Random Matrix Theory Notes*

November 29, 2022

Contents

*Lecture held by Prof. Felix Krahmer in WiSe2223

1 Introduction

1. Lecture
17.10.2022

1.1 Why singular values?

- Linear equations $y = Ax$ are *simplest possible approximation* for any continuous model.
- Taylor's theorem: for small variable range, we can obtain good approximation results under some regularity conditions.
- An important aspect here is *stability*:

  - How do small perturbations in x change y?

  - Reversely, how do small perturbations in y change the reconstruction quality for x?
- Suitable measure for the "quality" of A:

$$\textit{the condition number} \quad \frac{s_{\max}}{s_{\min}},$$

  where $s_{\max}$ and $s_{\min}$ are the maximal and minimal singular values of A.
- When we have the freedom to *design* A (possible via model parameters), we want the condition number to be small. Ideal situation: *approximate isometry*, i.e. all singular values ≈ 1 (after scaling or normalization).

1.2 Why random matrices?

(a) *Compressed sensing.*
  - For a *square matrix*, identity is perfectly well-conditioned.
  - For a *flat rectangular matrix* $A \in \mathbb{R}^{m \times N}$, we have a nontrivial kernel. This can be the worst conditioning (no unique solution).

  In signal processing, signals of interest are often modeled as being *approximately sparse*, i.e., most entries are very small, but we don't know where the large entries are.
  Question: Can we choose A such that all k-columns submatrices ($k < m$) are approximate isometries? How many rows do we need? The simplest solution would be N rows (identity), but that's not desirable because we want to minimize the amount of data we need to access. To reduce it, there are some *deterministic* algorithms which roughly require k^2 rows. However, random construction would only require $k \log N$ rows.

(b) *Dimension reduction.* Assuming that we have p points $\{x_1, ..., x_p\} \in \mathbb{R}^N$, can we project $\mathbb{R}^N \to \mathbb{R}^n$ using a matrix A such that the geometry is approximately preserved

$$(1 - \varepsilon)\|x_i - x_j\| \leqslant \|Ax_i - Ax_j\| \leqslant (1 + \varepsilon)\|x_i - x_j\|.$$

  - Deterministic method must adapt to the points $\{x_1, ..., x_p\}$. No single matrix will work for all sets. (Isometry $\lightning$ Dimension reduction)
  - Random methods work for every set with high probability, and no adaptation is necessary (just set of failures is different). This is of huge advantage for high-dimensional data processing.

1.3 Asymptotic and non-asymptotic regimes

Random matrix theory studies properties of $N \times n$ matrices A chosen from some distribution on the set of all matrices.

Observation: $N, n \to \infty \implies$ spectrum of A stabilizes

Mathematical formulation: *Limit laws* (Random matrix version of central limit theorem)

Example (Bai-Yin law)**.** Let $A \in \mathbb{R}^{n\times n}$ with i.i.d. standard normal random entries. Then

$$\frac{s_{\max}(A)}{2\sqrt{n}} \xrightarrow{n\to\infty} 1 \quad \text{a.s.}$$

This is not enough for finite dimensions, because we have no information about the rate. We need a non-asymptotic version: In every dimension, one has

$$s_{\max}(A) \leqslant C\sqrt{n} \quad \text{w.p. at least } 1 - C' \exp(-n)$$

for absolute constants C, C'. Although this version is less precise (due to the absolute constant C), it is more quantitative, i.e. we have exponentially small probability of failure for fixed dimension.

1.4 Guiding paradigm

$$\boxed{\text{Tall random matrices should act as approximate isometries.}}$$

More precisely, an $N \times n$ random matrix A with $N \gg n$ should satisfy

$$(1-\delta)k\|x\|_2 \leqslant \|Ax\|_2 \leqslant (1+\delta)k\|x\|_2 \quad \text{with high probability}$$

with k a normalization factor and $\delta \ll 1$. Equivalently,

$$(1-\delta)k \leqslant s_{\min}(A) \leqslant s_{\max}(A) \leqslant (1+\delta)K$$

yet equivalently

$$\text{condition number} \quad \frac{s_{\max}(A)}{s_{\min}(A)} \leqslant \frac{1+\delta}{1-\delta} \approx 1.$$

1.5 In this class

We study (tall) random matrices with *independent rows* or *independent columns* and either *strong moment assumptions* (**Subgaussian**) or *no moment assumption* except finite variance (**heavy-tailed**).

Applications:

- (Compressed sensing)
- Dimension reduction
- Estimation of covariance matrices

2 Preliminaries

2.1 Matrices and their singular values

We mostly study tall $A \in \mathbb{R}^{N\times n}$ or $\mathbb{C}^{N\times n}$ with $N \geqslant 1n > 1$ (for flat matrices, consider adjoint).

Definition 2.1 (Singular values)**.** The numbers

$$s_1(A) \geqslant s_2(A) \geqslant \cdots \geqslant s_n(A) \geqslant 0$$

such that $s_i^2(A)$ are the eigenvalues of A^*A are called the *singular values* of A. We also write for the extreme singular values

$$s_{\max}(A) := s_1(A), \quad s_{\min}(A) := s_n(A).$$

Observations:

- $s_{\max}(A)$ and $s_{\min}(A)$ are the smallest $M \in \mathbb{R}$ and the largest $m \in \mathbb{R}$ s.t.

$$m\|x\|_2 \leqslant \|Ax\|_2 \leqslant M\|x\|_2 \quad \forall x \in \mathbb{R}^n.$$

- For flat matrices A, $s_{\min}(A)$ is 0.
- Geometric interpretation: Extreme singular values control the distortion of the Euclidian geometry under the action of A.

Definition 2.2 (Spectral norm)**.** We define the *spectral norm* (or *operator norm*)

$$\|A\| = \|A\|_{l_2^n \to l_2^N} = \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Ax\|_2}{\|x\|_2} = \sup_{x\in\mathbb{S}^{n-1}} \|Ax\|_2.$$

Then it holds that $s_{\max}(A) = \|A\|$ and $s_{\min} = \frac{1}{\|A^\dagger\|}$, where $A^\dagger$ is the pseudo-inverse of A. Further note that

$$\|A\| = \underbrace{\|s\|_\infty}_{\max_i |s_i| = |s_1|} \qquad \text{where } s = (s_1, ..., s_n).$$

Similarly, we define

Definition 2.3 (Schatten norm)**.** Let $A \in \mathbb{R}^{N\times n}$ or $\mathbb{C}^{N\times n}$ with singular values $(s_1, ..., s_n) =: S$. Let $1 \leqslant p \leqslant \infty$. Then the Schatten-$p$-norm is defined as

$$\|A\|_{S^p} := \|s\|_p.$$

The Schatten-2-norm is also called the Frobenius norm and denoted by $\|\cdot\|_F$.

