High-Dimensional Probability for Data Science

Based on the PhD working group lectures

Daniele Zago

November 28, 2021

Contents

Lecture 1: Concentration inequalities

2021-11-20

The object of the first lectures is trying to characterize deviations of sums of random variables X_i w.r. to their expected value $\mathbb{E}$. These *concentration inequalities* take for instance the form of

$$\mathbb{P}(|S-\mu|>t) \leq \text{Bound},$$

where the bound is tighter than what we usually obtain using the standard inequalities that are presented in a first course in probability. In particular, we are not looking for asymptotic results as in the central limit theorem, but rather for estimates which are valid for any sample size N.

1.1 Hoeffding's inequality

Let us begin by recalling two standard inequalities which are going to be especially useful in the following sections.

Theorem 1 (Markov's inequality)
Let $X \geq 0$ be a random variable with finite expected value, $\mathbb{E}[X]<\infty$, then

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}, \quad \textit{for all } t>0.$$

A straightforward consequence of Markov's inequality can be obtained by replacing the random variable X with $|X-\mu|$ and squaring both sides inside the probability operator, which yields the following inequality.

Corollary 1 (Chebyshev's inequality)
If X is a random variable with finite variance, $\mathbb{V}[X]<\infty$, then

$$\mathbb{P}(|X-\mu| \geq t) \leq \frac{\mathbb{V}[X]}{t^2}.$$

Remark Many of the arguments that we make in this lecture will be based on the following trick: for any random variable X and for any $\lambda>0$,

$$\begin{aligned} \mathbb{P}(X-\mu \geq t) &= \mathbb{P}(e^{\lambda(X-\mu)} \leq e^{\lambda t}) && \text{(monotone)} \\ &\leq e^{-\lambda t}\mathbb{E}[e^{\lambda(X-\mu)}] && \text{(Markov)} \end{aligned}$$

Now, since it holds for any choice of $\lambda>0$ we can obtain the tightest bound by optimizing w.r. to λ,

$$\mathbb{P}(X-\mu \geq t) \leq \inf_{\lambda>0} e^{-\lambda t}\mathbb{E}[e^{\lambda(X-\mu)}],$$

and since X is usually a sum of random variables, its characteristic function can be decomposed into a product and evaluated quite easily.

Theorem 2 (Hoeffding's inequality)
Let $X_1, \dots, X_N$ be i.i.d Rademacher($\frac{1}{2}$) random variables and $a_1, \dots, a_N \in \mathbb{R}$, then for any $t > 0$ we have

$$\mathbb{P}\Big(\sum_{i=1}^{N} a_i X_i \geq t\Big) \leq \exp\left(-\frac{t^2}{2\|a\|_2^2}\right)$$

Sample size Unlike standard concentration inequalities based on the central limit theorem, this inequality gives an exact bound for any value of N.

Tightness Moreover, we can see that the tail behaviour, i.e. $\mathbb{P}(Y \geq t)$, is Gaussian-like in t, which means that this bound is extremely tight.

Proof.
Suppose that $\|a\|_2 = 1$, otherwise we can rescale t accordingly. For $\lambda > 0$, we have

$$\mathbb{P}\Big(\sum_{i=1}^{N} a_i X_i \geq t\Big) \overset{\text{Markov}}{\leq} e^{-\lambda t}\mathbb{E}[e^{\lambda \sum_{i=1}^{N} a_i X_i}]$$

$$= e^{-\lambda t} \prod_{i=1}^{N} \underbrace{\mathbb{E}[e^{\lambda a_i X_i}]}_{\frac{1}{2}e^{\lambda a_i} + \frac{1}{2}e^{-\lambda a_i}} \qquad \text{(Indep.)}$$

$$= e^{-\lambda t} \prod_{i=1}^{N} \cosh(\lambda a_i) \qquad (\tfrac{1}{2}e^{x} + \tfrac{1}{2}e^{-x} = \cosh(x))$$

$$\leq e^{-\lambda t} e^{\frac{\lambda^2}{2}\sum_{i=1}^{N} a_i^2} \qquad (\cosh(x) \leq e^{\frac{x^2}{2}}, \text{ see here})$$

Now, if we want to find the optimal bound, $\lambda_{\text{opt}} = \inf_{\lambda > 0} e^{-\lambda t + \frac{\lambda^2}{2}\|a\|_2^2}$, we first notice that the function inside the exponent is parabolic in λ,

$$f(\lambda) = -\lambda t + \frac{\lambda^2}{2}\|a\|_2^2 \overset{\text{parabola}}{\Longrightarrow} \lambda_{\text{opt}} = \frac{t}{\|a\|_2^2} \implies f(\lambda_{\text{opt}}) = -\frac{t^2}{2\|a\|_2^2}.$$

Therefore, by substituting the optimal λ we obtain the proof of Hoeffding's inequality,

$$\mathbb{P}\Big(\sum_{i=1}^{N} a_i X_i \geq t\Big) \leq e^{-\frac{t^2}{2\|a\|_2^2}}.$$

□

Exercise Restate Hoeffding's inequality for $X_1, \dots, X_N \overset{\text{iid}}{\sim} \text{Ber}(\frac{1}{2})$, using the fact that $Z_i = 2X_i - 1$ with $Z_i \sim \text{Rademacher}(\frac{1}{2})$.

Exercise Use Hoeffding's inequality for Bernoulli random variables to prove that by tossing a coin N times we have the exact bound

$$\mathbb{P}\Big(\text{at least } \frac{3}{4} \text{ heads}\Big) \leq e^{-N/8}.$$

Remark We can get a double bound from the above 2 by using $\mathbb{P}(|S| \geq t) \leq \mathbb{P}(S \geq t) + \mathbb{P}(-S \geq t)$, and observing that the Rademacher r.v. is symmetric $S = -S$. Therefore, both bounds are equal and the following two-sided inequality can be stated.

Theorem 3 (Two-sided Hoeffding's inequality)
Let $X_1, \ldots, X_N$ be i.i.d Rademacher r.v.'s, then for all $t \geq 0$ and for all $a \in \mathbb{R}^N$,

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{N} a_i X_i\Big| \geq t\Big) \leq 2\exp\left(-\frac{t^2}{2\|a\|_2^2}\right).$$

We now turn to the more general problem of bounded random variables, which include as a special case the setting of Bernoulli r.v.'s with varying parameter p_i.

