High-Dimensional Probability

March 19, 2019

Contents

Week 1: Hoeffding and Chernoff Bounds and Key Distributions

Speaker Courtney Paquette, March 7

Intuition: Why should we study in optimization concentration inequalities? Consider the problem of *matrix completion*. Consider a fixed $n \times n$ matrix X with $\text{rank}(X) = r$ where $r << n$.

- Each entry X_{ij} is revealed to us independently with some probability $p \in (0, 1)$ and hidden with probability $1 - p$. Let $Y \in \mathbb{R}^{n \times n}$

$$Y_{ij} = \delta_{ij} X_{ij} \quad \text{where} \quad \delta_{ij} \sim \text{Ber}(p) \text{ are independent.}$$

- If $p = \frac{m}{n^2}$, then we are shown m entries of X on average.

How can we infer X from Y?
How many entries m do we need to see in order for matrix completion to be possible?

1 Review of Basic Probability

The material, in this section, is covered in most introductory graduate probability courses (see *e.g.* Vershynin's High-Dimensional Probability, Durrett's Probability: Theory and Examples). The purpose of this section is to introduce the basic notation as well as some classic probability theorems.

Throughout, we denote $(\Omega, \mathcal{F}, \mathbf{Pr})$, a probability measure space such that $\mathcal{F}$ is the σ-algebra containing sets from Ω and $\mathbf{Pr}$ is the probability measure on $\mathcal{F}$. A random variable $X : (\Omega, \mathcal{F}, \mathbf{Pr}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}), dx)$ is a mapping from the probability space to $\mathbb{R}$. Here $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra generated by open sets on $\mathbb{R}$ and m is the Lebesgue measure. Recall, the following operations

$$\begin{aligned} \text{Expectation of } X \qquad & \mathbf{E}[X] := \int_\Omega X \, dPr \\ \text{Variance of } X \qquad & \text{Var}[X] := \mathbf{E}[(X - \mathbf{E}[X])^2] \\ \text{Covariance of } X \text{ and } Y \qquad & \text{cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]. \end{aligned}$$

1.1 Important distributions and moment generating functions

As we recall from a basic probability course, the *distribution* of a random variable X, $F_X(t)$, is, intuitively, information about what values X takes and with what probabilities:

$$F_X(t) = \Pr(X \leq t) = \int_{-\infty}^{t} f(x) \, dx$$

where f is the *density* function of X. Below I list some examples of common density and distribution functions

- **Normal or Gaussian distribution:** $X \sim N(0, 1)$ (mean $\mu = 0$ and variance, $\sigma^2 = 1$) has density

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R}$$

- **Bernoulli distribution:** $X \sim \text{Ber}(p)$ with some fixed parameter $p \in (0, 1)$ means

$$\Pr(X = 1) = p \quad \text{and} \quad \Pr(X = 0) = 1 - p.$$

Recall, the $\mathbf{E}[X] = p$ and $\text{Var}(X) = p(1 - p)$.

- **Binomial distribution:** $X \sim$ Binom(N,p) is $X = \sum_{i=1}^{N} X_i$ where $X_i \sim$ Ber(p). Note $\mathbf{E}[X] = Np$ and $\mathrm{Var}(X) = Np(1-p)$
- **Poisson distribution:** $Z \sim \mathrm{Pois}(\lambda)$ if it takes values $0, 1, 2, \ldots$ with probabilities

$$\Pr(Z = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

Definition 1.1 (Moment Generating Function (MGF))**.** *The MGF of a random variable X is*

$$\begin{aligned} M_X(t) &= \mathbf{E}[e^{tX}], \quad t \in \mathbb{R} \\ &\approx 1 + t\mathbf{E}[X] + \frac{t^2\mathbf{E}[X^2]}{2!} + \frac{t^3\mathbf{E}[X^3]}{3!} + \ldots + \frac{t^n\mathbf{E}[X^n]}{n!} + \ldots \\ &= \int_{-\infty}^{\infty} e^{tx} f(x)\,dx \quad f(x) \text{ is the density function of } X \end{aligned}$$

Remarks:

- $\frac{d^p}{dt^p} M_X(t)|_{t=0} = \mathbf{E}[X^p]$– pth moment of X
- (Uniqueness) $M_X(t)$ is "two-sided" Laplace transform: If two random variables X and Y have MGFs that are finite and equal in some neighborhood of 0 then they have the same distributions, $P(X \le x) = P(Y \le y)$.
- Suppose X_i are independent. Then $M_{\sum_i X_i}(t) = \mathbf{E}[e^{t\sum_i X_i}] = \prod_i \mathbf{E}[e^{X_i t}]$

Examples of MGF. Now we give several examples of moment generating functions. For all of our examples, we will have very convenient bounds of the form

$$M_X(t) = \mathbf{E}[e^{tX}] \le \exp\left(\frac{C^2t^2}{2}\right) \quad \text{for all } t \in \mathbb{R} \tag{1.1}$$

for some $C \in \mathbb{R}$, depending on the distribution X.

- **Normal distribution:** $Z \sim N(0, \sigma^2)$ then

$$\mathbf{E}[\exp(tZ)] = \exp\left(\frac{t^2\sigma^2}{2}\right)$$

  (Check this! It's a direct calculation.)

- **Rademacher r.v. (symmetric Bernoulli):** $X = 1$ with prob. $1/2$ and $X = -1$ with prob. $1/2$. Note that Rademacher r.v. and Bernoulli r.v. are related by the identity $2X - 1$. Then

$$\mathbf{E}[e^{tX}] \le \exp\left(\frac{t^2}{2}\right).$$

  Proof. First, note that $\mathbf{E}[X^k] = 0$ whenever k is odd and $\mathbf{E}[X^k] = 1$ whenever k is even. Using the Taylor expansion of e^x, we have

$$\mathbf{E}[e^{tX}] = \sum_{k=0}^{\infty} \frac{t^k\mathbf{E}[X^k]}{k!} = \sum_{k=0,2,4,} \frac{t^k}{k!} = \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!}.$$

Finally, we use that $(2k)! \geq 2^k \cdot k!$ for all $k = 0, 1, 2 \ldots$, so that

$$\mathbf{E}[e^{tX}] \leq \sum_{k=0}^{\infty} \frac{(t^2)^k}{2^k \cdot k!} = \sum_{k=0}^{\infty} \left(\frac{t^2}{2}\right)^k \cdot \frac{1}{k!} = \exp\left(\frac{t^2}{2}\right).$$

□

I have also included the moment generating functions which do not satisfy the growth condition in (1.1).

- **Poisson distribution:** Suppose $Z \sim \text{Pois}(\lambda)$. It is a direct calculation to compute the moment generating function:

$$\mathbf{E}[e^{tZ}] = e^{-\lambda} \sum_{k=0}^{\infty} e^{tk} \cdot \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^t\lambda)^k}{k!} = e^{\lambda(e^t-1)}.$$

- **Exponential distribution:** Suppose $X \sim \exp(\lambda)$. Again, by a direct calculation, we compute its MGF, for all $t < \lambda$,

$$\mathbf{E}[e^{tX}] = \int_0^{\infty} \lambda e^{-\lambda x} e^{tx}\, dx = \frac{\lambda}{t-\lambda} e^{(t-\lambda)x}\Big|_0^{\infty} = \frac{\lambda}{\lambda - t}.$$

1.2 Limit theorems

In data science, we often encounter large sums of (identically distributed, independent (iid)) random variables. Recall, the variable of a sum of independent random variables is

$$\text{Var}\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} \text{Var}(X_i).$$

In particular, if X_i have the same distribution with mean μ and variance σ^2, then dividing both sides by N, we have

$$\text{Var}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{\sigma^2}{N}$$

Thus the variance of the *sample mean*, $\frac{1}{N}\sum_{i=1}^{N} X_i$ of the sample $\{X_1, X_2, \ldots, X_N\}$ shrinks to 0 as $N \to \infty$. This indicates for large N, we expect that the sample mean concentrates tightly about its expectation μ.