2.2 Nets

Recall that $s_{\max}(A) = \|A\| = \sup_{x\in\mathbb{S}^{n-1}} \|Ax\|_2$ is a supremum over infinitely many x. To analyze the distribution of $s_{\max}$ for a random matrix A, we need to discretize this expression.

Definition 2.4 (ε-net, covering number)**.** Let (X, d) be a metric space and let $\varepsilon > 0$. A subset N_ε is called an *ε-net* of X if every point $x \in X$ can be approximated to an accuracy of ε by some point $y \in N_\varepsilon$, i.e., s.t. $d(x, y) \leqslant \varepsilon$.
The minimal cardinality of an ε-net of X, if finite, is denoted $N(X, \varepsilon)$ and is called the *covering number* of X at scale ε.

Note that $N(X, \varepsilon)$ is finite if and only if X is compact. 2. Lecture 24.10.2022

Lemma 2.5 (Covering number of the sphere)**.** Consider the vector space $\mathbb{R}^n$ equipped with the norm $|||\cdot|||$ and $S = \{x \in \mathbb{R}^n : |||x||| = 1\}$ to be the associated unit sphere. Then for every $\varepsilon > 0$, one has

$$N(s, \varepsilon) \leqslant \left(1 + \frac{2}{\varepsilon}\right)^n$$

Proof. We use a volume argument. Fix $\varepsilon > 0$ and choose N_ε to be a maximal ε-separated subset of S, i.e., N_ε is such that $|||x - y||| \geqslant \varepsilon$ for all $x, y \in N_\varepsilon$, $x \neq y$, and no superset of S containing N_ε has this property. This can be constructed by iteratively adding points. At the end, no additional points can be added. As a result, every point in S has distance $< \varepsilon$ to the nearest point in N_ε. Therefore, N_ε is an ε-net.
Claim: Balls of radii $\frac{\varepsilon}{2}$ centered at the points in N_ε are disjoint. Indeed, if two balls overlap, then the distance of the centers is $< \varepsilon$ ↯.
All such balls lie in $(1 + \frac{\varepsilon}{2})B$, where

$$B = \{x \in \mathbb{R}^n : |||x||| < 1\}$$

We now compare the volumens: because $|N_\varepsilon|$ balls of radius $\frac{\varepsilon}{2}$ are contained in one ball of radius $1 + \frac{\varepsilon}{2}$, we have

$$|N_\varepsilon| \operatorname{vol}\left(\frac{\varepsilon}{2} B\right) \leqslant \operatorname{vol}\left(\left(1 + \frac{\varepsilon}{2}\right) B\right).$$

Using $\operatorname{vol}(rB) = r^n \operatorname{vol}(B)$, we get

$$|N_\varepsilon| \cdot \left(\frac{\varepsilon}{2}\right)^n \operatorname{vol}(B) \leqslant \left(1 + \frac{\varepsilon}{2}\right)^n \operatorname{vol}(B),$$

and therefore

$$|N_\varepsilon| \leqslant \left(1 + \frac{2}{\varepsilon}\right)^n.$$

■

Nets can help to estimate spectral norms. Remember that spectral norms are defined as a supremum over an infinite set of points. How can we be sure we don't miss the maximizer?
Idea: Estimate action of A on all points in N_ε and generalize to all points on the sphere via a perturbation argument.

Lemma 2.6. Let A be a $N \times n$ matrix and let N_ε be an ε-net of S^{n-1} w.r.t. the l_2-norm for some $\varepsilon \in [0,1)$. Then

$$\max_{x \in N_\varepsilon} \|Ax\|_2 \leqslant \|A\| \leqslant (1-\varepsilon)^{-1} \max_{x \in N_\varepsilon} \|Ax\|_2.$$

Proof. First note that lower bound follows from the definition.
Upper bound: By compactness, there exists $x_0 \in S^{n-1}$ s.t. $\|Ax_0\|_2 = \|A\|$. Choose $y \in N_\varepsilon$ which approximates x_0 as $\|x_0 - y\|_2 \leqslant \varepsilon$. By the triangle inequality, we have

$$\begin{aligned} \|A\| = \|Ax_0\|_2 &\leqslant \|Ay\|_2 + \|A(x_0 - y)\| \\ &\leqslant \max_{x \in N_\varepsilon} \|Ax\|_2 + \|A\| \cdot \underbrace{\|x_0 - y\|_2}_{\leqslant \varepsilon} \end{aligned}$$

Consequently, we have

$$\|A\|(1-\varepsilon) \leqslant \max_{x \in N_\varepsilon} \|Ax\|$$

and thereby the claim. ■

The same trick works for symmetric matrices and the associated quadratic form. First note that for a symmetric matrix $A \in \mathrm{Sym}(n)$, there exists $Q \in O(n)$ with $A = Q\Sigma Q^T$. Therefore,

$$\sup_{\|x\|_2=1} \langle Ax, x\rangle \overset{Q^T \in O(n)}{=} \sup_{\|x\|_2=1} \langle Q\Sigma Q^T x, Q^T x\rangle = \sup_{\|Q^T x\|_2=1} \langle \Sigma x, x\rangle = \max_i |\Sigma_{ii}| = \|A\|.$$

Lemma 2.7. Let A be a symmetric $n \times n$ matrix and let N_ε be an ε-net of S^{n-1} w.r.t. $\|\cdot\|_{l_2}$ for some $\varepsilon \in [0,1)$. Then

$$\|A\| = \sup_{x \in S^{n-1}} |\langle Ax, x\rangle| \leqslant (1-2\varepsilon)^{-1} \max_{x \in N_\varepsilon} |\langle Ax, x\rangle|$$

Proof. Choose $x_0 \in S^{n-1}$ s.t. $\|A\| = \langle Ax_0, x_0\rangle$ and choose $y \in N_\varepsilon$ which approximates x_0 as $\|x_0 - y\|_2 \leqslant \varepsilon$. By the triangle inequality, we have

$$\begin{aligned} |\langle Ax_0, x_0\rangle - \langle Ay, y\rangle| &= |\langle Ax_0, x_0 - y\rangle + \langle A(x_0 - y), y\rangle| \\ &\leqslant (\|A\|\|x_0\|_2)\|x_0 - y\|_2 + (\|A\|\|x_0 - y\|_2)\|y\|_2 \\ &\leqslant 2\varepsilon\|A\|. \end{aligned}$$

Therefore,

$$|\langle Ay, y\rangle| \geqslant |\langle Ax_0, x_0\rangle| - 2\varepsilon\|A\| = (1-2\varepsilon)\|A\|.$$

The claim follows by taking maximum over y. ■

2.3 Non-asymptotic results in one dimension

Last time, we talked about asymptotic vs. non-asymptotic theory. Most results in probability theory are asymptotic. Here we introduce some non-asymptotic variants.