Theorem 4 (Hoeffding's inequality for bounded r.v.'s)
Let $X_1, X_2, \ldots, X_N$ be independent but not identically distributed r.v.'s, such that $X_i \in [m_i, M_i]$ and $\mathbb{E}[X_i] < \infty$. Then, for all $t \geq 0$ the following inequality holds,

$$\mathbb{P}\Big(\sum_{i=1}^{N}(X_i - \mathbb{E}[X_i]) \geq t\Big) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{N}(M_i - m_i)^2}\right).$$

Proof.
(Exercise 2.2.7 in the book) The difficult part is achieving the constant 2 in the numerator, therefore we start with a different constant and then use a trick to get it. Let $\lambda > 0$, then by the same argument as before we can write

$$\begin{aligned}\mathbb{P}(\sum_{i=1}^{N}(X_i - \mathbb{E}[X_i]) \geq t) &\leq e^{-\lambda t}\mathbb{E}[e^{\lambda \sum_i X_i - \mathbb{E}[X_i]}] \\ &= e^{-\lambda t}\prod_i \mathbb{E}[e^{\lambda(X_i - \mathbb{E}[X_i])}] \\ &\leq e^{-\lambda t + \sum_i \lambda(M_i - m_i)}\end{aligned}$$

This is not as easy to optimize as before since we don't have a quadratic form, therefore we need a subtle trick to transform it into a more easily handled problem.

Trick In order to replace "$\cosh x \leq e^{x^2/2}$" we can use the following trick: Let Y be a r.v. with $\mathbb{E}[Y] = 0$ (our case of $X - \mathbb{E}[X]$) and $Y \in [a, b]$, then for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}[e^{\lambda Y}] \leq e^{\lambda^2 \frac{(b-a)^2}{2}}.$$

This is based on a symmetrization of Y by introducing another independent random variable $Y' \overset{\mathrm{d}}{=} Y$ and $Z \sim \text{Rademacher}(\frac{1}{2})$ from which we have $\mathbb{E}[e^{-\lambda Y'}] \overset{\text{Jens.}}{\leq} e^{-\lambda \mathbb{E}[Y]} = 1$, therefore

$$\mathbb{E}[e^{\lambda Y}] \leq \mathbb{E}[e^{\lambda Y}]\cdot\mathbb{E}[^{-\lambda Y'}] = \mathbb{E}[e^{\lambda(Y-Y')}] = \mathbb{E}[e^{\lambda Z(Y-Y')}] = \mathbb{E}[\cosh(\lambda(Y-Y'))] \leq \mathbb{E}[e^{\lambda^2 \frac{(Y-Y')^2}{2}}] = e^{\frac{\lambda^2(b-a)^2}{2}}.$$

Using this trick, we can optimize the equation using

$$\mathbb{P}\Big(\sum_{i=1}^{N}(X_i - \mathbb{E}[X_i]) \geq t\Big) \leq e^{-\lambda t}\prod_i e^{\lambda^2\frac{(M_i-m_i)^2}{2}}$$

$$= \exp\Big(-\lambda t + \frac{\lambda^2}{2}\sum_i \frac{(M_i-m_i)^2}{2}\Big).$$

We can optimize with $\lambda > 0$ and get the minimum with a different constant than 2. Finding this other minimum requires more work.

□

Example (Book 2.2.9 – Boosting a randomized algorithm)
We have an algorithm that gives the right answer out of two classes with a probability $\frac{1}{2}+\delta$, with $\delta > 0$. We run this algorithm N (odd) times and take the majority vote to get the final classification.

Problem Find the minimal N such that $\mathbb{P}(\text{correct answer}) \geq 1-\varepsilon$ for $\varepsilon \in (0,1)$ fixed.

Solution Consider the following r.v. $X_1, \ldots, X_N$ be the indicator of the wrong answer

$$X_i = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ run is wrong} \\ 0 & \textit{otherwise} \end{cases}$$

then, using theorem 4 with $t = N\delta$, $M_i = 1$ and $m_i = 0$ we can bound the probability of wrong answer as

$$\mathbb{P}\Big(X_1 + \ldots + X_N \geq \frac{N}{2}\Big) = \mathbb{P}\Big(\sum_{i=1}^{N}(X_i - (\frac{1}{2}-\delta)) \geq N\delta\Big) \overset{4}{\leq} \exp\left(-\frac{2N^{\not 2}\delta^2}{\not N}\right).$$

Therefore, in order to have the required bounded probability we need

$$-2N\delta^2 \leq \log\varepsilon \iff \boxed{N \geq \frac{1}{2\delta^2}\log\frac{1}{\varepsilon}}.$$

1.2 Chernoff's inequality

Consider the last Hoeffding's inequality (theorem 4), then for a sum of random variables we can write the Gaussian tail using the CLT as approximately

$$\mathbb{P}(|Z| \geq t) \leq 2e^{-\frac{t^2}{2}}.$$

Chernoff's inequality is useful in regimes of sums in order to prove a bound that is again independent from the central limit theorem. The following theorem is a merged result of Theorem 2.3.1, Exercise 2.3.2 and Exercise 2.3.5 in the book.

Theorem 5 (Chernoff's inequality)
Let $X_1, \dots, X_N$ be such that $X_i \overset{iid}{\sim} Bern(p_i)$ and consider the cumulative sum $S_N = \sum_i X_i$ with expected value $\mu = \mathbb{E}[S_N] = \sum_i p_i$. Then, the following inequalities hold:

$$\mathbb{P}(S_N \geq t) \leq e^{-\mu} \cdot \left(\frac{e\mu}{t}\right)^t \qquad \text{for } t > \mu,$$

$$\mathbb{P}(S_N \leq t) \leq e^{-\mu} \left(\frac{e\mu}{t}\right)^t \qquad \text{for } t < \mu,$$

$$\text{“Small deviations”:} \quad \mathbb{P}(|S_N - \mu| \geq \delta\mu) \leq 2e^{-C\mu\delta^2} \qquad \text{for } \delta \in (0, 1],$$

where C is a universal constant (i.e. does not depend on the other quantities).

