for any $j \in \{1, 2, \ldots, n-1\}$, and any $\{y_1, \ldots, y_{j+1}\}$ in $\mathbb{R}$, the following holds for some $d_{j+1} \geq 0$.*

$$\left|\mathbb{E}f\left(y_1, \ldots, y_{j+1}, X_{j+2}, \ldots, X_m\right) - \mathbb{E}f\left(y_1, \ldots, y_j, X_{j+1}, \ldots, X_m\right)\right| \leq d_{j+1}$$

Then for any $t > 0$, the following tail bound applies to $f(X_1, \ldots, X_n)$.

$$\mathbb{P}\left(\left|f(X_1, \ldots, X_n) - \mathbb{E}f(X_1, \ldots, X_n)\right| > t\right) \leq 2\exp\left(-\frac{t^2}{\sum_{j=1}^{n} d_j^2}\right)$$

Remark 3.4. The strength of this theorem comes from the fact that we require very little from the function f. For instance, if f is merely continuous, and the X_i are bounded random variables, then this inequality applies, and gives us a strong concentration result. ◇

Proof of Corollary 3.3. Let $\mathcal{F}_j$ be the σ-algebra generated by X_1 to X_j. Define the martingale M_j in the following manner.

$$M_j = \mathbb{E}\left(f(X_1, \ldots, X_n) \mid\ \mathcal{F}_j\right)$$

It is easy to check that the sequence of M_j's actually forms a martingale. Furthermore, one can verify the condition required by the martingale in Theorem 3.2.

$$\begin{aligned}|M_j - M_{j-1}| &= \left|E\left[f(X_1, \ldots, X_n) \mid\ \mathcal{F}_j\right] - E\left[f(X_1, \ldots, X_n) \mid\ \mathcal{F}_{j-1}\right]\right| \\ &\leq \sup_{\{y_1, \ldots y_j\}} \Big|E\left[f(y_1, \ldots, y_j, X_{j+1} \ldots, X_n) \mid\ \mathcal{F}_j\right] \\ &\quad - E\left[f(y_1, \ldots, y_{j-1}, X_j, \ldots X_n) \mid\ \mathcal{F}_{j-1}\right]\Big| \\ &\leq d_j\end{aligned}$$

Using Theorem 3.2, we get the claimed tail bound. □

3.2 Applications of Azuma's martingale inequality

In this section, we consider several applications of Azuma's martingale inequality and the average bounded differences inequality.

Random vectors in Banach spaces Let $\{X_1, \ldots, X_n\}$ be independent random vectors in a normed space $(Y, \|\cdot\|)$. If for all $j \leq n$, $\|A_j\| \leq a_j$ almost surely, then the following holds.

$$\mathbb{P}\left(\left|\left\|\sum_{j=1}^{n} X_j\right\| - \mathbb{E}\left\|\sum_{j=1}^{n} X_j\right\|\right| > t\right) \leq 2\exp\left(-\frac{t^2}{16\sum_{j=1}^{n} a_j^2}\right)$$

This follows from Azuma's martingale inequality by letting M_j be the random variable obtained by taking the conditional expectation of $\left\|\sum_{j=1}^{n} X_j\right\|$ conditioned on the σ-algebra generated by $\{X_1, \ldots, X_j\}$. This sequence of real valued random variables forms a martingale, and from the fact that $\|X_j\|$ is bounded above by a_j almost surely, we get that that the martingale difference is bounded above by $2a_j$ almost surely, and the result follows.

Observe that this proof tells us absolutely nothing about $\mathbb{E}\left\|\sum_{j=1}^{n} X_j\right\|$, but only gives a concentration inequality about the mean. This will be the norm when proving concentration inequalities using Azuma's martingale inequality.

Randomized bin packing Another practical application is the *randomized bin packing problem*. Suppose $\{X_1, \ldots, X_n\}$ are numbers from the interval $[0, 1]$, which we will call weights. What is the minimum number $N(X_1, \ldots, X_n)$ of partitions (or bins) we need for these weights such that the sum of weights in each partition (or bin) does not exceed 1? This is an optimization problem of great importance to computer scientists, and a naïve algorithm to compute the minimum number of bins is takes super-exponential time as a function of n.

Consider now a variant of this problem where the weights $\{X_1, \ldots, X_n\}$ are i.i.d random variables. We are interested in the minimal number of bins $N(X_1, \ldots, X_n)$ as n tends to ∞. One has a result in the spirit of the law of large numbers.