Theorem 1.2 (Strong law of large numbers)**.** *Let $X_1, X_2, \ldots,$ be a sequence of iid r.v. with mean μ. Consider the sum*

$$S_N = X_1 + X_2 + \ldots + X_N.$$

Then as $N \to \infty$

$$\frac{S_N}{N} \to \mu \quad \textit{a.s.}$$

The next theorem, the central limit theorem, makes one step further. It identifies the limiting distribution of the (properly scaled) sum of X_i's as the normal distribution.

Theorem 1.3 (Central limit theorem). *Let $X_1, X_2, \ldots$ be a sequence of iid random variables with mean μ and variance σ^2. Suppose S_N is defined as in Theorem 1.2 and we normalize the random variable S_N (i.e. mean is 0 and unit variance) as follows*

$$Z_N = \frac{S_N - \mathbf{E}[S_N]}{\sqrt{\mathrm{Var}(S_N)}} = \frac{1}{\sigma\sqrt{N}} \sum_{i=1}^{N} (X_i - \mu).$$

Then as $N \to \infty$,

$$Z_N \to N(0,1) \qquad \textit{in distribution.}$$

(Remark: this means $\Pr(Z_N \geq t) \to \Pr(Z \geq t) = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2}\, dx$.)

2 Concentration Inequalities

A basic question in probability, statistics, and machine learning is the following: given a r.v. X with expectation $\mathbf{E}[X]$, how likely is X to be close to its expectation? Concentration inequalities quantify how a r.v. X deviates from its mean. They usually take the form

$$\Pr(|X - \mathbf{E}[X]| > t) \leq \text{ something small .}$$

Often we see two types of behaviors depending on whether we are close to the mean or far away. Our first bound is perhaps the most basic of all probability inequalities.

Lemma 2.1 (Markov's inequality). *Let $X \geq 0$ be a non-negative r.v. Then for all $t \geq 0$*

$$\Pr(X \geq t) \leq \frac{\mathbf{E}[X]}{t}.$$

Proof. We have the following

$$t \Pr(X \geq t) = \int_{\{X \geq t\}} t \, d\Pr \leq \int_{\{X \geq t\}} X \, d\Pr \leq \int_\Omega X \, d\Pr = \mathbf{E}[X].$$

□

Essentially all other bounds on the probabilities are variations on Markov's inequality. The first variation, perhaps the most important one, uses second moments – variance – of a r.v. rather than its mean. This is known as Chebyshev's inequality.

Corollary 2.2 (Chebyshev's Inequality). *Let X be any random variable with Var$(X) < \infty$. Then for all $t \geq 0$*

$$\Pr(|X - \mathbf{E}[X]| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}.$$

Proof. The result is an immediate consequence of Markov's inequality. We note that $|X - \mathbf{E}[X]| \geq t$ implies $(X - \mathbf{E}[X])^2 \geq t^2$ (in fact the sets are equal). Thus, we have

$$\Pr(|X - \mathbf{E}[X]| \geq t) = \Pr((X - \mathbf{E}[X])^2 \geq t^2) \leq \frac{\mathbf{E}[(X - \mathbf{E}[X])^2]}{t^2} = \frac{\mathrm{Var}(X)}{t^2}$$

where the first inequality is Markov's inequality. □

A nice consequence of Chebyshev's inequality is that averages of r.v. with finite variance converge to their mean. Let us give an example of this fact. Suppose X_i are iid and satisfy $\mathbf{E}[X_i] = 0$. Then $\mathbf{E}[X_i] = 0$, while if we define $\bar{X} = \frac{1}{N}\sum_{i=1}^{n} X_i$ then

$$\mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X_1)}{N}.$$

In particular, for any $t \geq 0$, we have

$$\Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N} X_i\right| \geq t\right) \leq \frac{\mathrm{Var}(Z_1)}{Nt^2}$$

so that $\Pr(|\bar{X}| \geq t) \to 0$ for any $t > 0$. Remember N represents number of samples.

But often we expect to have even sharper – exponential – bounds on the probability that a r.v. X exceeds its expectation. Why do we expect this? Well $S_N = \sum_{i=1}^{N} X_i$ where X_i are iid random variables with mean 0 and variance σ^2. Then by normalizing S_N we know that it converges to a normal in distribution (central limit theorem):

$$\Pr\left(\frac{1}{N}S_N \geq t\right) = \Pr\left(\frac{S_N}{\sqrt{N}\sigma} \geq \frac{t\sqrt{N}}{\sigma}\right) \approx \Pr\left(Z \geq \frac{t\sqrt{N}}{\sigma}\right)$$

where $Z \sim N(0,1)$. Moreover, we have the following results for the tails of a normal distributed r.v.

Theorem 2.3 (Tails of Normal Distribution)**.** *Let $Z \sim N(0,1)$. Then for all $t \geq 0$, we have*

$$\left(\frac{1}{t} - \frac{1}{t^3}\right) \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \leq \Pr(Z \geq t) \leq \frac{1}{t} \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

In particular, for $t \geq 1$, the tail is bounded by the density

$$\Pr(Z \geq t) \leq \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

Therefore, we have the following results for any $t \geq \sigma/\sqrt{N}$

$$\text{Chebyshev} \qquad \Pr\left(\frac{1}{N} \cdot \sum_{i=1}^{N} X_i \geq t\right) \leq \frac{\sigma}{Nt^2}$$

$$\text{"CLT"} \qquad \Pr\left(\frac{1}{N} \cdot \sum_{i=1}^{N} X_i \geq t\right) \leq \frac{1}{\sqrt{2\pi}} \cdot e^{-t^2 N/2\sigma^2}$$

The CLT says it decays *exponentially* fast in N, which is much better than the linear decay that followed from Chebyshev's inequality. Unfortunately, this does not follow rigorously from the CLT. Although the approximation by the normal density is valid, the error in approximation can not be ignored and this error decays even slower than linearly in N.

Theorem 2.4 (Berry-Esseen CLT)**.** *In the setting of Theorem 1.3, for every N and every $t \in \mathbb{R}$ we have*

$$|\Pr(Z_N \geq t) - \Pr(Z \geq t)| \leq \frac{\rho}{\sqrt{N}}$$

where $\rho = \mathbf{E}|X_1 - \mu|^3/\sigma^3$ and $Z \sim N(0,1)$.

2.1 Chernoff bounds

Chernoff bounds use the MGF to obtain exponential deviation bounds.

Theorem 2.5 (Chernoff bounds)**.** *Let X be a r.v. Then for any $t \geq 0$, we have*

$$\Pr(X - \mathbf{E}[X] \geq t) \leq \min_{\lambda \geq 0} \mathbf{E}[e^{\lambda(X - \mathbf{E}[X])}]e^{-\lambda t} = \min_{\lambda \geq 0} M_{X-\mathbf{E}[X]}(\lambda)e^{-\lambda t}$$

and

$$\Pr(X - \mathbf{E}[X] \leq -t) = \min_{\lambda \geq 0} \mathbf{E}[e^{\lambda(\mathbf{E}[X] - X)}]e^{-\lambda t} \leq \min_{\lambda \geq 0} M_{\mathbf{E}[X]-X}(\lambda)e^{-\lambda t}.$$

Proof. I will only prove the first inequality since the second inequality is completely identical. The idea is to use Markov's inequality on the exponentiated version. For any $\lambda \geq 0$, we have $X - \mathbf{E}[X] \geq t$ iff $\exp(\lambda(X - \mathbf{E}[X])) \geq \exp(\lambda t)$. Hence, we have

$$\Pr(X - \mathbf{E}[X] \geq t) = \Pr(\exp(\lambda(X - \mathbf{E}[X])) \geq \exp(\lambda t)) \leq \mathbf{E}[e^{\lambda(X - \mathbf{E}[X])}]e^{-t\lambda}$$

where the last inequality follows from Markov's inequality (Lemma 2.1). As our choice of $\lambda > 0$ does not matter, we can take the best one by minimizing the right side of the bound (Note: the bound holds trivially for $\lambda = 0$). □