Proof.

1. The first step is always the same, let $\lambda > 0$ then

$$\mathbb{P}(S_N \geq t) = \mathbb{P}(e^{\lambda S_N} \geq e^{\lambda t}) \leq e^{-\lambda t}\mathbb{E}[e^{\lambda S_N}] = e^{-\lambda t} \prod_i \mathbb{E}[e^{\lambda X_i}]. \tag{1}$$

Now for a Bernoulli random variable, $\mathbb{E}[e^{\lambda X_i}] = (1 - p_i)e^0 + p_i e^\lambda = 1 + (e^\lambda - 1)p_i$, and we use the following identity:

$$1 + x \leq e^x \quad \text{for all } x > 0,$$

to write

$$\mathbb{E}[e^{\lambda X_i}] = 1 + \overbrace{(e^\lambda - 1)p_i}^{x} \leq \exp\big((e^\lambda - 1)p_i\big).$$

Going back to (1), we have the following bound for any $\lambda > 0$,

$$\mathbb{P}(S_N \geq t) \leq e^{-\lambda t} e^{(e^\lambda - 1)\sum_i p_i} = e^{-\lambda t + \mu(e^\lambda - 1)}.$$

Again, by optimizing over λ we find that the tightest bound from (1) is given by

$$f(\lambda) = -\lambda t + \mu(e^\lambda - 1) \implies \lambda_{\text{opt}} = \underset{\lambda > 0}{\operatorname{argmin}} f(\lambda) = \log \frac{t}{\mu},$$

from which we obtain the first Chernoff bound,

$$\mathbb{P}(S_N \geq t) \leq e^{-\mu} \left(\frac{e\mu}{t}\right)^t.$$

2. For the second inequality, proceed as before using

$$\mathbb{P}(S_N \leq t) \overset{\lambda \geq 0}{=} \mathbb{P}(e^{-\lambda S_N} \geq e^{-\lambda t}).$$

3. We can obtain the bound on $\mathbb{P}(|S_N - \mu| \geq \delta\mu)$ by using the fact that

$$\mathbb{P}(|S_N - \mu| \geq \delta\mu) \leq \mathbb{P}(S_N - \mu \geq \delta\mu) + \mathbb{P}(S_N - \mu \leq -\delta\mu) \overset{(1),(2)}{\leq} \dots$$

□

Theorem 6 (Poisson tail behaviour)
Let $Z \sim Pois(\gamma)$ with $\gamma > 0$, i.e. X has probability mass function $\mathbb{P}(X = k) = e^{-\gamma}\frac{\gamma^k}{k!}$, for $k = 0, 1, \ldots$. Then,

1. *For all $\delta \in (0, 1]$ theorem 5-3 holds*

$$\mathbb{P}(|Z - \gamma| \geq \delta\gamma) \leq 2e^{-C\lambda\delta^2}$$

2. *Let now $t > \gamma$, then the following bound holds*

$$\mathbb{P}(X \geq t) \leq e^{-\gamma}\left(\frac{e\gamma}{t}\right)^t \tag{A}$$

Remark These bound are extremely useful in practical applications and is similar to Chernoff's bound (theorem 5), which works instead for a sum of Bernoulli variables.

Remark 2 If $p_i = \frac{\gamma}{N}$, then $S_N \approx Z \sim \text{Pois}(\gamma)$ for $N \gg 1$ and the rate of convergence is very fast, therefore this result could also be obtained as a limit. However, the above theorem is *exactly* valid.

Proof.
(Execise) Prove equation (A) using the basic trick $\mathbb{P}(X \geq t) \leq e^{-\lambda t}\mathbb{E}[e^{\lambda X}]$, which can be computed explicitly, and then optimize over $\lambda > 0$. Briefly comment on why this bound is optimal.

□

Lecture 2: Subgaussian random variables

2021-11-20

In this lecture we generalize Hoeffding's inequality to subgaussian random variables, which are a class of distributions that enjoy nice properties and are fundamental in the high-dimensional setting. We begin by recalling some properties of the Gaussian distribution

Prop. 1 (Properties of the gaussian distribution)
Let $X \sim \mathcal{N}(0,1)$, then the following statements hold:

1. *We have a tail estimate for X given by*

$$\left(\frac{1}{t} - \frac{1}{t^3}\right)\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}} \leq \mathbb{P}(X \geq t) \leq \frac{1}{t}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}, \quad t > 0.$$

   This estimate in particular implies that

$$\mathbb{P}(X \geq t) \leq \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}} \qquad t \geq 1,$$

$$\mathbb{P}(|X| \geq t) \leq 2e^{-\frac{t^2}{2}} \qquad t \geq 0.$$

2. *Given $p \geq 1$, we have that*

$$\|X\|_{L^p} = \mathbb{E}[|X|^p]^{\frac{1}{p}} = \sqrt{2}\left(\frac{\Gamma(\frac{1+p}{2})}{\Gamma(\frac{1}{2})}\right)^{\frac{1}{p}}$$

3. *The moment-generating function of X is $\mathbb{E}[e^{\lambda X}] = e^{\frac{\lambda^2}{2}}$ for all $\lambda \in \mathbb{R}$.*

Corollary 2 (Bounded norm of a gaussian r.v.)
If $X \sim \mathcal{N}(0,1)$ there exists a $C > 0$ such that $\|X\|_{L^p} \leq C\sqrt{p}$ for all $p \geq 1$.

Proof.
Use Stirling's approximation for the Gamma function to obtain the bound.

□

With these properties we can now discuss another class of random variables, which include the Gaussian distribution.

2.1 Space of subgaussian random variables

We begin the analysis of subgaussian random variables by stating a sequence of equivalent properties that turn out to be equivalent to each other.

