Math 626: High Dimensional Probability

Taught by Mark Rudelson
Lecture notes by Sayantan Khan

January 29, 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)**.** *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.

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 $\left(\ell^2\right)^N$, given by inequalities involving the norms of elements in $\left(\ell^2\right) N$. 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 $\left(\ell^2\right)^N$ 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 A. Let $\{X_1, X_2, \ldots, X_m\}$ be i.i.d. samples of X. We define the sample covariance A_m in the following manner.

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

As m tends to infinity, the sample covariance A_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 A_m and A be matrices mapping some normed space X to some other normed space Y. We define the distance between A_m and A using the operator norm.

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

We can consider $w((A_m - A)v)$ 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 70s. 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.

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 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.

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

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.

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

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 $2^M \leq n < 2^{M+1}$.

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

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

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 (Khinchin's inequality)**.** *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 Khinchin'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 Khinchin's inequality, due to Kahane.

Theorem 2.11 (Kahane inequality)**.** *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 $\frac{t}{e\delta \left\|\mathbf{a}_2^2\right\|}$. 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.

2.2 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. □

2.3 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 to 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 $v \in V$. We will prove that the following inequality holds with high probability.

$$|\|Gv\|_2 - n| \leq \varepsilon$$

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

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).

[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).

[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).