Example with a sum of Radamacher r.v. Recall in Section 1.1, the MGF for a Radamacher r.v. X is

$$\mathbf{E}[e^{\lambda X}] \leq \exp\left(\frac{\lambda^2}{2}\right). \tag{2.1}$$

Suppose $X = \sum_{i=1}^{N} X_i$ where $X_i \in \{-1, 1\}$ is an independent random sign with $\mathbf{E}[X] = 0$. By the Chernoff bound, it becomes immediately clear that

$$\Pr(X \geq t) \leq \mathbf{E}[e^{\lambda X}]e^{-\lambda t} = \mathbf{E}[e^{\lambda X_1}]^N e^{-\lambda t} \leq \exp\left(\frac{N\lambda^2}{2}\right)e^{-\lambda t}$$

where the equality comes from X being a sum of independent r.v. As this holds for all $\lambda \geq 0$, we can minimizing this in λ, namely,

$$\min_{\lambda \geq 0} \frac{N\lambda^2}{2} - \lambda t.$$

Then $\lambda = t/N$, which gives

$$\Pr(X \geq t) \leq \exp\left(-\frac{t^2}{2N}\right).$$

In particular, taking $t = \sqrt{2N \log(1/\delta)}$, we have

$$\Pr\left(\sum_{i=1}^{N} X_i \geq \sqrt{2N \log\left(\frac{1}{\delta}\right)}\right) \leq \delta$$

So $X = \sum_{i=1}^{N} X_i = O(\sqrt{N})$ with extremely high probability – the sum of N independent random signs is essentially never larger than $\sqrt{N}$.

2.2 Hoeffding's Inequality

Hoeffding's inequality is a powerful technique for bounding the probability that sums of r.v. are too large or too small.

Lemma 2.6 (Hoeffding's Lemma)**.** *Let X be a bounded r.v. with $X \in [a, b]$. Then*

$$\mathbf{E}[\exp(\lambda(X - \mathbf{E}[X]))] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right) \qquad \text{for all } \lambda \in \mathbb{R}.$$

Proof. We prove a slightly weaker version of this lemma with a factor of 2 instead of 8. The idea is to use *symmetrization* with Jensen's inequality and our random sign MGF. First, let X' be an independent copy of X with the same distribution, so $X' \in [a, b]$ and $\mathbf{E}[X'] = \mathbf{E}[X]$ but X and X' are independent. Then we have

$$\mathbf{E}_X[\exp(\lambda(X - \mathbf{E}[X]))] = \mathbf{E}_X[\exp(\lambda(X - \mathbf{E}_{X'}[X']))] \leq \mathbf{E}_X[\mathbf{E}_{X'}[\exp(\lambda(X - X'))]]$$

where $\mathbf{E}_X$ and $\mathbf{E}_{X'}$ indicate expectations taken with respect to X and X'. Here the inequality follows from Jensen's inequality applied to the function $f(x) = e^{-x}$. The difference $X - X'$ is symmetric about zero, namely, $-(X - X')$ has the same distribution as $X - X'$. So if $S \in \{-1, 1\}$ is a random sign variable, then $S(X - X')$ has exactly the same distribution as $X - X'$. So we have

$$\begin{aligned}\mathbf{E}_{X,X'}[\exp(\lambda(X - X'))] &= \mathbf{E}_{X,X',S}[\exp(\lambda S(X - X'))] \\ &= \mathbf{E}_{X,X'}\big[\mathbf{E}_S[\exp(\lambda S(X - X')) \,|\, \sigma(X, X')]\big].\end{aligned}$$

This last inequality used the conditioning property that we saw from martingales. Now we can use the MGF of the random sign (2.1), which yields

$$\mathbf{E}_S[\exp(\lambda S(X - X')) \,|\, \sigma(X, X')] \leq \exp\left(\frac{\lambda^2(X - X')^2}{2}\right).$$

By our assumption $|X - X'| \leq (b - a)$ so $(X - X')^2 \leq (b - a)^2$. Thus, this give

$$\mathbf{E}_{X,X'}[\exp(\lambda(X - X'))] \leq \exp\left(\frac{\lambda^2(b-a)^2}{2}\right)$$

and the result immediately follows with a 2 instead of an 8. □

We now prove Hoeffding's inequality using (1) Chernoff bounds and (2) Hoeffding's lemma.

Theorem 2.7 (Hoeffding's inequality)**.** *Fix constants $-\infty < a < b < \infty$. Let $X_1, X_2, \ldots, X_N$ be independent bounded random variables with $X_i \in [a, b]$ for all i. Then*

$$\Pr\left(\frac{1}{N}\sum_{i=1}^{N}(X_i - \mathbf{E}[X_i]) \geq t\right) \leq \exp\left(-\frac{2Nt^2}{(b-a)^2}\right)$$

and

$$\Pr\left(\frac{1}{N}\sum_{i=1}^{N}(X_i - \mathbf{E}[X_i]) \leq -t\right) \leq \exp\left(-\frac{2Nt^2}{(b-a)^2}\right)$$

for all $t \geq 0$.

Note, in fact, this gives a two-sided bound since

$$\Pr(|X| \geq t) = \Pr(X \geq t) + \Pr(X \leq -t).$$

Therefore, we have from Hoeffding's inequality that

$$\Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N}(X_i - \mathbf{E}[X_i])\right| \geq t\right) \leq 2 \cdot \exp\left(-\frac{2Nt^2}{(b-a)^2}\right).$$

Proof. We only prove the result for the upper tail as the lower tail proof is identical. By Chernoff's bounds (Theorem 2.5), we immediately have

$$\begin{aligned}
\Pr\left(\frac{1}{N}\sum_{i=1}^{N}(X_i - \mathbf{E}[X_i]) \geq t\right) &= \Pr\left(\sum_{i=1}^{N}(X_i - \mathbf{E}[X_i]) \geq tN\right) \\
&\leq \min_{\lambda \geq 0} \ \mathbf{E}\left[\exp\left(\lambda \sum_{i=1}^{N}(X_i - \mathbf{E}[X_i])\right)\right] e^{-\lambda N t} \\
\text{(independence of } X_i \text{ and MGF product)} \quad &= \min_{\lambda \geq 0} \ e^{-\lambda N t} \prod_{i=1}^{N} \mathbf{E}[e^{\lambda(X_i - \mathbf{E}[X_i])}] \\
\text{(Hoeffding's lemma)} \quad &\leq \min_{\lambda \geq 0} \ e^{-\lambda N t} \prod_{i=1}^{N} \exp\left(\frac{\lambda^2 (b-a)^2}{8}\right).
\end{aligned}$$

We now need to compute the minimum of λ. This is the same as finding the following

$$\min_{\lambda \geq 0} \ \frac{N\lambda^2(b-a)^2}{8} - \lambda N t.$$

Taking derivatives, we get that $\lambda = \frac{4t}{(b-a)^2}$. Plugging this in for λ and simplifying yields

$$\Pr\left(\frac{1}{N}\sum_{i=1}^{N}(X_i - \mathbf{E}[X_i]) \geq t\right) \leq \min_{\lambda \geq 0} \ e^{-\lambda N t} \prod_{i=1}^{N} \exp\left(\frac{\lambda^2 (b-a)^2}{8}\right) = \exp\left(-\frac{2Nt^2}{(b-a)^2}\right).$$

□

Week 2: Singular values of random matrices
Speaker Elliot Paquette, March 14

3 Singular values of random matrices

Theorem 3.1 (General Hoeffding Inequality)**.** *Let $\{X_i\}_{i=1}^N$ be independent random variables and $X_i \in [a_i, b_i]$ a.s. where $a_i, b_i \in \mathbb{R}$ and $a_i < b_i$. Then*

$$\Pr\left(\sum_{i=1}^N (X_i - \mathbf{E}[X_i]) > t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^N (b_i - a_i)^2}\right)$$

Proof. The same as Hoeffding's inequality for bounded random variables. □

3.1 Largest singular value of a random matrix

Assumption 3.2. *Let A be a random $m \times n$ matrix such that $\{A_{ij}\}_{i\in[m],j\in[n]}$ are independent and bounded, $|A_{ij}| \leq 1$ and $\mathbf{E}[A_{ij}] = 0$.*

Theorem 3.3. *There is an absolute constant $C > 0$ so that*

$$\Pr(\|A\|_{op} \geq C\sqrt{m+n}) \to 0 \quad \text{as } \max\{m, n\} \to \infty.$$

Remark 1. We state some observations:

- The theorem says $\|A\|_{op} \approx \sqrt{m+n}$
- This is tight because $\|Ae_1\|^2 = \sum_{i=1}^m A_{i1}^2$ so $\mathbf{E}[\|Ae_1\|^2]$ is of order m if $\mathbf{E}[A_{i1}^2] \geq c > 0$

We recall some facts about the operator norm

$$\|A\|_{op} = \sup_{\|x\|_2=1} \|Ax\|_2 = \sup_{\|x\|_2=\|y\|_2=1} x^T A y.$$

Unfortunately, the spheres $\mathbb{S}^{n-1}$ and $\mathbb{S}^{m-1}$ have uncountable number of points and so we need to **reduce to a finte set**.