Theorem 7 (Equivalence of properties for subgaussian r.v.'s)
Let X be a generic random variable, then the following properties are equivalent:

1. *(Tail of X) There exists a $K_1 > 0$ such that $\mathbb{P}(|X| > t) \leq 2e^{-t^2/k_1^2}$ for all $t \geq 0$.*
2. *(Moments of X) There exists a $k_2 > 0$ such that $\|X\|_{L^p} \leq k_2\sqrt{p}$ for all $p \geq 1$.*
3. *(MGF of X^2) There exists a $k_3 > 0$ such that $\mathbb{E}[e^{\lambda^2 X^2}] \leq e^{k_3^2\lambda^2}$ for $|\lambda| \leq \frac{1}{k_3}$.*
4. *(MGF of X^2) There exists a $k_4 > 0$ such that $\mathbb{E}[e^{X^2/k_4^2}] \leq 2$.*

In addition, if $\mathbb{E}[X] = 0$ we can add another equivalent property:

5. *(MFG of X) There exists a $k_5 > 0$ such that $\mathbb{E}[e^{\lambda X}] \leq e^{k_5\lambda^2}$ for all $\lambda \in \mathbb{R}$.*

Moreover, the above constants $k_1, \dots, k_5$ differ by a constant factor, i.e. if one property holds then all properties hold and $\exists C_{ij} > 0$ such that

$$k_i \leq C_{ij}k_j \quad \text{for all } i, j, \text{ with a } C_{ij} \text{ that does not depend on } X.$$

Proof.
Long and boring.

□

Remark *5.* really needs that $\mathbb{E}[X] = 0$, otherwise it does not work independently of X.

Given the usefulness of these bounds, it's important to isolate the class of r.v.'s that share these properties.

Def. (Subgaussian r.v.)
A r.v. X is called ***subgaussian*** if it satisfies one of the equivalent properties in theorem 7.

Theorem 8 (Subgaussian random variables form a vector space)
The set of subgaussian random variables is a vector space, which means that

$$X, Y \text{ subgaussian} \implies X + Y \text{ is subgaussian}$$

$$X \text{ subgaussian} \implies \alpha X \text{ is subgaussian}$$

Proof.
We aim to prove that $\mathbb{E}[e^{\frac{(X+Y)^2}{(a+b)^2}}] \leq 2$, we can consider

$$\frac{X+Y}{a+b} = \frac{a}{a+b}\frac{X}{a} + \frac{b}{b+a}\frac{Y}{b},$$

use the fact that e^{x^2} is convex to conclude that

$$e^{\frac{(x+y)^2}{(a+b)^2}} \leq \frac{a}{a+b}e^{\frac{x^2}{a^2}} + \frac{b}{a+b}e^{\frac{y^2}{b^2}}.$$

□

Def. (Subgaussian norm)
Let X be a subgaussian r.v., then we define the ***subgaussian norm of X*** as

$$\|X\|_{\psi_2} := \inf\Big\{t > 0 : \mathbb{E}[e^{X^2/t^2}] \leq 2\Big\}.$$

Remark Take $t = k_4$ and we see that the set over which the inf is taken is never empty.

Remark 2 By dominated convergence this infimum is a minimum.

Prop. 2 (Subgaussian norm is indeed a norm)
$\|\cdot\|_{\psi_2}$ is a norm on the space of subgaussian r.v.'s.

Proof.
Everything is simple, except for the triangle inequality which is not straightforward.

□

Finally, we have a last observation which

Prop. 3 (Subgaussian r.v.'s are a Banach space)
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $V = \{X \text{ r.v subgaussian on } \Omega\}$ and $\|\cdot\|_{\psi_2}$ as defined above. Then, $(V, \|\cdot\|_{\psi_2})$ is a Banach space.

Since we have that the optimal constant for property *4.* is given by the subgaussian norm $\|X\|_{\psi_2}$, then we have the following updated set of inequalities in terms of $k_4 = \|X\|_{\psi_2}^2$:

1. $\mathbb{P}(|X| > t) \leq 2e^{-\frac{Ct^2}{\|X\|_{\psi_2}^2}}$ for all $t \geq 0$.
2. $\|X\|_{L^p} \leq C\|X\|_{\psi_2}\sqrt{p}$ for all $p \geq 1$.
3. $\mathbb{E}[e^{\frac{X^2}{\|X\|_{\psi_2}^2}}] \leq 2$.
4. If $\mathbb{E}[X] = 0$, then $\mathbb{E}[e^{\lambda X}] \leq e^{C\lambda^2\|X\|_{\psi_2}^2}$.

Prop. 4 (Bounded r.v.'s are subgaussian)
If X is a bounded random variable then X is subgaussian.

Proof.
$\|X\|_{\psi_2} \leq \dfrac{\|X\|_\infty}{\log 2}$.

□

Non-subgaussian r.v.'s Poisson, exponential, Pareto, Cauchy, ...
For subgaussian random variables we have something similar to the property of Gaussian random variables

Prop. 5 (Sums of subgaussians)
Let $X_1, \ldots, X_N$ be i.i.d subgaussian random variables with $\mathbb{E}[X_i] = 0$ for all i. Then, $\sum_{i=1}^N X_i$ is subgaussian and

$$\Big\| \sum_{i=1}^N X_i \Big\|_{\psi_2}^2 \leq C \sum_{i=1}^N \|X_i\|_{\psi^2}^2.$$

Moreover, since $\|\cdot\|_{\psi_2}^2$ is a norm, we also have the following bound for free:

$$\Big\| \sum_{i=1}^N X_i \Big\|_{\psi_2}^2 \leq C \sum_{i=1}^N \|X_i\|_{\psi^2}.$$

Proof.
Since $\mathbb{E}[X_i] = 0$ then also $\mathbb{E}[\sum_i X_i] = 0$ and we use property *5.* to show

$$\begin{aligned} \mathbb{E}[e^{\lambda \sum_i X_i}] &\overset{5.}{\leq} \prod_i e^{C\lambda^2 \|X_i\|_{\psi_2}^2} \\ &= e^{C\lambda^2 \sum_i \|X_i\|_{\psi_2}^2}. \end{aligned}$$

Moreover, since the best constant is k_4 we have the norm.

□

2.2 General Hoeffding's inequality

Subgaussian random variables are extremely useful since we have a general form of the Hoeffding's inequality without passing through Rademacher or boundedness.