Proposition 2.8 (Hoeffding's inequality)**.** Let $X_1, ..., X_n$ be a sequence of independent, real-valued random variables s.t. $\mathbb{E}(X_l) = 0$ and $|X_l| \leqslant B_l$ a.s. for all $l = 1, ..., n$ for some $B_l > 0$. Then

$$P\left(\sum_{l=1}^{n} X_l > t\right) \leqslant \exp\left(-\frac{t^2}{2\sum_{l=1}^{n} B_l^2}\right) \quad \forall t > 0,$$

and consequently

$$P\left(\left|\sum_{l=1}^{n} X_l\right| > t\right) \leqslant 2\exp\left(-\frac{t^2}{2\sum_{l=1}^{n} B_l^2}\right) \quad \forall t > 0.$$

Proposition 2.9 (Bernstein type inequality)**.** Let $X_1, ..., X_n$ be independent mean-zero random variables such that for all $l \in [n]$

$$\mathbb{E}[|x_l|^m] \leqslant m! K^{m-2}\frac{\sigma_l}{2} \quad \forall m \in \mathbb{N}, m \geqslant 2$$

for some $K > 0$ and $\sigma_l > 0$. Then

$$P\left(\left|\sum_{l=1}^{n} X_l\right| > t\right) \leqslant 2\exp\left(-\frac{t^2}{2(\sigma^2 + Kt)}\right) \quad \forall t > 0$$

where $\sigma^2 := \sum_{l=1}^{n} \sigma_l^2$.

Both results are about sums of independent random variables. Now, we demonstrate one example for sum of *dependent* random variables. One of the simplest dependence could be the product of two independent random variables.

Chaos: Consider $X = \sum a_{ij} X_i X_j$ where a_{ij} are deterministic coefficients and X_i are i.i.d. random variables. There will be necessarily dependencies as we have more than $\binom{n}{2}$ choices but we only have n variables. Even more specific: consider **Rademacher** random variables ε_i with $P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = \frac{1}{2}$ and **Rademacher chaos** $\sum_{i \neq j} a_{ij}\varepsilon_i\varepsilon_j$. Note that the diagonal terms are deterministic.

However, one problem we face is that ε_i and ε_j are the "same" variables, and can not be treated separately. We would like to decouple them and compare to the case where ε_i and ε_j' are independent sequences.

Lemma 2.10 (Decoupling)**.** Let $\xi = (\xi_1, ..., \xi_M)$ be a sequence of independent random variables with $\mathbb{E}[\xi_j] = 0 \ \forall j \in [M]$. Let $A_{jk},\ j,k \in [M]$ be a doubly indexed sequence of elements in a vector space X. Let $F : X \to \mathbb{R}$ be a convex function. Then

$$\mathbb{E}\left[F\left(\sum_{j,k=1, j\neq k}^{M} \xi_j\xi_k A_{jk}\right)\right] \leqslant \mathbb{E}\left[F\left(4\sum_{j,k=1}^{M} \xi_j\xi_k' A_{jk}\right)\right],$$

where ξ_j' is an independent copy of ξ_j.

Motivation: Condition on ξ', use concentration inequality for ξ. 3. Lecture 31.10.2022

Proof. We introduce a sequence $\delta = (\delta_j)_{j=1}^M$ of independent random variables δ_j via $P(\delta_j = 0) = P(\delta_j = 1) = \frac{1}{2}$. Then for $j \neq k$

$$\mathbb{E}[\delta_k(1-\delta_j)] = \frac{1}{4}.$$

This yields

$$\begin{aligned} E &:= \mathbb{E}\left[F\left(\sum_{j\neq k}^{M} \xi_j \xi_k A_{jk}\right)\right] \\ &= \mathbb{E}\left[F\left(4\sum_{j\neq k}^{M} \mathbb{E}_\delta[\delta_j(1-\delta_k)]\xi_j\xi_k A_{jk}\right)\right] \\ &\overset{\text{Jenson}}{\leqslant} \mathbb{E}_\xi\left[\mathbb{E}_\delta\left[F\left(4\sum_{j\neq k}^{M} \delta_j(1-\delta_k)\xi_j\xi_k A_{jk}\right)\right]\right]. \end{aligned}$$

Now let

$$\sigma(\delta) := \{j = 1, ...M : \delta_j = 1\}.$$

Then by Fubini, we have

$$E \leqslant \mathbb{E}_\delta\left[\mathbb{E}_\xi\left[F\left(4\sum_{j\in\sigma(\delta)}\sum_{k\notin\sigma(\delta)} \xi_j\xi_k A_{jk}\right)\right]\right].$$

Now, conditionally on δ, a random variable ξ_i appears <u>either</u> only as the first factor (if $i \in \sigma(s)$) <u>or</u> only as the second factor (if $i \notin \sigma(s)$). So if we replace all ξ_k, $k \notin \sigma(\delta)$ by independently identically distributed ξ_k', the first factor is never changed, the second factor is always changed and the value of the expectation is not changed. Hence we can write

$$E \leqslant \mathbb{E}_\delta\left[\mathbb{E}_\xi\left[\mathbb{E}_{\xi'}\left[F\left(4\sum_{j\in\sigma(\delta)}\sum_{k\notin\sigma(\delta)} \xi_j\xi_k' A_{jk}\right)\right]\right]\right].$$

Hence, there exists δ_0 and $\sigma = \delta(\delta_0)$ s.t.

$$E \leqslant \underbrace{\mathbb{E}_\xi\left[\mathbb{E}_{\xi'}\left[F\left(4\sum_{j\in\sigma(\delta_0)}\sum_{k\notin\sigma(\delta_0)} \xi_j\xi_k' A_{jk}\right)\right]\right]}_{(*)}.$$

(Otherwise, E can not be smaller than the expectation of δ.) Our <u>goal</u> now is to introduce missing terms. We want to use the fact $\mathbb{E}\xi_i = 0$ and pull our expectation using Jensen.