Definition 3.4 (Covering number and ε-net)**.** Let $K \subseteq \mathbb{R}^n$ be compact and $\varepsilon > 0$. Define the *covering number* $\mathcal{N}(K, \varepsilon)$ to be the cardinality of the smallest ε-net of K. A set $W \subset K$ is an *ε-net of K* if for any $y \in K$, there exists a $w \in W$ so that $\|w - y\| \leq \varepsilon$.

Definition 3.5 (Packing number and ε-separated)**.** Let $K \subseteq \mathbb{R}^n$ be compact and $\varepsilon > 0$. Define the *packing number* $\mathcal{P}(K, \varepsilon)$ to be the largest cardinality of an ε-separated set in K. A set $W \subset K$ is an *ε-separated* if for all $w_1, w_2 \in W$, $\|w_1 - w_2\| > \varepsilon$.

Lemma 3.6.

$$\mathcal{P}(K, 2\varepsilon) \leq \mathcal{N}(K, \varepsilon) \leq \mathcal{P}(K, \varepsilon).$$

Proof. We only prove the second inequality and leave the first one to the reader. Let $W \subseteq K$ be an inclusion maximal ε-separated set. Then for all $x \in K \setminus W$, $\min_{w\in W} \|x - w\|_2 \leq \varepsilon$. Therefore, the set W is an ε-net. □

In particular, for our problem, we want to estimate the covering number of the n-dimensional sphere, $\mathbb{S}^{n-1}$. The above lemma says, in particular, that we can bound the covering number by the packing number. The lemma below estimates the covering number of the sphere.

Lemma 3.7 (Packing number of the sphere)**.** *Fix $\varepsilon > 0$. If $K = \mathbb{S}^{n-1}$, the sphere in n-dimensional space, then*

$$\mathcal{N}(K, \varepsilon) \leq \left(1 + \frac{2}{\varepsilon}\right)^n.$$

Proof. The idea is to estimate the covering number of the sphere. Then by the previous lemma we have an estimate of the packing number of the sphere which gives an estimate of the covering number. For each point x in the ε-separated set of K, if we draw a ball of radi $\varepsilon/2$ centered at x then all these $\varepsilon/2$-balls are disjoint. We denote $B(x, r)$ to be a ball centered at x with radius r. Hence we can compute the volume of these balls:

$$\mathcal{P}(K, \varepsilon)\text{Vol}(B(0, \tfrac{\varepsilon}{2})) \leq \text{Vol}\Big(\bigcup_{x \in \mathbb{S}^{n-1}} B(x, \tfrac{\varepsilon}{2})\Big) \leq \text{Vol}(B(0, 1 + \tfrac{\varepsilon}{2})).$$

Here the last inequality follows because $\bigcup_{x \in \mathbb{S}^{n-1}} B(x, \varepsilon/2)$ forms a $\varepsilon/2$-neighborhood of $\mathbb{S}^{n-1}$. Rearranging the above inequalities gives

$$\mathcal{P}(K, \varepsilon) \leq \frac{\text{Vol}(B(0, 1 + \frac{\varepsilon}{2}))}{\text{Vol}(B(0, \frac{\varepsilon}{2}))} = \frac{(1 + \frac{\varepsilon}{2})^n}{(\frac{\varepsilon}{2})^n}.$$

□

Next, we return to the proof for singular values.

Proof of Theorem 3.3. The idea is to replace all x, y on the sphere by all x, y in an ε-net. Let $\mathcal{N}_m, \mathcal{N}_n$ be ε-nets of $\mathbb{S}^{m-1}$ and $\mathbb{S}^{n-1}$ respectively. We now will prove the result for ε-nets. By the union bound,

$$\Pr\left(\exists (x, y) \in \mathcal{N}_m \times \mathcal{N}_n \,:\, x^T A y > t\sqrt{m+n}\right) \leq |\mathcal{N}_m||\mathcal{N}_n| \max_{(x,y) \in \mathcal{N}_m \times \mathcal{N}_n} \Pr(x^T A y > t\sqrt{m+n}). \tag{3.1}$$

We observe that $x^T A y = \sum_{i,j} x_i A_{ij} y_j$ and $x_i A_{ij} y_j$ is bounded in $|\cdot|$ by $|x_i y_i|$. Here we used the assumption that $|A_{ij}| \leq 1$. Now we are in a position to use Hoeffding's inequality (Theorem 3.1)

$$\Pr\left(x^T A y > t + \mathbf{E}[x^T A y]\right) = \Pr\left(x^T A y > t\right) \leq \exp\left(\frac{-2t^2}{4\left\|x\right\|_2^2 \left\|y\right\|_2^2}\right).$$

Here we used the assumption that $\mathbf{E}[x^T A y] = 0$. From Lemma 3.7, we bound the cardinality of the packing numbers, $|\mathcal{N}_m||\mathcal{N}_n| \leq (1 + \frac{2}{\varepsilon})^{m+n}$. This, together with (3.1), gives the bound

$$\Pr\left(\exists (x, y) \in \mathcal{N}_m \times \mathcal{N}_n \,:\, x^T A y > t\sqrt{m+n}\right) \leq \left(1 + \frac{2}{\varepsilon}\right)^{m+n} \exp\left(\frac{-2t^2(m+n)}{4\left\|x\right\|_2^2 \left\|y\right\|_2^2}\right).$$

By choosing t appropriately, we can make the RHS small. Therefore, for any $\varepsilon > 0$, there exists a constant $C_\varepsilon > 0$ so that on $\mathcal{N}_m \times \mathcal{N}_n$, $x^T A y \leq C_\varepsilon \sqrt{m+n}$ with probability $\to 1$ as $\max\{m, n\} \to \infty$. We have shown the result when x, y live on the ε-nets so now we need to make it hold for any x, y living on the sphere.

Suppose $\varepsilon < 1/4$ Then for any $(x, y) \in \mathbb{S}^{m-1} \times \mathbb{S}^{n-1}$, there exists $(u, v) \in \mathcal{N}_m \times \mathcal{N}_n$ so that $\|u - x\| < 1/4$ and $\|v - y\| < 1/4$. Choose x, y to be the vectors on the sphere which attain sup $x^T Ay$. It follows

$$x^T Ay = u^T Av + (x - u)^T Av + x^T A(y - v) \leq u^T Av + 1/4x^T Ay + 1/4x^T Ay.$$

Here we scaled $x - u$ to be on the sphere then used that x, y attain the sup. Since $x^T Ay = \|A\|_{op}$, we get by rearranging

$$\tfrac{1}{2} \|A\|_{op} \leq u^T Av \leq C_{1/4}\sqrt{m + n},$$

since u, v are in the ε-net. The result follows.

□

Week 3: Low rank approximation and sub-gaussian distributions

Speaker Courtney Paquette, March 20

3.2 Low rank approximation

Question: Given a matrix $X \in \mathbb{R}^{m\times n}$, can one tell if X is completely random noise or does it contain a signal?