Theorem 9 (General Hoeffding's inequality)
Let $X_1, \ldots, X_N$ be independent subgaussian random variables with $\mathbb{E}[X_i] = 0$ for all i. Then, for each $t \geq 0$ we have a tail estimate

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

Proof.
Using the previous Prop. 5, we have that $X := \sum_{i=1}^N X_i$ is a subgaussian r.v. and we can write

$$\mathbb{P}(|X| > t) \leq 2e^{-\frac{Ct^2}{\|X\|_{\psi_2}^2}}, \quad \text{for all } t \geq 0.$$

Using the bound on the norm given by Prop. 5 and taking for instance .

□

Corollary 3 (General Hoeffding's inequality 2)
Let $X_1, X_2, \ldots, X_n$ be independent subgaussian random variables with $\mathbb{E}[X_i] = 0$, and let $a_1, a_2, \ldots, a_n \in \mathbb{R}$. Then,

$$\mathbb{P}\Big(\big|\sum_{i=1}^{N} a_i X_i\big| \geq t\Big) \leq 2\exp\left(-\frac{ct^2}{k^2\|a\|_2^2}\right),$$

where $k = \max_i \|X_i\|_{\psi_2}^2$.

Proof.
Use again the same properties, recall the homogeneity property of the norm and then bound using the maximum of the $|a_i|$'s.

□

Note We can also apply the theorem to general $X_1, \ldots, X_N$ independent and subgaussian but we need to replace X_i by $X_i - \mathbb{E}[X_i]$ beforehand.

Recall that $\|X - \mathbb{E}[X]\|_{L^2} \leq \|X\|_{L^2}$. This does not hold for the subgaussian norm, however we do have a lemma in this direction.

Lemma 1 (Centering of a subgaussian r.v.)
Let X be subgaussian, then $X - \mathbb{E}[X]$ is subgaussian (vector space) and

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

Proof.
$\|\cdot\|_{\psi_2}$ is a norm, therefore

$$\begin{aligned}
\|X - \mathbb{E}[X]\|_{\psi_2} &\leq \|X\|_{\psi_2} + \|\mathbb{E}[X]\|_{\psi_2} \\
&= \|X\|_{\psi_2} + |\mathbb{E}[X]| \cdot \|1\|_{\psi_2} \\
&\leq \|X\|_{\psi_2} + \|X\|_{L^1} \cdot \|1\|_{\psi_2} && (|\mathbb{E}[X]| \leq \mathbb{E}[|X|] = \|X\|_{L^1}) \\
&\leq \|X\|_{\psi_2} + C \cdot \|X\|_{\psi_2} \cdot \sqrt{1} \cdot \|1\|_{\psi_2} && (\text{using } \mathit{2.}) \\
&\leq K\|X\|_{\psi_2}.
\end{aligned}$$

□

Lecture 3: Geometry of random vectors

2021-11-21

If we consider a Gaussian distribution, $X \sim \mathcal{N}(0,1)$, then we might be interested in the length of $\|X\|_2^2$. However, $X^2 \sim \chi_1^2$ but this is not a subgaussian distribution:

$$\mathbb{P}(X^2 > t) = \mathbb{P}(|X| > \sqrt{t}) \geq C\left(\frac{1}{t^{1/2}} - \frac{1}{t^{3/2}}\right)\frac{1}{\sqrt{2\pi}}e^{-t/2} \not\lesssim 2e^{-t^2/k_1^2},$$

which violates the lower bound *1.* of the subgaussian random variable.

3.1 Subexponential random variables

We start with a characterization of a set of properties:

Theorem 10 (Equivalence of properties for subexponential r.v.'s)
Let X be a generic random variable, then the following properties are equivalent:

1. (Tail of X) *There exists a $k_1 > 0$ such that $\mathbb{P}(|X| > t) \leq 2e^{-t/k_1}$ for all $t \geq 0$.*
2. (Moments of X) *There exists a $k_2 > 0$ such that $\|X\|_{L^p} \leq k_2 p$ for all $p \geq 1$.*
3. (Mgf of $|X|$) *There exists a $k_3 > 0$ such that $\mathbb{E}[e^{\lambda|X|}] \leq e^{k_3\lambda}$ for $|\lambda| \leq \frac{1}{k_3}$.*
4. (Mgf of $|X|$) *There exists a $k_4 > 0$ such that $\mathbb{E}[e^{|X|/k_4}] \leq 2$.*

In addition, if $\mathbb{E}[X] = 0$ we can add another equivalent property:

5. (Mfg of X) *There exists a $k_5 > 0$ such that $\mathbb{E}[e^{\lambda X}] \leq e^{k_5\lambda^2}$ for all $|\lambda| \leq \frac{1}{k_5}$.*

Moreover, the above constants $k_1, \ldots, k_5$ differ by a constant factor, i.e. if one property holds then all properties hold and $\exists C_{ij} > 0$ such that

$$k_i \leq C_{ij}k_j \quad \textit{for all } i,j, \textit{ with a } C_{ij} \textit{ that does not depend on } X.$$

Remark Property *5.* changes in condition since the mgf might not exist for all $\lambda \in \mathbb{R}$.

Def. (Subexponential r.v.'s)
A random variable X satisfying one (and therefore all) of the above properties is called ***subexponential***.

Def. (Subexponential norm)
Given X subexponential r.v., we define the ***subexponential norm*** as

$$\|X\|_{\psi_1} = \inf\{t > 0 : \mathbb{E}[e^{|X|/t}] \leq 2\}.$$

Prop. 6
The set of subexponential random variables equipped with the $\|\cdot\|_{\psi_1}$ norm is a Banach space.

Example (Subgaussian $\implies$ subexponential)
Any subgaussian random variable is also subexponential, for example take any property above in theorem 10 and check it.

Example (Exponential)
The exponential r.v. is subexponential, indeed

$$X \sim \text{Exp}(\gamma) \implies \mathbb{P}(X \geq t) = e^{-\gamma t}$$

Example (Poisson)
The Poisson r.v. is subexponential, since

$$X \sim \text{Pois}(\gamma) \implies \mathbb{E}[e^{\lambda X}] = e^{\gamma} e^{\gamma e^{\lambda}} \leq e^{k_5 \lambda}.$$

There is a deep connection between subexponential and subgaussian random variables, summarized by the following lemma.