$$
\begin{aligned}
(*) &= \mathbb{E}_\xi\left[\mathbb{E}_{\xi'}\left[F\left(4\sum_{j\in\sigma}\left(\sum_{k\notin\sigma}\xi_j\xi_k' A_{jk} + \sum_{k\in\sigma}\xi_j \underbrace{\mathbb{E}\xi_k'}_{=0} A_{jk}\right)\right)\right]\right] \\
&\overset{\text{Jenson}}{\leqslant} \mathbb{E}_\xi\left[\mathbb{E}_{\xi'}\left[F\left(4\sum_{j\in\sigma}\sum_{k=1}^{M}\xi_j\xi_k' A_{jk}\right)\right]\right] \\
&\overset{\text{Fubini}}{=} \mathbb{E}_{\xi'}\left[\mathbb{E}_\xi\left[F\left(4\sum_{k=1}^{M}\left(\sum_{j\in\sigma}\xi_j\xi_k' A_{jk} + \sum_{j\notin\sigma}\underbrace{(\mathbb{E}\xi_j)}_{=0}\xi_k' A_{jk}\right)\right)\right]\right] \\
&\overset{\text{Jenson}}{\leqslant} \mathbb{E}\left[F\left(4\sum_{j=1}^{M}\sum_{k=1}^{M}\xi_j\xi_k' A_{jk}\right)\right]
\end{aligned}
$$

and thereby the claim. ■

Theorem 2.11 (Tail estimates for Rademacher chaos)**.** Let $A \in \mathbb{R}^{M\times M}$ be a symmetric matrix with zero diagonal and ε a Rademacher factor. Consider the Rademacher chaos

$$
X = \sum_{j,k=1}^{M} \varepsilon_j\varepsilon_k A_{jk}.
$$

Then

$$
P(|X| \geqslant t) \leqslant \begin{cases} 2\exp\left(-\frac{3t^2}{128\|A\|_F^2}\right) & \text{if } \ 0 < t \leqslant \frac{4\|A\|_F^2}{3\|A\|} \\ 2\exp\left(-\frac{t}{32\|A\|}\right) & \text{if } \ t > \frac{4\|A\|_F^2}{3\|A\|} \end{cases}.
$$

Proof. Consider moment generating function

$$
\begin{aligned}
\mathbb{E}[\exp(\theta x)] &= \mathbb{E}\left[\theta\sum_{j\neq k}\varepsilon_j\varepsilon_k A_{jk}\right] \\
&\overset{\text{Lem. 2.10}}{\leqslant} \mathbb{E}\left[\exp\left(4\theta\sum_{j\neq k}\varepsilon_j\varepsilon_k' A_{jk}\right)\right] \\
&= \mathbb{E}_\varepsilon\Bigg[\underbrace{\mathbb{E}_{\varepsilon'}\left[\exp\left(4\theta\sum_{j\neq k}\varepsilon_j\varepsilon_k' A_{jk}\right)\right]}_{=\prod_{k=1}^{M}\mathbb{E}\exp\left(\varepsilon_k\cdot 4\cdot\theta\sum_{j\neq k}\varepsilon_j A_{jk}\right)}\Bigg] \\
&\overset{(*)}{\leqslant} E_\varepsilon\left[\prod_k \exp\left(8\theta^2\left|\sum_j \varepsilon_j A_{jk}\right|^2\right)\right] \\
&= E_\varepsilon\left[\exp\left(8\theta^2\sum_k\left|\sum_j\varepsilon_j A_{jk}\right|^2\right)\right]
\end{aligned}
$$

where $(*)$ results from

$$
\mathbb{E}(\exp(\theta y)) \leqslant \exp\left(\frac{\theta^2 y_{\max}^2}{2}\right)
$$

when $y \leqslant y_{\max}$ a.s. and $E[y] = 0$ with $y_{\max} = 4\theta \left|\sum_{j\neq k} \varepsilon_j A_{jk}\right|$.
By symmetry of A,

$$\sum_k \left(\sum_j \varepsilon_j A_{jk}\right)^2 = \sum_k \sum_j \varepsilon_j A_{jk} \sum_l \varepsilon_l A_{lk} = \varepsilon^* A^2 \varepsilon.$$

The matrix $B := A^2 = A^* A$ is symmetric and positive semidefinite.
Goal: Estimate moment generating function for positive semi-definite chaos: for $\kappa > 0$

$$\begin{aligned}\mathbb{E}[\exp(\kappa\varepsilon^* B\varepsilon)] &= \mathbb{E}\exp\left(\kappa \sum_j B_{jj} + \kappa \sum_{j\neq k} B_{jk}\varepsilon_j\varepsilon_k\right)\\ &\overset{\text{decoupling}}{\leqslant} \exp(\kappa \operatorname{tr}(B))\mathbb{E}\left[\exp\left(4\kappa \sum_{j\neq k} B_{jk}\varepsilon_j\varepsilon'_k\right)\right]\\ &\leqslant \exp(\kappa \operatorname{tr}(B))\mathbb{E}\exp\left(8\kappa^2 \sum_k \left(\sum_j \varepsilon_j B_{jk}\right)^2\right).\end{aligned}$$

Now use positive semidefiniteness

$$\sum_k \left(\sum_j \varepsilon_j B_{jk}\right)^2 = \varepsilon^* B^2 \varepsilon = \varepsilon^* P D^2 P^* \varepsilon = \sum_i \lambda_i(B)((P^*\varepsilon)_i)^2 \overset{\lambda_i \geqslant 0}{\leqslant} \lambda_{\max}\varepsilon^* PDP\varepsilon = \|B\|\varepsilon^* B\varepsilon.$$

Hence

$$\begin{aligned}\mathbb{E}[\exp(\kappa\varepsilon^* B\varepsilon)] &\leqslant \exp(\kappa \operatorname{tr}(B))\mathbb{E}[\exp\left(8\kappa^2\|B\|\varepsilon^* B\varepsilon\right)]\\ &= \exp(\kappa \operatorname{tr}(B))\mathbb{E}\left[(\exp(\varepsilon^* B\varepsilon))^{8\kappa^2\|B\|}\right]\\ &\underset{\text{if } 8\kappa^2\|B\|<1}{\overset{\text{Jenson}}{\leqslant}} \exp(\kappa \operatorname{tr}(B))(\mathbb{E}[\exp(\kappa\varepsilon^* B\varepsilon)])^{8k\|B\|}.\end{aligned}$$