Consider the same assumption as last time (Assumption 3.2).

Assumption 3.8. *Let A be a random $m \times n$ matrix such that $\{A_{ij}\}_{i\in[m],j\in[n]}$ are independent and bounded, $|A_{ij}| \le 1$ and $\mathbf{E}[A_{ij}] = 0$.*

Last week, we showed that under Assumptions 3.2 and for any $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$ the following

$$\Pr(x^T Ay \le -t) \le \exp\left(\frac{-t^2}{2\|x\|_2^2 \|y\|_2^2}\right) \quad \text{for all } t > 0$$

$$\text{and} \qquad \Pr\left(\sup_{x\in\mathbb{S}^{m-1}, y\in\mathbb{S}^{n-1}} x^T Ay \ge C'\sqrt{m+n}\right) \le \exp\left(-C''(m+n)\right). \tag{3.2}$$

Consider adding a rank one perturbation to A, namely, adding λuv^T for some u, v unit vectors and $\lambda \in \mathbb{R}$ independent of A. By conditioning, we may assume that λ is deterministic. Here think

$$X = A + \lambda uv^T.$$

Question: How big should λ be so that you can see the signal in the singular values of X?

Proposition 3.9 (Largest singular value of perturbed A)**.** *Suppose $A \in \mathbb{R}^{m\times n}$ is an $(m\times n)$-matrix satisfies Assumption 3.2. Let $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$ be any unit vectors. Then there exists an absolute constant C so that if $\lambda \ge C\sqrt{m+n}$ the following occurs*

$$\Pr\left(\sigma_{\max}(A + \lambda uv^T) \le C'\sqrt{m+n}\right) \le \exp(-C''(m+n))$$

for some constant $C'' > 0$.

Proof. From the definition of $\sigma_{\max}$, we have the following

$$\begin{aligned}
&\sigma_{\max}(A + \lambda uv^T) \ge u^T(A + \lambda uv^T)v^T = u^T Av + \lambda \\
\Rightarrow \quad &\Pr\left(\sigma_{\max}(A + \lambda uv^T) \le C'\sqrt{m+n}\right) \le \Pr\left(u^T Av + \lambda \le C'\sqrt{m+n}\right) \\
&\qquad = \Pr\left(u^T Av \le (C'-C)\sqrt{m+n}\right) \\
&\text{(Hoeffding's Inequality)} \qquad \le \exp\left(\frac{-(C'-C)^2(m+n)}{2\|u\|^2\|v\|^2}\right).
\end{aligned}$$

Here we assume that $C' - C < 0$. □

Remark: This says if the signal is large, namely $\sqrt{m+n}$, then the largest singular value of the matrix X is bigger than $\sqrt{m+n}$. At the moment, we haven't finished answering our question because we do not know what happens to the other singular values of X. Recall, last time, we saw that all the singular values of A basically are less than $O(1)\sqrt{m+n}$. We will see that, in fact, for the other singular values of X they w.h.p. lie are less than $\sqrt{m+n}$.

We recall the *Courant-Fisher's min-max theorem*: Suppose $\sigma_{\max} = \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$ are the ordered singular values of $A \in \mathbb{R}^{m\times n}$ where $m \geq n$. Then

$$\sigma_i = \min_{\dim E = i} \left\{ \max_{x\in E, \|x\|_2 = 1} \|Ax\|_2 \right\}$$

In particular, we have that for the second largest eigenvalue

$$\sigma_2 = \inf_w \sup_{x\perp w, \|x\|_2=1} \|Ax\|_2 . \tag{3.3}$$

Proposition 3.10. *Suppose $A \in \mathbb{R}^{m\times n}$ is an $(m\times n)$-matrix satisfies Assumption 3.2. Let $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$ be any unit vectors and $\lambda \geq 0$. Then the following holds*

$$\Pr\left(\sigma_2(A+\lambda uv^T) \geq C'\sqrt{m+n}\right) \leq \exp\left(-C''(m+n)\right)$$

for some constant $C' > 0$.

Proof. We just need to use equation (3.2) which we proved last week. By using (3.3), we have the following chain of inequalities

$$\begin{aligned}
\sigma_2(A+\lambda uv^T) = \inf_w \sup_{x\perp w} \|Ax\|_2 &\leq \sup_{\substack{x\perp u \text{ or } y\perp v \\ x\in\mathbb{S}^{m-1}, y\in\mathbb{S}^{n-1}}} x^T(A+\lambda uv^T)y \\
&\leq \sup_{\substack{x\perp u \text{ or } y\perp v \\ x\in\mathbb{S}^{m-1}, y\in\mathbb{S}^{n-1}}} x^T Ay \\
&\leq \sup_{x\in\mathbb{S}^{m-1}, y\in\mathbb{S}^{n-1}} x^T Ay.
\end{aligned}$$

The result follows from equation (3.2). □

Remarks: The last two propositions, together, show that when the signal is large, $\lambda \geq C\sqrt{m+n}$, we can distinguish the signal in the singular values of the matrix $A+\lambda uv^T$ from pure noise. In fact, although we didn't prove this, when the signal is smaller than $C\sqrt{m+n}$ and provided all other things held equally (e.g. we don't know the signal is sparse), we *can not* distinguish the signal from pure noise. If the signal is sparse, then one can see the signal with a lower threshold than $C\sqrt{m+n}$, but this requires a lot more work.

Although I only showed the rank one case, a similar result holds for rank k. Namely, if one adds a rank k matrix and the signals are larger than $C\sqrt{m+n}$, one can distinguish the top k signals.

We can also say something about the smallest singular value; however proving this result is much more difficult. When $m >> n$, it is clear that the smallest singular value is bounded away from 0. Using the net argument for the largest singular value, we have $\sigma_{\min} \sim \sqrt{m}$. When $m = n$, there is no gap, but it is still invertible, namely $\sigma_{\min} \sim n^{-1/2}$ with high probability.

Proposition 3.11 (Smallest singular value, $m = n$)**.** *Suppose $A \in \mathbb{R}^{n\times n}$ satisfies Assumption 3.2. Then for every $\varepsilon \geq 0$ one has*

$$\Pr(\sigma_{\min}(A) \leq \varepsilon n^{-1/2}) \leq C\varepsilon + c^n,$$

where $C > 0$ and $c \in (0,1)$.

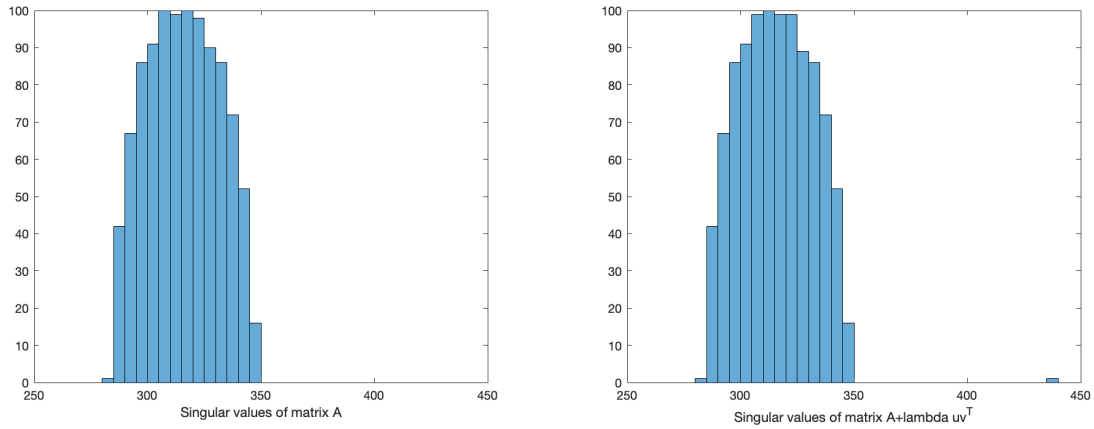

Figure 1: Histogram of the singular values of the matrices A and $A + \lambda uv^T$ where A is $(100,000 \times 1000)$ matrix generated so that each entry is $N(0,1)$ and independent. The vectors u and v are just the one vectors and $\lambda = 0.3$. You can see how the largest singular value pops out and all the other singular values lie within $330 \approx \sqrt{m+n}$.