Lemma 2 (Subgaussian square is subxeponential)
A r.v. X is subgaussian $\iff$ X^2 is subexponential, moreover

$$\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2$$

Proof.
If we consider the subexponential norm, we have

$$\begin{aligned}\|X^2\|_{\psi_1} &= \inf\{t > 0 : \mathbb{E}[e^{X^2/t}] \leq 2\} \\ &= \inf\{k^2 > 0 : \mathbb{E}[e^{X^2/k^2}] \leq 2\}.\end{aligned}$$

□

Lemma 3 (Product of subgaussians)
Let X, Y be subgaussian r.v.'s not necessarily independent, then $X \cdot Y$ is subexponential and

$$\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2}.$$

Proof.
Without loss of generality we take $\|X\|_{\psi_2} = \|Y\|_{\psi_2} = 1$ (by bilinearity), then we have to prove that

$$\|XY\|_{\psi_1} \leq 1.$$

Equivalently, we have $\|X\|_{\psi_2} = 1 = \|Y\|_{\psi_2}$ that implies

$$\mathbb{E}[e^{X^2}] \leq 2, \quad \mathbb{E}[e^{Y^2}] \leq 2,$$

we want to prove that

$$\mathbb{E}[e^{|XY|}] \leq 2.$$

We use the fact that $ab \leq \frac{a^2+b^2}{2}$ by Young's inequality, therefore

$$|XY| \leq \frac{X^2}{2} + \frac{Y^2}{2},$$

$$\mathbb{E}[e^{|XY|}] \overset{\text{Y.}}{\leq} \mathbb{E}[e^{\frac{X^2}{2}} e^{\frac{Y^2}{2}}] \overset{\text{Y.}}{\leq} \frac{\mathbb{E}[e^{X^2}]}{2} + \frac{\mathbb{E}[e^{Y^2}]}{2} \leq \frac{2}{2} + \frac{2}{2} = 2.$$

□

Prop. 7 (Centering)

There exists a $C > 0$ such that for all X subexponential,

$$\|X - \mathbb{E}[X]\|_{\psi_1} \leq C\|X\|_{\psi_1}$$

Proof.
Analogous to subgaussian.

□

We consider now an inequality for subexponential random variables, which implies a part on subgaussian random variables.

Remark Consider a bounded r.v. X, then its moment-generating function is

$$\begin{aligned}\mathbb{E}[e^{\lambda X}] &\overset{\lambda \approx 0}{\approx} \mathbb{E}[1 + \lambda X + \frac{\lambda^2}{2} X^2 + o(\lambda^2 X^2)] \\ &= 1 + \frac{\lambda^2}{2}\mathbb{E}[X^2] + o(\lambda^2) \\ &\approx e^{\frac{\lambda^2}{2}\mathbb{E}[X^2]}\end{aligned}$$

This property is very similar to property *5.* of subexponential and subgaussian random variables.

Theorem 11 (Bernstein's inequality)

Let $X_1, X_2, \ldots, X_n$ be independent, mean-zero subexponential r.v.'s. Then, for all $t > 0$ we have

$$\mathbb{P}\Big(\big|\sum_{i=1}^{N} X_i\big| \geq t\Big) \leq 2\exp\left(-c \cdot \min\left\{\frac{t^2}{\sum_{i=1}^{N} \|X_i\|_{\psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}}\right\}\right).$$

Proof.
We use property *5.* to write

$$\begin{aligned}\mathbb{P}(S \geq t) &\leq e^{-\lambda t} \prod_{i=1}^{N} \mathbb{E}[e^{\lambda X_i}] \\ &\overset{5.}{\leq} e^{-\lambda t} \prod_{i=1}^{N} e^{C\lambda^2 \|X_i\|_{\psi_1}^2} \qquad (\text{for } |\lambda| \leq \frac{C}{\|X_i\|_{\psi_1}}) \\ &\leq e^{-\lambda t + C\lambda^2 \sum_i \|X_i\|_{\psi_1}^2}\end{aligned}$$

Now, if in the worst case $\widehat{\lambda} = \frac{C}{\|X_i\|_{\psi_1}}$ is to the right of the minimum of the parabola, we have to take it instead of minimizing the parabola.

$$\lambda_{\text{opt}} = \begin{cases} \frac{t}{2C\sum_i \|X_i\|_{\psi_1}^2} & \text{if } \widehat{\lambda} \geq \frac{t}{2C\sum_i \|X_i\|_{\psi_1}} \\ \widehat{\lambda} & \text{if } 0 < \widehat{\lambda} < \dots \end{cases}$$

□

Replacing X_i with $a_i X_i$ in Bernstein's inequality above, we get the more general bound.

Theorem 12 (Bernstein's inequality for weighted sums)
Let $X_1, X_2, \dots, X_n$ be independent, mean-zero subexponential r.v.'s. Then, for all $t > 0$ we have

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{N} a_i X_i\Big| \geq t\Big) \leq 2\exp\left(-c \cdot \min\left\{\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right\}\right),$$

where $K = \max_i \|X_i\|_{\psi_1}$.

Corollary 4 (Special case of Bernstein's inequality)
Choosing $a_i = \frac{1}{N}$ in theorem 12 we have a quantitative law of large numbers for subexponential random variables,

$$\mathbb{P}\Big(\Big|\frac{1}{N}\sum_{i=1}^{N} X_i\Big| \geq t\Big) \leq 2\exp\left\{-cN \cdot \min\left\{\frac{t^2}{K^2}, \frac{t}{K}\right\}\right\},$$

where $K = \max_i \|X_i\|_{\psi_1}$.