Theorem 3.5. *There exists a real number $\gamma \in [0, 1]$, depending on the distribution of X_i such that the following holds almost surely.*

$$\lim_{n\to\infty} \frac{N(X_1, \ldots, X_n)}{n} = \gamma$$

Let M_j be the random obtained by taking the conditional expectation of $F(X_1, \ldots, X_n)$ conditioned on the σ-algebra $\mathcal{F}_j$ generated by $\{X_1, \ldots, X_j\}$. This is a martingale, and we now verify that $|M_{j+1} - M_j|$ is at most 1.

$$|M_{j+1} - M_j| = |\mathbb{E}\left(N(X_1, \ldots, X_n) \mid \mathcal{F}_{j+1}\right) - \mathbb{E}\left(N(X_1, \ldots, X_n) \mid \mathcal{F}_j\right)|$$

To see that the right hand side is at most 1, it will suffice to condition further on the σ-algebra generated by $\{X_{j+2}, \ldots, X_n\}$. After doing so, we see that the difference is bounded pointwise by 1.

$$|N(x_1, \ldots, x_n) - N(x_1, \ldots, x_j, X_{j+1}, x_{j+2}, \ldots, x_n)| \leq 1$$

The above inequality follows from the fact that changing one weight can at most decrease or increase the number of bins by 1. If the weight increases, we need at most one additional bin. If the weight decreases, and the number of bins goes down by more than 1, then increasing the weight again would increase the number of bins by more than 1, which cannot happen. This proves the claim that the martingale difference is bounded above by 1. Using Azuma's martingale inequality, we get the following concentration inequality for N.

$$\mathbb{P}\left(|N(X_1,\ldots,X_n) - \mathbb{E}N(X_1,\ldots,X_n)| > t\right) \leq 2\exp\left(-\frac{t^2}{n}\right)$$

Letting $t = \frac{\gamma n}{2}$, we get that the probability that $N(X_1,\ldots,X_n)$ is less than $\frac{\gamma n}{2}$ is very small.

$$\begin{aligned}\mathbb{P}\left(N(X_1,\ldots,X_n) \leq \frac{\gamma n}{2}\right) &\leq 2\exp\left(-frac\gamma^2 n^2 n\right)\\ &= 2\exp\left(-\gamma^2 n\right)\end{aligned}$$

Functional and geometric concentration for the discrete cube Another application of Azuma's martingale inequality is functional and geometric concentration for the discrete cube. The discrete cube D_n is $\{0,1\}^n$ with the distance function d_H, where $d_H(x,y) \coloneqq \frac{1}{n}\sum_{i=1}^n |x_i - y_i|$. Let μ be the uniform probability measure on D_n. Let $f : D_n \to \mathbb{R}$ be 1-Lipschitz. If we consider the coordinates to be independent random variables, we see that the 1-Lipschitz condition translates to the average difference of f being less than or equal to 1. We then get the following concentration result for f.

$$\mu\left(|f(x) - \mathbb{E}f(x)| > t\right) \leq 2\exp\left(-nt^2\right) \tag{3.6}$$

This inequality gives us *geometric measure concentration.*

Theorem 3.7 (Geometric measure concentration)**.** *Let $A \subset D_n$ be a set such that $\mu(A) \geq \frac{1}{2}$. Let $\varepsilon > 0$, and let A_ε be the ε neighbourhood of A. Then the complement A_ε^C of A_ε has very little measure.*

$$\mu(A_\varepsilon^C) \leq 2\exp\left(-\varepsilon^2 n\right)$$

This is quite unlike what we are used to in the case of a solid cube in $\mathbb{R}^n$, with Lebesgue measure. In this case, as n approaches ∞, the measure of $\mu(A_\varepsilon^C)$ does not approach 0.

Proof of geometric measure concentration. Let $f(x) = d_H(x,A)$. Clearly, f is 1-Lipschitz and we can use (3.6). We have that $\mu(A_\varepsilon^C) = \mu(f(x) > \varepsilon)$. To use (3.6), we need to estimate $\mathbb{E}f(x)$. If $x \in A$, $f(x) = 0$.

$$\begin{aligned}\frac{1}{2} &\leq \mu(A)\\ &\leq \mu\left(|f(x) - \mathbb{E}f(x)| \geq \mathbb{E}f(x)\right)\\ &\leq 2\exp\left(-n\left(\mathbb{E}f(x)\right)^2\right)\end{aligned}$$

Thus, $\mathbb{E}f(x) \leq \frac{C}{\sqrt{n}}$ for some constant C.