Proof. See Rudelson and Vershynin, "The Littlewood-offord problem and invertibility of random matrices" *Advances in Mathematics* Vol 218, Issue 2, 600–633, 2008 □

3.3 Principal component analysis (PCA)

This idea of distinguishing a signal is of utmost importance in data science. To illustrate this, we introduce an important concept from probability, statistics, and data science called *principal component analysis*. Before continuing, let's introduce a few basic concepts about high-dimensional distributions.

Definition 3.12 (Mean, covariance, and second moment of random vector)**.** The concept of *mean* for a random variable generalizes in a straightforward way for a random vector X taking values in $\mathbb{R}^n$. The notion of variance is replaced in high dimensions by the *covariance matrix* of a random vector $X \in \mathbb{R}^n$, defined as follows

$$\text{cov}(X) = \mathbf{E}[(X-\mu)(X-\mu)^T] = \mathbf{E}[XX^T] - \mu\mu^T, \quad \text{where } \mu = \mathbf{E}[X].$$

Thus $\text{cov}(X)$ is an $n \times n$ symmetric, positive-semidefinite matrix. The diagonal entries of $\text{cov}(X)$ are precisely the variances of X_i. More generally, the entries of $\text{cov}(X)$ are the *covariances* of the pairs of coordinates of $X = (X_1, \ldots, X_n)$:

$$\text{cov}(X)_{ij} = \mathbf{E}[(X_i - \mathbf{E}[X_i])(X_j - \mathbf{E}[X_j])].$$

We also define the *second moment matrix* of a random vector X as

$$\Sigma = \Sigma(X) = \mathbf{E}[XX^T].$$

The second moment matrix is a higher dimensional generalization of the second moment $\mathbf{E}[Z^2]$. By translation, we can assume that X has mean zero and thus the covariance and second moment matrices are equal. Like the covariance matrix, the second moment matrix Σ is an $n \times n$ symmetric

positive semi-definite matrix. The spectral theorem for such matrices says that all eigenvalues λ_i of Σ are real and non-negative. Moreover Σ can be expressed via spectral decomposition as

$$\Sigma = \sum_{i=1}^{n} \lambda_i u_i u_i^T,$$

where $u_i \in \mathbb{R}^n$ are the eigenvectors of Σ. We will assume that λ_i are in decreasing order.

In data science, the spectral decomposition of Σ is of utmost importance when the distribution of a random vector X represents data. The eigenvector u_1 corresponding to the largest eigenvalue λ_1 defines the first *principal direction*. This is the direction which the distribution is most extended and explains most of the variability in the data. The next eigenvector u_2 (corresponding to the second largest singular value) defines the next principal direction; it best explains the remaining variability in the data... It should be noted from data we only see an estimate of the covariance matrix

$$\Sigma_m = \frac{1}{m} \sum_{i=1}^{m} X_i X_i^T$$

where $X_i \in \mathbb{R}^n$ represents data. We know by law of large numbers $\Sigma_m \to \Sigma$ a.e. as the sample size increases. This leads to the quantitative question: how large must the sample size m be to guarantee $\Sigma_m \approx \Sigma$ with high probability?

It often happens (as we just saw) with real data that only a few eigenvalues λ_i are large and thus can be considered informative; the remaining eigenvalues are small and considered noise. thus, a few principal directions explain most variability in the data. Even though the data is presented in high-dimensional space $\mathbb{R}^n$, such data is essentially low dimensional, namely it clusters near the low-dimensional subspace E spanned by the first few principal components.

The most basic data analysis algorithm, called *principal component analysis* (PCA), computes the first few principal components and then projects the data in $\mathbb{R}^n$ onto the subspace E spanned by them. This reduces the dimension of the data and simplifies the data analysis.

3.4 Isotropic random vectors

In a basic probability course it is useful to often assume that random variables have zero mean and unit variances. This is also true in higher dimensions, where the notion of isotropy generalizes the unit variance.

Definition 3.13 (Isotropic random variables)**.** A random vector $X \in \mathbb{R}^n$ is called *isotropic* if

$$\Sigma(X) = \mathbf{E}[XX^T] = I_n$$

where I_n denotes the identity matrix in $\mathbb{R}^n$.

Remark: We state some general properties of isotropic random variables.

- (Characterization of isotropy) A random vector $X \in \mathbb{R}^n$ is isotropic iff $\mathbf{E}[\langle X, x\rangle^2] = \|x\|_2^2$ for all $x \in \mathbb{R}^n$. This says X is isotropic iff all one-dimensional marginals of X have unit variance. Informally this means that an isotropic distribution is extended evenly in all directions.

- (Almost orthogonality of independent vectors) We have $\|X\|_2^2 = n$ and if X, Y are two isotropic, independent random vectors then $\mathbf{E}[\langle X, Y\rangle^2] = n$. To see the almost orthogonality, we can normalize X and Y, $\bar{X} = X/\|X\|$ and $Y/\|Y\|$. Since $\|X\|_2 \sim \sqrt{n}$ and $\|Y\| \sim \sqrt{n}$ and $\langle X, Y\rangle \sim \sqrt{n}$ w.h.p. it implies that $|\langle \bar{X}, \bar{Y}\rangle| \sim \frac{1}{\sqrt{n}}$.

4 Important distributions for concentration inequalities

4.1 Sub-gaussian distributions

It would be useful to extend these concentration inequalities to a wider class of distributions – at least we might expect Gaussian to belong to this class.

Question: which random variables X must obey a concentration inequality like Hoeffding's?

If we let the sum be a single term, then Hoeffding's inequality says

$$\Pr(|X| \geq t) \leq 2e^{-ct^2}, \quad \text{for some constant } c > 0.$$

This immediate tells us: if we want Hoeffding's inequality to hold, we must assume the r.v. X has sub-gaussian tails. This class of such distributions is called *sub-gaussian*. It contains Gaussian, Bernoulli, and any bounded distribution (i.e. $\|X\|_\infty < \infty$). And, as we will see shortly, concentration inequalities can indeed be proved for all sub-gaussian distributions. This makes the family of sub-gaussian distributions a natural and in many cases canonical class where one can develop various high dimensional probability theory.

The next proposition states a relationship between the sub-gaussian tail decay, the growth of moments, and the growth of the moment generating function. The proof is also useful since it shows how to transform one type of information about r.v. into another.

Theorem 4.1 (Sub-gaussian properties)**.** *Let X be a r.v. Then the following properties are equivalent; the parameters $K_i > 0$ appearing in these properties differ from each other by at most an absolute constant factor ($K_j \leq CK_i$ for any two properties).*

1. *The tails of X satisfy*

$$\Pr(|X| \geq t) \leq 2\exp(-t^2/K_1^2) \quad \textit{for all } t \geq 0.$$

2. *The moments of X satisfy*

$$\|X\|_{L^p} = (\mathbf{E}|X|^p)^{1/p} \leq K_2\sqrt{p}, \quad \textit{for all } p \geq 1.$$

3. *The MGF of X^2 satisfies*

$$\mathbf{E}[\exp(\lambda^2 X^2)] \leq \exp(K_3^2\lambda^2) \quad \textit{for all } \lambda \textit{ such that } |\lambda| \leq \tfrac{1}{K_3}$$

4. *The MGF of X^2 is bounded at some point, namely*

$$\mathbf{E}[\exp(X^2/K_4^2)] \leq 2.$$

Moreover, if $\mathbf{E}[X] = 0$, then properties (1)-(4) are also equivalent to the following one

5. *The MGF of X satisfies*

$$\mathbf{E}[\exp(\lambda X)] \leq \exp(K_5^2\lambda^2) \quad \textit{for all } \lambda \in \mathbb{R}.$$

Remarks: The constant 2 that appears in some properties above does not have special meaning; it can be replaced by any constant. The constant K_i's depend on the distribution.