Remark If we have subexponential random variables with mean zero, then we can avoid using K and simply write the following two-regime inequality by replacing t with $t/\sqrt{N}$,

$$\mathbb{P}\Big(\Big|\frac{1}{N}\sum_{i=1}^{N} X_i\Big| \geq t\Big) \leq \begin{cases} 2\exp\left(-ct^2\right) & \text{if } t \leq C\sqrt{N} \quad \text{SMALL DEVIATIONS} \\ 2\exp\left(-t\sqrt{N}\right) & \text{if } t \geq C\sqrt{N} \quad \text{LARGE DEVIATIONS} \end{cases}$$

where C and c can depend on $\|X\|_{\psi_1}$, but does not if they are i.i.d random variables.

3.2 Random vectors in high dimensions

Theorem 13 (Concentration of the norm)
Let $X \in \mathbb{R}^n$ be a random vector with independent subgaussian coordinates X_i such that $\mathbb{E}[X_i^2] = 1$. Then,

$$\left\| \|X\|_2 - \sqrt{n} \right\|_{\psi_2} \leq Ck^2, \tag{2}$$

where $k = \max_i \|X_i\|_{\psi_2}$.

Proof.
We can apply Bernstein inequality to see that by centering X_i^2,

$$\|X_i^2 - 1\|_{\psi_1} \overset{\text{center.}}{\leq} C\|X_i^2\|_{\psi_1} = C\|X_i\|_{\psi_2}^2 \leq CK^2,$$

and therefore

$$\mathbb{P}\left(\frac{1}{n}\|X\|_2^2 - 1 \geq u\right) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{(X_i^2 - 1)}_{\text{subexp}} \overset{\text{Cor.4}}{\leq} \leq 2\exp\left(-c \cdot n \cdot \min\left\{\frac{u}{C^2K}, \frac{u}{CK^2}\right\}\right).$$

Now, since $K \geq 1$ we have that $K^4 \geq K^2$ and by renaming the absolute constants,

$$\mathbb{P}\left(\frac{1}{n}\|X\|_2^2 - 1 \geq u\right) = 2\exp\left(-\frac{cn}{k^4} \cdot \min\left\{u^2, u\right\}\right).$$

Trick If we take $z \geq 0$ and $\delta \geq 0$, then a trivial trick yields

$$|z - 1| \geq \delta \implies |z^2 - 1| \geq \max\left\{\delta, \delta^2\right\}$$

...

□

Remark $\mathbb{E}[\|X\|_2^2] = \mathbb{E}[\sum_i X_i^2] = n$ so it's not surprising to see $\sqrt{n}$ above.

Equivalent Recall by the properties that

$$(2) \iff \mathbb{P}\left(\left|\|X\|_2 - \sqrt{n}\right| \geq t\right) \leq 2\exp\left(-\frac{-ct^2}{k^4}\right), \quad \text{for all } t \geq 0.$$

What is surprising is that t ***does not depend on n***, i.e. we can find a bound independent of n such that

$$\sqrt{n} - t_0 \leq \|X\|_2 \leq \sqrt{n} + t_0.$$

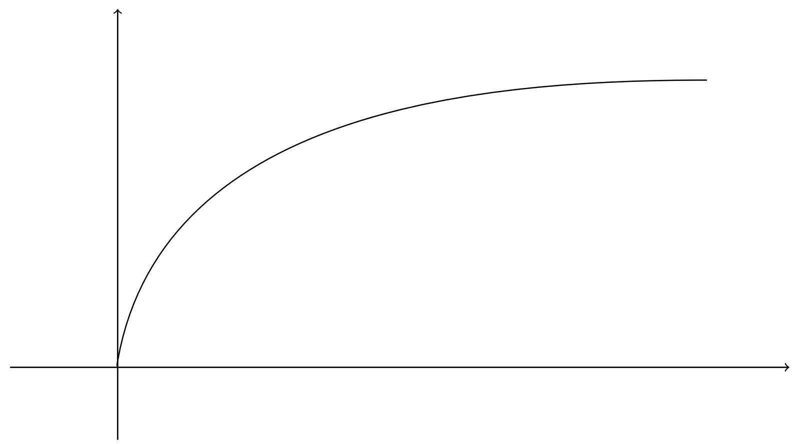

Figure 1: errorOfOrderOneSquareRoot

Consequences As an exercise, we have

$$\begin{cases} \sqrt{n} - CK^2 \le \mathbb{E}[\|X\|_2] \le \sqrt{n} + CK^2 \\ \mathbb{V}[\|X\|_2] \le CK^4 \end{cases}$$

Def. (Covariance matrix)
Let X be random vector in $\mathbb{R}^n$ with $\mathbb{E}[X] = \mu$, then the ***covariance matrix of X*** is

$$\mathrm{Cov}(X) = \mathbb{E}[(X-\mu)(X-\mu)^\top] = \mathbb{E}[XX^\top] - \mu\mu^\top,$$

where $\mathrm{Cov}(X)_{ij} = \mathrm{Cov}(X_i, X_j)$.

Def. (2nd-moment)
The ***second-moment matrix of X*** is

$$\Sigma(X) = \mathbb{E}[XX^\top],$$

where $\Sigma_{ij} = \mathbb{E}[X_i X_j]$.

Remark If $\mathbb{E}[X] = 0$, then $\mathrm{Cov}(X) = \Sigma(X)$. For all X random vectors, $\mathrm{Cov}(X)$ and $\Sigma(X)$ are symmetric positive semidefinite matrices.

Lecture 4: Concentration of measure

2021-11-22

Def. (Isotropy)
A random vector $X \in \mathbb{R}^n$ is called ***isotropic*** if

$$\Sigma(X) = \mathbb{E}[XX^\top] = \mathbb{1}_n.$$

Reduction to isotropy

a) Let Z be an isotropic mean-zero r.v. in $\mathbb{R}^n$, fix $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathcal{M}_{n\times n}(\mathbb{R})$, $\Sigma \geq 0$ then

$$X := \mu + \Sigma^{1/2} Z$$

has mean μ and $\operatorname{Cov}(X) = \Sigma$.

b) If X is a r.v. then $Z := \Sigma^{-1/2}(x - \mu)$ is an isotropic mean-zero r.v.