We thus get the following.

$$\mu\left(f(x) > t + \frac{C}{\sqrt{n}}\right) \leq 2\exp(-nt^2)$$

Setting t to the appropriate value gives us the result. □

Random allocations Suppose we have n balls, which we randomly and independently put into m bins. One can ask various statistical questions about the setup, e.g. number of empty bins, longest stretch of empty bins and so on. The problem we consider is the following: let Z be the number of empty bins. We can write Z as $\sum_{j=1}^{m} Z_j$, where Z_j is the indicator for the event that the j^{th} bin is empty. We can easily evaluate $\mathbb{E}Z$: this will be $\sum_{j=1}^{m} \mathbb{E}Z_j = m\left(1-\frac{1}{m}\right)^n \approx m\exp\left(-\frac{n}{m}\right)$. Suppose now that $\frac{n}{m}$ is approximately a constant r. We want concentration inequalities for Z, and Z is a sum of identically distributed random variables. However, the Z_j are not independent, which means most of our results don't apply here. To deal with this, we consider the random variables $\{X_j\}$, which is the number of the bin the j^{th} ball was put in. Clearly, Z is a function of $\{X_1, \ldots, X_n\}$: like the previous examples, we can create a martingale by taking the conditional expectations of Z conditioned on the σ-algebra generated by $\{X_1, \ldots, X_j\}$. Furthermore, the martingale differences are bounded above by 1 as well, which means by Azuma's martingale inequality, we get the following concentration inequality.

$$\mathbb{P}\left(|Z - \mathbb{E}Z| > t\right) \leq 2\exp\left(-\frac{t^2}{4n}\right) \tag{3.8}$$

We can do better though: note that the martingale difference bound we obtained was a pointwise bound, and the difference in expectation might be smaller. Let Δ_j denote the martingale difference.

$$\Delta_j = \mathbb{E}\left[Z \mid X_1, \ldots, X_j\right] - \mathbb{E}\left[Z \mid X_1, \ldots, X_{j-1}\right]$$

Recall now that $Z = \sum_{j=1}^{m} Z_j$. Let $1 \leq k \leq m$. Then given $X_1, \ldots, X_j$, if at least one of them equals k, $Z_k = 0$. Otherwise Z_k is a Bernoulli random variable with $\mathbb{P}(Z_k = 1 \mid X_1, \ldots, X_j) = \left(1-\frac{1}{m}\right)^{n-j}$. This lets evaluate the conditional expectations of Z_k.

$$\mathbb{E}\left[Z_k \mid X_1, \ldots, X_j\right] = \left(1-\frac{1}{m}\right)^{n-j} \prod_{l=1}^{j} \left(1 - \mathbb{1}_k(X_l)\right)$$

We similarly have the $j-1^{\text{th}}$ term.

$$\mathbb{E}\left[Z_k \mid X_1, \ldots, X_{j-1}\right] = \left(1-\frac{1}{m}\right)^{n-j+1} \prod_{l=1}^{j-1} \left(1 - \mathbb{1}_k(X_l)\right)$$

Let X_j' be an independent copy of X_j. Then we can rewrite the above expression in the following manner.

$$\begin{aligned}
\mathbb{E}\left[Z_k \mid X_1, \ldots, X_{j-1}\right] &= \left(1-\frac{1}{m}\right)^{n-j} \mathbb{E}_{X_j'}\left(1 - \mathbb{1}_k(X_j')\right) \prod_{l=1}^{j-1} \left(1 - \mathbb{1}_k(X_l)\right) \\
&= \left(1-\frac{1}{m}\right)^{n-j} \mathbb{E}_{X_j'}\left[\left(1 - \mathbb{1}_k(X_j')\right) \prod_{l=1}^{j-1} \left(1 - \mathbb{1}_k(X_l)\right)\right]
\end{aligned}$$

We can write the martingale difference Δ_j by taking the difference of the two terms we derived.

$$\begin{aligned}\Delta_j &= \left(1-\frac{1}{m}\right)^{n-j} \cdot \left[\prod_{l=1}^{j}(1-\mathbb{1}_k(X_l)) - \mathbb{E}_{X'_j}\left[\left(1-\mathbb{1}_k(X'_j)\right)\prod_{l=1}^{j-1}(1-\mathbb{1}_k(X_l))\right]\right] \\ &= \left(1-\frac{1}{m}\right)^{n-j} \mathbb{E}_{X'_j}\left[\left(\mathbb{1}_k(X'_j)-\mathbb{1}_k(X_j)\right)\prod_{l=1}^{j-1}(1-\mathbb{1}_k(X_l))\right]\end{aligned}$$