Definition 4.2 (Sub-gaussian r.v.)**.** A r.v. X that satisfies one of the properties in Theorem 4.1 is called a *sub-gaussian* random variable.

The *sub-gaussian norm* of X, denoted by $\|X\|_{\psi_2}$ is defined to be the smallest K_4 in Property (4):

$$\|X\|_{\psi_2} = \inf\{t \geq 0 \,:\, \mathbf{E}[\exp(X^2/t^2)] \leq 2\}.$$

Lemma 4.3. *The sub-gaussian norm is a norm on the space of sub-gaussian random variables.*

Proof. It is clear that the scaling and zero properties of a norm hold. We just need to prove the triangle inequality. Let X, Y be two sub-gaussian random variables. Fix $a, b > 0$. Set $f(x) = e^{x^2}$. Since f is convex and increasing, we have that

$$\begin{aligned}
\text{(increasing)} \qquad f\left(\frac{|X+Y|}{a+b}\right) &\leq f\left(\frac{|X|+|Y|}{a+b}\right) \\
&= f\left(\frac{a}{a+b}\cdot\frac{|X|}{a} + \frac{b}{a+b}\cdot\frac{|Y|}{b}\right) \\
\text{(convexity)} \qquad &\leq \frac{a}{b+a} f\left(\frac{|X|}{a}\right) + \frac{b}{a+b} f\left(\frac{|Y|}{b}\right).
\end{aligned}$$

We set $a = \|X\|_{\psi_2}$ and $b = \|Y\|_{\psi_2}$ and take expectations on both sides

$$\begin{aligned}
&\mathbf{E}\left[\exp\left(\frac{(X+Y)^2}{(\|X\|_{\psi_2}+\|Y\|_{\psi_2})^2}\right)\right] \\
&\qquad \leq \frac{\|X\|_{\psi_2}}{\|X\|_{\psi_2}+\|Y\|_{\psi_2}} \mathbf{E}\left[\exp\left(\frac{X^2}{\|X\|_{\psi_2}^2}\right)\right] + \frac{\|Y\|_{\psi_2}}{\|X\|_{\psi_2}+\|Y\|_{\psi_2}} \mathbf{E}\left[\exp\left(\frac{Y^2}{\|Y\|_{\psi_2}^2}\right)\right] \\
&\qquad \leq \frac{\|X\|_{\psi_2}}{\|X\|_{\psi_2}+\|Y\|_{\psi_2}} \cdot 2 + \frac{\|Y\|_{\psi_2}}{\|X\|_{\psi_2}+\|Y\|_{\psi_2}} \cdot 2 = 2
\end{aligned}$$

Thus, the constant $\|X\|_{\psi_2} + \|Y\|_{\psi_2}$ is in the set $\{t \geq 0 \,:\, \mathbf{E}[\exp((X+Y)^2/t^2)] \leq 2\}$. The result follows. □

We can restate the constants in Theorem 4.1 in terms of the norm $\|X\|_{\psi_2}$. It states that every sub-gaussian r.v. X satisfies the following bounds:

$$\Pr(|X| \geq t) \leq 2\exp(-ct^2/\|X\|_{\psi_2}^2) \quad \text{for all } t \geq 0 \tag{4.1}$$

$$\|X\|_{L^p} \leq C\|X\|_{\psi_2}\sqrt{p} \quad \text{for all } p \geq 1 \tag{4.2}$$

$$\mathbf{E}(\exp(X^2/\|X\|_{\psi_2}^2)) \leq 2 \tag{4.3}$$

$$\text{if } \mathbf{E}X = 0 \text{ then } \mathbf{E}(\exp(\lambda X)) \leq \exp(C\lambda^2\|X\|_{\psi_2}^2) \quad \text{for all } \lambda \in \mathbb{R} \tag{4.4}$$

Here $c, C > 0$ are absolute constants (i.e they do NOT depend on the distribution). Moreover, up to absolute constant factors, $\|X\|_{\psi_2}$ is the smallest possible number that makes each of these inequalities valid.

Examples of sub-gaussian distributions.

- **Gaussian.** $X \sim N(0,1)$ is a sub-gaussian random variable with $\|X\|_{\psi_2} \leq C$, where C is an absolute constant. This follows from Theorem 2.3 where we explicitly computed the tails of a Gaussian. More generally, if $X \sim N(0, \sigma^2)$ then X is sub-gaussian with

$$\|X\|_{\psi_2} \leq \sigma C.$$

- **Bernoulli.** Let X be a Radamacher r.v. Since $|X| = 1$, it follows that X is sub-gaussian with

$$\|X\|_{\psi_2} = \frac{1}{\sqrt{\ln 2}}.$$

- **Bounded.** Any bounded r.v. X is sub-gaussian with

$$\|X\|_{\psi_2} \leq C \|X\|_\infty$$

where $C = 1/\sqrt{\ln(2)}$.

Remarks.

- Poisson, exponential, Cauchy distributions are not sub-gaussian
- Centering a sub-gaussian r.v. is a sub-gaussian r.v.

Lemma 4.4. *If X is a sub-gaussian r.v. then $X - \mathbf{E}[X]$ is sub-gaussian and*

$$\|X - \mathbf{E}[X]\|_{\psi_2} \leq C \|X\|_{\psi_2}$$

where C is an absolute constant.

- Sum of sub-gaussian r.v. is a sub-gaussian r.v.

Lemma 4.5. *Let $X_1, X_2, \ldots, X_N$ be independent, mean zero, sub-gaussian r.v. Then $\sum_{i=1}^N X_i$ is also a sub-gaussian r.v. and*

$$\left\| \sum_{i=1}^N X_i \right\|_{\psi_2}^2 \leq C \sum_{i=1}^N \|X_i\|_{\psi_2}^2$$

where C is an absolute constant.

Proof. We use property (5) of Theorem 4.1. Let us analyze the MGF of the sum. For any $\lambda \in \mathbb{R}$, we have

$$\begin{aligned}
\mathbf{E}\left[\exp(\lambda \sum_{i=1}^N X_i)\right] &= \prod_{i=1}^N \mathbf{E}[\exp(\lambda X_i)] \quad \text{(by independence)} \\
&\leq \prod_{i=1}^N \exp(C\lambda^2 \|X_i\|_{\psi_2}^2)] \quad \text{(by sub-gaussian property (4.1))} \\
&\leq \exp(\lambda^2 K^2) \quad \text{where } K^2 := C \sum_{i=1}^N \|X_i\|_{\psi_2}^2.
\end{aligned}$$

Therefore, the sum is sub-gaussian since this is a characterization of sub-gaussian. Moreover by the definition of the sub-gaussian norm and the equivalences of and , we immediately have

$$\left\| \sum_{i=1}^{N} X_i \right\|_{\psi_2} \leq C_1 K$$

where C_1 is an absolute constant. The proposition is proved. □

The concept of sub-gaussian distributions can be extended higher dimensions through projections onto lines.

Definition 4.6 (Sub-gaussian random vectors)**.** A random vector $X \in \mathbb{R}^n$ is called *sub-gaussian* if the one-dimensional marginals $\langle X, x \rangle$ are sub-gaussian random variables for all $x \in \mathbb{R}^n$. The *sub-gaussian norm* of X is defined as

$$\|X\|_{\psi_2} = \sup_{x \in \mathbb{S}^{n-1}} \|\langle X, x \rangle\|_{\psi_2}.$$

Remarks: A simple example of a sub-gaussian random vector is a random vector with independent sub-gaussian coordinates.

- Sub-gaussian distributions with independent coordinates ⇒ sub-gaussian. Let $X = (X_1, X_2, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with independent, mean zero, sub-gaussian coordinates X_i. Then X is sub-gaussian random vector and $\|X\|_{\psi_2} \leq C \max_{i \leq n} \|X_i\|_{\psi_2}$.