Lemma 4 (Characterization of isotropy)
A random vector $X \in \mathbb{R}^n$ is isotropic if and only if

$$\mathbb{E}[\langle X, x\rangle^2] = \|x\|_2^2, \quad \forall x \in \mathbb{R}^n, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ is the scalar product in $\mathbb{R}^d$.

Proof.
LHS of (1) is

$$\mathbb{E}\Big[\Big(\sum_i X_i x_i\Big)\Big(\sum_j X_j x_j\Big)\Big] = \sum_i \sum_j x_i x_j \mathbb{E}[X_i X_j].$$

Since $\sum_i x_i^2 = \|x\|2^2$, we have (1) $\iff \mathbb{E}[X_i X_j] = \delta_{ij}$, therefore $\iff X$ is isotropic.

□

Lemma 5 (Norm of isotropic r.v.'s)
Let X be an isotropic r.v. in $\mathbb{R}^n$, then $\mathbb{E}[\|X\|_2^2] = n$. Moreover, if X and Y are independent isotropic r.v.'s in $\mathbb{R}^n$, then $\mathbb{E}[\langle X, Y\rangle^2] = n$.

Proof.
For the first equality, we have

$$\begin{aligned}
\mathbb{E}[\|X\|_2^2] &= \mathbb{E}[\overbrace{X^\top X}^{1\times 1}] \\
&= \mathbb{E}[\operatorname{tr} XX^\top] && \text{(cyclic)} \\
&= \operatorname{tr} \mathbb{E}[XX^\top] && \text{(linearity)} \\
&= \operatorname{tr} I_n && \text{(isotropy)} \\
&= n.
\end{aligned}$$

□

Order of magnitude if we define $\overline{X} = \frac{X}{\|X\|_2}$ and $Y = \frac{Y}{\|Y\|_2}$ with $X \perp\!\!\!\perp Y$ isotropic, then we have that

$$\begin{cases} \|X\|_2 \sim \sqrt{n} \\ \|Y\|_2 \sim \sqrt{n} \\ |\langle X, Y\rangle| \sim \sqrt{n} \end{cases}$$

and therefore

$$\left|\langle \overline{X}, \overline{Y}\rangle\right| = \frac{|\langle X, Y\rangle|}{\|X\|\|Y\|} \sim \frac{\sqrt{n}}{\sqrt{n}\sqrt{n}} \sim \frac{1}{\sqrt{n}}.$$

Example (Standard multivariate Gaussian)
Let $X = (X_1, X_2, \ldots, X_n)$ with $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, then $X \sim \mathcal{N}(0, I_n)$ and $I_n = \mathrm{Cov}(X)$. Hence, X is an isotropic random vector. Recall theorem 13, then the norm of X has concentration bound

$$\mathbb{P}\left(\left|\|X\|_2 - \sqrt{n}\right| \geq t\right) \leq 2e^{-\frac{ct^2}{k^4}}.$$

We can apply the concentration of the norm to the standard Gaussian vector $X \sim \mathcal{N}(0, I_n)$ using another universal constant to include $\|Z\|_{\psi_2}$ since they are i.i.d marginals,

$$X \sim \mathcal{N}(0, I_n) \implies \mathbb{P}\left(\left|\|X\|_2 - \sqrt{n} \geq t\right|\right) \leq 2e^{-Ct^2}.$$

Link between Gaussian distribution and Hausdorff measure on S^{n-1}.

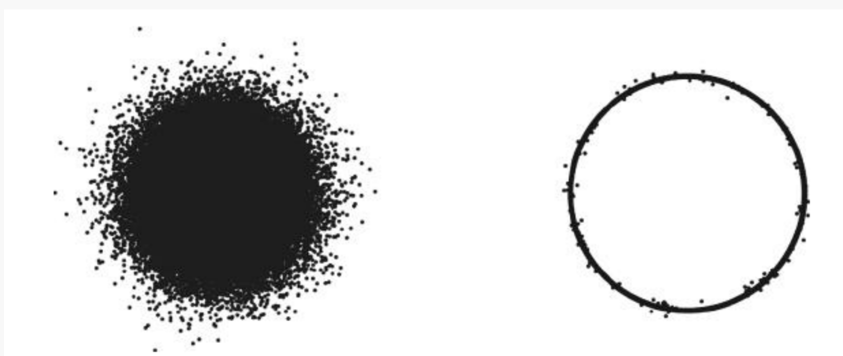

Figure 2: Gaussian point cloud in two dimensions and its visualization in high dimensions. The standard normal distribution is very close to a Unif$(\sqrt{n}S^{n-1})$ distribution on the sphere of radius $\sqrt{n}$.

Theorem 14 (Cramér-Wald)
If X, Y are random vectors in $\mathbb{R}^n$ and $\langle X, \vartheta\rangle \overset{d}{=} \langle Y, \vartheta\rangle$ for all $\vartheta \in \mathbb{R}^n$, then $X \overset{d}{=} Y$

Proof.
No.

□

Def. (Subgaussian random vector)
A random vector $X \in \mathbb{R}^n$ is called ***subgaussian*** if the one-dimensional marginals $\langle X, \vartheta \rangle$ are subgaussian random variables for all $\vartheta \in \mathbb{R}^n$.

Def. (Subgaussian norm of random vectors)
The ***subgaussian norm*** of a subgaussian random vector X is defined as

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

Prop. 8 (Subgaussian marginals)
$X \in \mathbb{R}^n$ is a subgaussian random vector if and only if $X_1, \dots, X_n$ are subgaussian random variables.

Lemma 6 (Bound on the subgaussian norm)
Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent, mean-zero and subgaussian coordinates. Then, X is a subgaussian random vector and

$$\|X\|_{\psi_2} \leq C \max_{1 \leq i \leq n} \|X_i\|_{\psi_2}.$$

Prop. 9 (Sum of subgaussian vectors)
Let $X_1, \dots, X_n$ be independent mean-zero subgaussian random vectors. Then $Z = \sum_{i=1}^n X_i$ is a subgaussian random vector and

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

Example (Examples of subgaussian random vectors)

Theorem 15 (Uniform distribution on a sphere)
Let $X \sim Unif(\sqrt{n} S^{n-1})$, then X is subgaussian and

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

Proof.
TODO

□