Observe now that the above expression is bounded above pointwise by $a_j = \left(1-\frac{1}{m}\right)^{n-j}$, which is better than the previous estimate we had. We can work out what the terms of the improved concentration inequality will look like.

$$\begin{aligned}\sum_{j=1}^{n} a_j^2 &= \sum_{j=1}^{n}\left(1-\frac{1}{m}\right)^{2(n-j)} \\ &= \frac{1-\left(1-\frac{1}{m}\right)^{2(n+2)}}{1-\left(1-\frac{1}{m}\right)^{2}} \\ &\approx \frac{1-\exp\left(-\frac{2n}{m}\right)}{\frac{2}{m}-\frac{1}{m^2}} \\ &\approx \frac{1-\exp(-2r)}{\frac{2}{m}}\end{aligned}$$

We use this term in Azuma's martingale inequality.

$$\mathbb{P}\left(|Z-\mathbb{E}Z|>t\right) \leq 2\exp\left(-\frac{t^2}{2m(1-e^{-2r})}\right) \tag{3.9}$$

Compare this to (3.8). If r is much bigger than 1, then (3.9) is much better, because the denominator is on the order of $2n$, which is better than $4n$. If r is much smaller than 1, the denominator in (3.9) is roughly $4n$, which is the same as (3.8), giving us no improvement.

The takeaway from this calculation is that the better we can estimate the martingale differences, the better concentration inequality we get. Another takeaway is that for when r is much smaller than 1, the number of bins vastly outnumbers the number of balls, in which case, it's not hard to imagine that Z_j are almost independent random variables. Treating them as actually independent random variables, and using Hoeffding's inequality to get concentration for their sum gives us approximately the same result we got.

References

[BC05] Bo Brinkman and Moses Charikar. "On the Impossibility of Dimension Reduction in l_1". In: *J. ACM* 52.5 (Sept. 2005), pp. 766–788. DOI: 10.1145/1089023.1089026 (cit. on p. 14).

[Dvo61] AP Dvoredsky. "Some results on convex bodies and Banach spaces". In: (1961) (cit. on p. 1).

[HW71] David Lee Hanson and Farroll Tim Wright. "A bound on tail probabilities for quadratic forms in independent random variables". In: *The Annals of Mathematical Statistics* 42.3 (1971), pp. 1079–1083 (cit. on p. 16).

[JL84] William B Johnson and Joram Lindenstrauss. "Extensions of Lipschitz mappings into a Hilbert space". In: *Contemporary mathematics* 26.189-206 (1984), p. 1 (cit. on p. 14).

[JN08] William B. Johnson and Assaf Naor. "The Johnson-Lindenstrauss lemma almost characterizes Hilbert space, but not quite". In: *arXiv e-prints*, arXiv:0807.1919 (July 2008), arXiv:0807.1919. arXiv: 0807.1919 [math.FA] (cit. on p. 14).

[Kah64] Jean-Pierre Kahane. "Sur les sommes vectorielles sigma plus minus un". In: *COMPTES RENDUS HEBDOMADAIRES DES SEANCES DE L ACADEMIE DES SCIENCES* 259.16 (1964), p. 2577 (cit. on p. 10).

[Khi23] Aleksandr Khintchine. "Über dyadische brüche". In: *Mathematische Zeitschrift* 18.1 (1923), pp. 109–116 (cit. on p. 9).

[Lie73] Elliott H Lieb. "Convex trace functions and the Wigner-Yanase-Dyson conjecture". In: *Advances in Mathematics* 11.3 (1973), pp. 267–288. DOI: https://doi.org/10.1016/0001-8708(73)90011-X (cit. on p. 22).

[LN04] James R Lee and Assaf Naor. "Embedding the diamond graph in L p and dimension reduction in L 1". In: *Geometric & Functional Analysis GAFA* 14.4 (2004), pp. 745–747 (cit. on p. 14).

[RV13] Mark Rudelson and Roman Vershynin. "Hanson-Wright inequality and sub-gaussian concentration". In: *Electron. Commun. Probab.* 18 (2013), 9 pp. DOI: 10.1214/ECP.v18-2865 (cit. on p. 17).

[Wri73] Farrol Tim Wright. "A bound on tail probabilities for quadratic forms in independent random variables whose distributions are not necessarily symmetric". In: *The Annals of Probability* (1973), pp. 1068–1070 (cit. on p. 16).