Consequently

$$\mathbb{E}\exp(\kappa\varepsilon^* B\varepsilon) \leqslant \exp\left(\frac{\kappa \operatorname{tr}(B)}{1 - 8\kappa\|B\|}\right)$$

provided $0 < \kappa < \frac{1}{8\|B\|}$. Specify $\theta := \sqrt{\frac{\kappa}{8}}$, i.e. $\kappa = 8\theta^2$. Then 4. Lecture 7.11.2022

$$\begin{aligned}\mathbb{E}[\exp(\theta X)] &\leqslant \exp\left(\frac{8\theta^2 \operatorname{tr}(A^2)}{1 - 64\theta^2\|A^2\|}\right), \quad 0 < \theta < \frac{1}{8\sqrt{\|A^2\|}}\\ &\underset{\|A^2\|=\|A\|^2}{\overset{\operatorname{tr}(A^2)=\|A\|_F^2}{=}} \exp\left(\frac{8\theta^2\|A\|_F^2}{1 - 64\theta^2\|A\|^2}\right), \quad 0 < \theta < \frac{1}{8\|A\|}\end{aligned}$$

by using the fact

$$\|A^2\| = \sup_{x\in S^{n-1}} \langle x, A^2 x\rangle \overset{A \text{ sym.}}{=} \sup_{x\in S^{n-1}} \langle Ax, Ax\rangle = \sup_{x\in S^{n-1}} \|Ax\|^2 = \|A\|^2$$

and

$$\operatorname{tr}(A^2) = \sum_{i,j} A_{ij}A_{ji} = \sum_{ij} A_{ij}^2 = \|A\|_F.$$

Now assume $0 < \theta < \frac{1}{16\|A\|}$. Then the denominator $\geqslant 1 - \frac{1}{4} = \frac{3}{4}$. Thus,

$$\begin{aligned}
P(X \geqslant t) &= P(\exp(\theta X) \geqslant \exp(\theta t)) \overset{\text{Mkv}}{\leqslant} \exp(-\theta t)\mathbb{E}[\exp(\theta X)] \\
&= \exp\left(-\theta t + \frac{8\theta^2\|A\|_F^2}{1 - 64\theta^2\|A\|^2}\right) \\
&\leqslant \exp\left(-\theta t + \frac{32}{3}\theta^2\|A\|_F^2\right)
\end{aligned}$$

We now calculator the optimal choice of θ:

$$\begin{aligned}
\frac{d}{d\theta}\left(-\theta t + \frac{32}{3}\theta^2\|A\|_F^2\right) &\overset{!}{=} 0 \\
\theta_{\text{opt}} &= \frac{3t}{64\|A\|_F^2}
\end{aligned}$$

Then

$$P(X \geqslant t) \leqslant \exp\left(-\frac{3t^2}{128\|A\|_F^2}\right).$$

Recall that we need $0 < \theta \leqslant \frac{1}{16\|A\|}$, i.e. estimate only works for $t \leqslant \frac{4\|A\|_F^2}{3\|A\|}$. For $t > \frac{4\|A\|_F^2}{3\|A\|}$, set $\theta = \frac{1}{16\|A\|}$ (as large as possible). Then

$$\begin{aligned}
P(X \geqslant t) &\leqslant \exp\left(-\theta t + \frac{32\theta^2\|A\|_F^2}{3}\right) \\
&\underset{\text{otherwise case 1}}{\overset{\theta \leqslant \theta_{opt}}{\leqslant}} \exp\left(-\theta t + \frac{\theta t}{2}\right) = \exp\left(-\frac{t}{32\|A\|}\right)
\end{aligned}$$

Observation: $-X = \sum \varepsilon_i \varepsilon_j A_{ij}$, so by the same proof, we get

$$P(-X \geqslant t) \leqslant \begin{cases} \exp\left(-\frac{3t^2}{128\|-A\|_F^2}\right), & t \leqslant \frac{4\|-A\|_F^2}{3\|-A\|} \\ \exp\left(-\frac{t}{32\|-A\|}\right), & t > \frac{4\|-A\|_F^2}{3\|-A\|}. \end{cases}$$

And since $\|A\| = \|-A\|$ and $\|A\|_F = \|-A\|_F$, we get

$$P(|X| \geqslant t) \leqslant P(X \geqslant t) + P(X \leqslant -t) = \begin{cases} 2\exp\left(-\frac{3t^2}{128\|A\|_F^2}\right) & \text{if } 0 < t \leqslant \frac{4\|A\|_F^2}{3\|A\|} \\ 2\exp\left(-\frac{t}{32\|A\|}\right) & \text{if } t > \frac{4\|A\|_F^2}{3\|A\|}. \end{cases}$$

and thereby the claim. ■

2.4 Subgaussian random variables

Motivation: Gaussian distributions are well behaved. What do we mean by that?
Recall: Let X be a standard normal random variable. Then the distribution of X has density

$$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right),$$

denoted by $\mathcal{N}(0,1)$. For a normal distribution with mean μ and variance σ^2, we have density

$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right).$$

Properties of the standard normal distribution:

- Well-behaved tails:
$$P(|X|>t)\leqslant 2\exp\left(-\frac{t^2}{2}\right),\quad t\geqslant 1$$
- Well-behaved moments:
$$(\mathbb{E}|X|^p)^{\frac{1}{p}}=O(\sqrt{p}),\quad p\geqslant 1$$
- Well-behaved moment-generating function:
$$\mathbb{E}[\exp(tX)]=\exp\left(\frac{t^2}{2}\right)$$

These three properties shall serve as our idea of a well-behaved distribution. A random variable (or more approximately, the associated distribution) satisfying them will be called subgaussian. We will see that the three properties are equivalent.
Fact (cf. Stiring's approximation)[1]:

$$p!\geqslant\left(\frac{p}{e}\right)^p$$

Lemma 2.12 (Equivalence of subgaussian properties)**.** Let X be a random variable. Then the following properties are equivalent with parameters $k_i>0$ differing from each other by at most an absolute constant factor. More precisely, there exists an absolute constant C, s.t. property i implies property j with parameter $k_j\leqslant Ck_i$ for any two properties $i,j=1,2,3$.

(1) Tails:
$$P(|X|>t)\leqslant\exp\left(1-\frac{t^2}{k_1^2}\right)\qquad\forall t\geqslant 0$$

(2) Moments (bounds on L^p norm):
$$(\mathbb{E}|X|^p)^{\frac{1}{p}}\leqslant k_2\sqrt{p}\qquad\forall p\geqslant 1$$

[1] Or proof by induction:

$$p!=p(p-1)!\geqslant p\frac{(p-1)^{p-1}}{e^{p-1}}=p^p\frac{\left(1-\frac{1}{p}\right)^{p-1}}{e^{p-1}}\overset{(*)}{\geqslant}\frac{p^n}{e^p},$$

where $(*)$ follows from

$$\ln\left(1-\frac{1}{p}\right)^{p-1}=(p-1)\ln\left(\frac{p-1}{p}\right)=-(p-1)\ln\left(\frac{p}{p-1}\right)=-(p-1)\ln\left(1+\frac{1}{p-1}\right)\overset{(*')}{\geqslant}-1\geqslant e^{-1},$$

and $(*')$ from the fact $\ln(1+x)<x$.