There are many examples of high-dimensional sub-gaussian distributions. We state some basic examples below:

- **Multivariate normal:** We know that a random vector $Z = (Z_1, Z_2, \ldots, Z_n)$ has the *standard normal distribution* in $\mathbb{R}^n$, denoted by $Z \sim N(0, I_n)$ if the coordinates Z_i are independent standard normal random variables $N(0, 1)$. The density of Z is then the product of the n standard normal densities. Note that the standard normal density is rotation invariant:

  Proposition 4.7 (Rotation invariance of standard normal)**.** *Consider a random vector $Z \sim N(0, I_n)$ and a fixed orthogonal matrix U then $UZ \sim N(0, I_n)$. Moreover for any fixed vector $u \in \mathbb{R}^n$, we have $\langle Z, u \rangle \sim N(0, \|u\|_2^2)$.*

  Let $\Sigma \in \mathbb{R}^{n \times n}$ be a positive-semidefinite matrix and $\mu \in \mathbb{R}^n$. A random vector X is said to be a *general normal distribution*, denoted $X \sim N(\mu, \Sigma)$ if $Z := \Sigma^{1/2}(X - \mu) \sim N(0, I_n)$. The density of $X \sim N(\mu, \Sigma)$ can be computed by the change of variables formula

$$f_X(x) = \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} \exp(-(x - \mu)^T \Sigma^{-1} (x - \mu)/2), \qquad x \in \mathbb{R}^n.$$

  The following proposition proves that the multivariate normal is sub-gaussian.

  Proposition 4.8 (Characterization of normal distribution)**.** *Let X be a random vector in $\mathbb{R}^n$. Show that X has a multivariate normal distribution if and only if every one-dimensional marginal $\langle X, \theta \rangle$, $\theta \in \mathbb{R}^n$ has a (univariate) normal distribution.*

  Another important property of multivariate normals is the following.

Proposition 4.9. *If $Z \sim N(0, I_n)$, then for any fixed vectors $u, v \in \mathbb{R}^n$, we have*

$$\mathbf{E}[\langle X, u\rangle\langle X, v\rangle] = \langle u, v\rangle.$$

- **Symmetric Bernoulli:** A random vector $X = (X_1, X_2, \ldots, X_n) \in \mathbb{R}^n$ is a symmetric Bernoulli distribution if the coordinates X_i are independent, symmetric Bernoulli random variables.
- **Spherical distribution:** We say a random vector X is spherically distributed, denoted $X \sim \text{Unif}(\sqrt{n}\mathbb{S}^{n-1})$ if the random vector is uniformly distributed on the Euclidean sphere in $\mathbb{R}^n$ with radius $\sqrt{n}$ and centered at the origin. To show this is sub-gaussian, we reduce this to the Gaussian distribution, $N(0, I_n)$ (i.e. $Z \sim N(0, I_d)$, then $Z/\left\|Z\right\|_2$ is uniformly distributed on the unit sphere. Similarly, the uniform distribution on the Euclidean ball $B(0, \sqrt{n})$ in $\mathbb{R}^n$ is sub-gaussian.
- **Uniform distribution over the unit cube** $[-1, 1]^n$**:** This follows from independence of coordinates.

Not all uniform distributions over convex sets are sub-gaussian. For instance, a n ball of the ℓ_1-norm in $\mathbb{R}^n$ ($\left\|x\right\|_1 \leq n$) is isotropic, but not sub-gaussian.

Theorem 4.10 (General Hoeffding's inequality)**.** *Let $X_1, \ldots, X_N$ be independent, mean zero, sub-gaussian random variables. Then for every $t \geq 0$, we have*

$$\Pr\left(\left|\sum_{i=1}^N X_i\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{\sum_{i=1}^N \left\|X_i\right\|_{\psi_2}^2}\right).$$

Theorem 4.11 (Khintchine's inequality)**.** *Let $X_1, \ldots, X_N$ be independent sub-gaussian random variables with zero means and unit variances, and let $(a_1, a_2, \ldots, a_N) \in \mathbb{R}^N$. Then*

$$\text{for } p \in [2, \infty) \qquad \left\|a\right\|_2 \leq \left\|\sum_{i=1}^N a_i X_i\right\|_{L^p} \leq CK\sqrt{p}\left\|a\right\|_2$$

$$\text{for } p = 1 \qquad c(K)\left\|a\right\|_2 \leq \left\|\sum_{i=1}^N a_i X_i\right\|_{L^p} \leq \left\|a\right\|_2.$$

Here $K = \max_i \left\|X_i\right\|_{\psi_2}$, C is an absolute constant, and $c(K) > 0$ is a quantity which may depend only on K.

4.2 Sub-exponential distributions

Although the sub-gaussian distribution is quite large, it leaves out a lot o f important distributions with heavier tails than gaussian (for instance exponential and poisson). As we saw, we are often interested in the norms of a random vector. Suppose $Z_i \sim N(0, 1)$ and consider the Euclidean norm

$$\left\|Z\right\|_2 = \left(\sum_{i=1}^n Z_i^2\right)^{1/2}.$$

On the one hand, $\left\|Z\right\|_2^2$ is a sum of independent random variables Z_i^2 so we expect some concentration to hold. On the other hand, even though Z_i are sub-gaussian Z_i^2 are not:

$$\Pr(Z_i^2 > t) = \Pr(|Z_i| > \sqrt{t}) \sim \exp(-(\sqrt{t})^2/2) = \exp(-t/2).$$

The tails behave more like an exponential distribution and are strictly heavier than sub-gaussian.

Definition 4.12 (Sub-exponential random variables)**.** A random variable $X \in \mathbb{R}$ is called *sub-exponential* if

$$\inf\{t > 0 : \mathbf{E}[\exp(|X|/t)] \leq 2\} < \infty.$$

One then can define the sub-exponential norm to be

$$\|X\|_{\psi_1} = \inf\{t > 0 : \mathbf{E}[\exp(|X|/t)] \leq 2\}.$$

Properties similar to those in the sub-gaussian random variables hold true. For us, the most important thing is that there is a clear relationship between sub-gaussians and sub-exponentials.

- A random variable X is sub-gaussian iff X^2 is sub-exponential: $\left\|X^2\right\|_{\psi_1} = \|X\|_{\psi_2}^2$.
- Product of sub-gaussians is sub-exponential: $\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2}$.
- Centering: $X - \mathbf{E}[X]$ is sub-exponential if X is sub-exponential.
- Examples of sub-exponentials include Poisson and Exponential distributions.

Theorem 4.13 (Bernstein's inequality)**.** *Let $X_1, \ldots, X_n$ be independent, mean zero, sub-exponential random variables. Then, for every $t \geq 0$, we have*

$$\Pr\left(|\sum_{i=1}^n X_i| \geq t\right) \leq 2\exp\left(-c\min\left\{\frac{t^2}{\sum_{i=1}^n \|X_i\|_{\psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}}\right\}\right)$$

In particular, this implies that

$$\Pr\left(|\frac{1}{n}\sum_{i=1}^n X_i| \geq t\right) \leq 2\exp\left(-cn\min\left\{\frac{t^2}{K^2}, \frac{t}{K}\right\}\right), \quad \text{where } K = \max_i \|X_i\|_{\psi_1}.$$

Remarks: Bernstein's inequality is a mixture of sub-gaussian and sub-exponential tails. The sub-gaussian tail is of course expected from the CLT, but the sub-exponential tails are too heavy to produce a single sub-gaussian tail everywhere. In fact, the sub-exponential tail is produced by exactly one X_i. To put Bernstein's inequality in to perspective,

$$\Pr\left(|\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i| \geq t\right) \leq \begin{cases} 2\exp(-ct^2), & t \leq C\sqrt{n} \\ 2\exp(-t\sqrt{n}), & t \geq C\sqrt{n}. \end{cases}$$

We see that in the *small deviation* regime $t \leq C\sqrt{n}$, we have sub-gaussian tail bound as if the sum has a normal distribution with constant variance. This domain widens as N increases and the central limit theorem become more powerful. For *large deviations* where $t \geq C\sqrt{n}$, the sum has heaver sub-exponential tail bound, which be due to the contribution of a single term X_i.