(3) Super-exponential moments:

$$\mathbb{E}\left[\exp\left(\frac{X^2}{k_3^2}\right)\right] \leqslant e$$

Moreover, if $\mathbb{E}X = 0$, then properties 1-3 are also equivalent to the following one:

(4) Moment generating function:

$$\mathbb{E}[\exp(tX)] \leqslant \exp(t^2 k_4^2) \qquad \forall\, t \in \mathbb{R}$$

Proof. First note that all properties are *homogeneous*, i.e., X satisfies the property with k_i iff $\widetilde{X} = \frac{X}{k_i}$ satisfies the properties with $\widetilde{k_i} = 1$. For example for property

$$P(|X| > t) \leqslant \exp\left(1 - \frac{t^2}{k_1^2}\right) \qquad \forall\, t \geqslant 0$$

is equivalent to

$$P(|\widetilde{X}| > \widetilde{t}) = P(|X| > \widetilde{t} \cdot k_1) \leqslant \exp\left(1 - \widetilde{t}^2\right).$$

So we can also always assume $k_i = 1$ and show $k_j \leqslant C$.

- <u>$1 \implies 2$:</u> Assume property 1. Note that

$$\begin{aligned}
\mathbb{E}|X|^p &= \mathbb{E}\left(\int_0^\infty \mathbb{1}_{\{y<|X|^p\}} dy\right) \\
&\overset{\text{Fubini}}{=} \int_0^\infty \mathbb{E}[\mathbb{1}_{\{\|X\|^p > y\}}] dy \\
&= \int_0^\infty P(|X|^p \geqslant y) dy \\
&\overset{y=t^p}{=} \int_0^\infty P(|X| > t) p t^{p-1} dt \\
&\overset{\text{prop 1}}{\leqslant} \int_0^\infty \exp(1-t^2) p t^{p-1} dt \\
&= \left(\frac{ep}{2}\right) \Gamma\left(\frac{p}{2}\right) \leqslant \left(\frac{ep}{2}\right)\left(\frac{p}{2}\right)^{\frac{p}{2}}
\end{aligned}$$

  So property 1 with $k_1 = 1$ implies property 2 with $k_2 = \sup_p \frac{1}{\sqrt{2}} \left(\frac{ep}{2}\right)^{\frac{1}{p}}$. 5. Lecture 14.11.2022

- <u>$2 \implies 3$:</u> Let $c > 0$ and assume $k_2 = 1$. Writing the Taylor series of the exponential function

$$\begin{aligned}
\mathbb{E}[\exp(cX^2)] = 1 + \sum_{p=1}^\infty \frac{c^p \mathbb{E}[X^{2p}]}{p!} &\overset{\text{prop 2}}{\leqslant} 1 + \sum_{p=1}^\infty c^p \frac{(2p)^{\frac{1}{2}\cdot 2p}}{p!} \\
\overset{p! \geqslant \left(\frac{p}{e}\right)^p}{\leqslant} 1 + \sum_{p=1}^\infty c^p \frac{(2p)^p}{\left(\frac{p}{e}\right)^p} &= 1 + \sum_{p=1}^\infty (2ce)^p \\
\overset{\text{if } 2ce<1}{=} \frac{1}{1-2ce} = \frac{e}{e(1-2ce)} &\overset{\text{if } 1-2ce \geqslant \frac{1}{e}}{\leqslant} e
\end{aligned}$$

This estimation holds if $1 - 2ce \geqslant \frac{1}{e}$, i.e. $c \leqslant \frac{e-1}{2e^2}$. Under this assumption, we have

$$\mathbb{E}\exp\left(\frac{e-1}{2e^2}X^2\right) \leqslant e.$$

Because the choice of c was arbitrary, property 3 holds with $k_3 = \sqrt{\frac{2e^2}{e-1}}$.

- <u>$3 \implies 1$:</u> Note that

$$\begin{aligned} P(|X| > t) &= P(\exp(X^2) \geqslant \exp(t^2)) \\ &\overset{\text{Mkv}}{\leqslant} \exp(-t^2)\mathbb{E}[\exp(X^2)]. \end{aligned}$$

So property 3 with $k_3 = 1$ implies property 1 with k_1 with $k_1 = 1$.

- <u>$2 \implies 4$:</u> Taylor series imply

$$\begin{aligned} \mathbb{E}[\exp(tX)] &= 1 + t\underbrace{\mathbb{E}[X]}_{=0 \text{ by ass.}} + \sum_{p=2}^{\infty} \frac{t^p \mathbb{E}[X^p]}{p!} \\ &\overset{\text{prop } 2}{\leqslant} 1 + \sum_{p=2}^{\infty} \frac{t^p p^{\frac{p}{2}}}{p!} \\ &\overset{p! \geqslant (\frac{p}{e})^p}{\leqslant} 1 + \sum_{p=2}^{\infty} \left(\frac{et}{\sqrt{p}}\right)^p \end{aligned} \tag{2.1}$$

We need to compare (2.1) with

$$\exp(k_4^2 t^2) = 1 + \sum_{k=1}^{\infty} \frac{(k_4|t|)^{2k}}{k!} \overset{p! \leqslant p^p}{\geqslant} \sum_{k=1}^{\infty} \left(\frac{k_4|t|}{\sqrt{k}}\right)^{2k} \tag{2.2}$$

Note that there are no odd terms in (2.2). Thus, we need to control $\left(\frac{et}{\sqrt{p}}\right)^p, p \geqslant 3$ and odd. If $\frac{e|t|}{\sqrt{p}} \leqslant 1$, then

$$\left(\frac{e|t|}{\sqrt{p}}\right)^p \leqslant \left(\frac{e|t|}{\sqrt{p}}\right)^{p-1} \leqslant \left(\frac{e|t|}{\sqrt{p-1}}\right)^{p-1}.$$

If $\frac{e|t|}{\sqrt{p}} > 1$, then

$$\left(\frac{e|t|}{\sqrt{p}}\right)^p \leqslant \left(\frac{e|t|}{\sqrt{p}}\right)^{p+1} \leqslant \left(\frac{e|t|}{\sqrt{\frac{p+1}{2}}}\right)^{p+1} = \left(\frac{\sqrt{2}e|t|}{\sqrt{p+1}}\right)^{p+1}.$$

Therefore,

$$\left(\frac{e|t|}{\sqrt{p}}\right)^p \leqslant \left(\frac{e|t|}{\sqrt{p-1}}\right)^{p-1} + \left(\frac{\sqrt{2}e|t|}{\sqrt{p+1}}\right)^{p+1} \tag{2.3}$$

Hence,

$$
\begin{aligned}
\mathbb{E}[\exp(tX)] &\overset{(2.1)}{\leqslant} 1+\sum_{p=2}^{\infty}\left(\frac{et}{\sqrt{p}}\right)^p \\
&\leqslant 1+\sum_{p\in 2\mathbb{N}}\left(\frac{e|t|}{\sqrt{p}}\right)^p+\left(\frac{e|t|}{\sqrt{p+1}}\right)^{p+1} \\
&\overset{(2.3)}{\leqslant} 1+\sum_{p\in 2\mathbb{N}}\left(\frac{e|t|}{\sqrt{p}}\right)^p+\left(\frac{e|t|}{\sqrt{p}}\right)^p+\left(\frac{\sqrt{2}e|t|}{\sqrt{p+2}}\right)^{p+2} \\
&\leqslant 1+\sum_{p\in 2\mathbb{N}}\underbrace{\left(2+(\sqrt{2})^p\right)}_{\leqslant(2\sqrt{2})^p}\left(\frac{e|t|}{\sqrt{p}}\right)^p \\
&\overset{p=2k}{\leqslant} 1+\sum_{k\in\mathbb{N}}\left(\frac{2\sqrt{2}e|t|}{\sqrt{2k}}\right)^{2k} \\
&\overset{(2.2)}{\leqslant} \exp(k_4^2t^2)
\end{aligned}
$$

provided $k_4 \geqslant 2e$.

- <u>$4 \implies 1$:</u> Note that Markov implies $\forall\,\lambda>0$

  $$P(X\geqslant t)=P(\exp(\lambda X)\geqslant\exp(\lambda t))\leqslant\exp(-\lambda t)\mathbb{E}[\exp(\lambda X)]\overset{\text{prop 4}}{\leqslant}\exp(-\lambda t+\lambda^2)$$

  with $k_4=1$. By choosing $\lambda=\frac{t}{2}$ we get

  $$P(X\geqslant t)\leqslant\exp\left(-\frac{t^2}{4}\right).$$

  and therefore

  $$P(|X|\geqslant t)\leqslant 2\exp\left(-\frac{t^2}{4}\right)\leqslant\exp\left(1-\frac{t^2}{4}\right).$$

This completes the proof. ■

Remark. Note that

(i) Constants 1 and e in property 1 and 3 are chosen for convenience, any number >0 or >1, respectively, will do.

(ii) $4 \implies 1$ does not use $\mathbb{E}X=0$ and thus the condition is only for necessity.

Definition 2.13 (Subgaussian random variables)**.** A random variable that satisfies one of the equivalent properties 1-3 in Lemma 2.12 is called a subgaussian random variable. The **subgaussian norm** of X, denoted as $\|X\|_{\psi_2}$, is defined as the smallest k_2 in property 2. In other words

$$\|X\|_{\psi_2}=\sup_{p\geqslant 1}p^{-\frac{1}{2}}(\mathbb{E}[|X|^p])^{\frac{1}{p}}$$

Remark. The subgaussian norm is indeed a norm:

(i) $\|X\|_{\psi_2} = 0 \iff X = 0$ a.s.
"$\Leftarrow$" direct calculation
"$\Rightarrow$" If not $X = 0$ a.s., then $\exists \varepsilon > 0$ s.t. $P(|X| > \varepsilon) = p > 0$. Thus

$$\|X\|_{\psi_2} \overset{p=1}{\geqslant} \mathbb{E}[|X|] \geqslant \varepsilon P(|X| \geqslant \varepsilon) = p - \varepsilon > 0,$$

which is a contradiction.

(ii) $\|\lambda X\|_{\psi_2} = \sup_{p\geqslant 1} p^{-\frac{1}{2}}(\mathbb{E}|\lambda X|^p)^{\frac{1}{p}} = |\lambda|\|X\|_{\psi_2}$.

(iii) Triangle inequality:

$$\begin{aligned}
\|X+Y\|_{\psi_2} &= \sup_{p\geqslant 1} p^{-\frac{1}{2}}(\mathbb{E}[|X+Y|^p])^{\frac{1}{p}} \\
&\underset{\text{for } L^p}{\overset{\text{Minkovski}}{\leqslant}} \sup_{p\geqslant 1} p^{-\frac{1}{2}}\left((\mathbb{E}[|X|^p])^{\frac{1}{p}} + (\mathbb{E}[|Y|^p])^{\frac{1}{p}}\right) \\
&\leqslant \sup_{p\geqslant 1} p^{-\frac{1}{2}}(\mathbb{E}[|X|^p])^{\frac{1}{p}} + \sup_{p\geqslant 1} p^{-\frac{1}{2}}(\mathbb{E}[|Y|^p])^{\frac{1}{p}} \\
&= \|X\|_{\psi_2} + \|Y\|_{\psi_2}
\end{aligned}$$

Thus, the class of subgaussian random variables on a given probability space is thus a normed space.

We can now reformulate Lemma 2.12 into the language of subgaussian norm. By Lemma 2.12 , there exist universal constants c, C s.t. a subgaussian random variable satisfies

$$P(|X| > t) \leqslant \exp\left(1 - \frac{ct^2}{\|X\|_{\psi_2}^2}\right) \qquad \forall t > 0, \tag{2.4}$$

$$(\mathbb{E}[|X|^p])^{\frac{1}{p}} \leqslant \|X\|_{\psi_2}\sqrt{p} \qquad \forall p \geqslant 1, \tag{2.5}$$

$$\mathbb{E}\left[\frac{cX^2}{\|X\|_{\psi_2}^2}\right] \leqslant e, \tag{2.6}$$

$$\mathbb{E}[\exp(tX)] \leqslant \exp\left(Ct^2\|X\|_{\psi_2}^2\right) \qquad \forall t \in \mathbb{R} \text{ if } \mathbb{E}[X] = 0. \tag{2.7}$$

Moreover, up to absolute constant factors, $\|X\|_{\psi_2}$ is the smallest possible number in the properties of Lemma 2.12.

Example. Examples for subgaussian random variables.

(i) (Gaussian) If X is a centered standard normal random variable with variance σ^2, then X is subgaussian with $\|X\|_{\psi_2} \leqslant C \cdot \sigma$.

(ii) (Bounded RV) Let X be such that $|X| \leqslant M$ a.s. Then X is subgaussian with $\|X\|_{\psi_2} \leqslant M$. Indeed, $(\mathbb{E}[|X|^p])^{\frac{1}{p}} \leqslant M \leqslant \sqrt{p}M$. In particular, a Rademacher random variable $P(\varepsilon = 1) = P(\varepsilon = -1) = \frac{1}{2}$ satisfies

$$\|\varepsilon\|_{\psi_2} = 1.$$

(We have equality for $p = 1$).

Last time we have seen that normal distribution is well-behaved. One more nice property for Gaussian is rotation invariance, which makes it easy to work in high dimensions. Given a finite number of independent centered normal random variables X_i, their sum $\sum_i X_i$ is also a centered random variable with $\mathrm{Var}\left(\sum_i X_i\right) = \sum_i \mathrm{Var}(X_i)$. Idea: multiplicativity of moment generating function same works for subgaussians.

Lemma 2.14 (Rotation invariance)**.** Consider a finite number of independent centered subgaussian random variables X_i. Then $\sum_i X_i$ is also a centered subgaussian random variable. Moreover,

$$\left\|\sum_i X_i\right\|_{\psi_2}^2 \leqslant C \sum_i \|X_i\|_{\psi_2}^2,$$

where C is an absolute constant.

Proof. One of the equivalent properties if $\mathbb{E}X = 0$ is:

$$\mathbb{E}[\exp(tX)] \leqslant \exp\left(Ct^2\|X\|_{\psi_2}^2\right) \qquad \forall t \in \mathbb{R}.$$

So for $t \in \mathbb{R}$

$$\begin{aligned}\mathbb{E}\left[\exp\left(\sum_i X_i\right)\right] = \mathbb{E}\left[\prod_i \exp(tX_i)\right] &\overset{\text{indep.}}{=} \prod_i \exp(tX_i) \\ &\leqslant \prod_i \exp(Ct^2\|X_i\|_{\psi_2}^2) = \exp(t^2K^2)\end{aligned}$$

where $K^2 = C\sum_i \|X_i\|_{\psi_2}^2$. Then, by Lemma 2.12 ($4 \Rightarrow 2$), we have

$$\left\|\sum_i X_i\right\|_{\psi_2} \leqslant C_1 K = C_1 C \sum_i \|X_i\|_{\psi_2}^2,$$

where C_1 is an absolute constant. ∎

A direct consequence of rotation invariance is Hoeffdings-type inequality for sums of independent subgaussian random variables. 6. Lecture 21.11.2022

Proposition 2.15 (Hoeffding-type inequality for sub-gaussian)**.** Let $X_1, ...X_N$ be independent centered sub-gaussian random variables. Let

$$K = \max_i \|X_i\|_{\psi_2}.$$

Then for every $a = (a_1, ..., a_N) \in \mathbb{R}^N$ and $t > 0$, we have

$$P\left(\left|\sum_{i=1}^N a_i X_i\right| > t\right) \leqslant e \cdot \exp\left(-\frac{ct^2}{K^2\|a\|_2^2}\right)$$

where $c > 0$ is an absolute constant.

Proof. First note that linear combinations of subgaussian random variables are again subgaussian:

$$\left\|\sum_i a_i X_i\right\|_{\psi_2}^2 \overset{\text{rotation inv}}{\leqslant} C\sum_i \|a_i X_i\|_{\psi_2}^2 \leqslant CK^2\|a\|_2^2.$$

The tail decay follows from Lemma 2.12. ■

The same works for moments instead of tails.

Corollary 2.16 (Khintchine-type inequality)**.** Let $X_1, ...X_N$ be independent centered subgaussian random variables. Then for $p \geqslant 2$ and any sequence of coefficients $a \in \mathbb{R}^N$

$$\left(\mathbb{E}\left|\sum_i a_i X_i\right|^p\right)^{\frac{1}{p}} \leqslant C \cdot \sqrt{p} \cdot \|a\|_2, \tag{2.8}$$

where C is an absolute constant. Furthermore, if the X_i's also have unit variance, then

$$\left(\mathbb{E}\left|\sum_i a_i X_i\right|^p\right)^{\frac{1}{p}} \geqslant \|a\|_2. \tag{2.9}$$

Proof. The inequality (2.8) follows directly from Lemma 2.12 by using a similar argument as in Proposition 2.15. For the inequality (2.9), we estimate

$$\begin{aligned}
\left(\mathbb{E}\left|\sum_i a_i X_i\right|^p\right)^{\frac{1}{p}} &= \left(\mathbb{E}\left|\sum_i a_i X_i\right|^{\frac{p}{2}\cdot 2}\right)^{\frac{1}{p}} \\
&\underset{x\mapsto x^{\frac{p}{2}}}{\overset{\text{Jensen}}{\geqslant}} \left(\mathbb{E}\left|\sum_i a_i X_i\right|^2\right)^{\frac{p}{2}\cdot\frac{1}{p}} \\
&= \left(\mathbb{E}\left[\sum_{i,j} a_i a_j X_i X_j\right]\right)^{\frac{1}{2}} \\
&\overset{\text{indep.}}{=} \left(\mathbb{E}\left[\sum_i a_i^2 X_i^2\right] + \underbrace{\sum_{i\neq j} a_i a_j \mathbb{E}[X_i]\mathbb{E}[X_j]}_{=0}\right)^{\frac{1}{2}} \\
&= \left(\sum_i a_i^2 \underbrace{\mathbb{E}X_i^2}_{=1}\right)^{\frac{1}{2}} = \|a\|_2
\end{aligned}$$

and hereby the claim. ■

2.5 Subexponential random variables

What if variables are not subgaussian? We talk about heavy-tailed random variables. In homework, we will see similar equivalence between tails, moments and super-exponential moments for any tail decay with rate $\exp(t^\alpha)$ with $\alpha \geqslant 1$. The "heaviest" tail in this framework would be $\alpha = 1$. Similar to the normal distribution as a prototype distribution for subgaussian, a prototype distribution for this case is the exponential distribution given by

$$P(X \geqslant t) = e^{-t}, \qquad t \geqslant 0.$$

Lemma 2.17. Let X be a random variable. Then the following properties are equivalent with parameters $K_i > 0$ differing from each other by at most an absolute constant factor:

(1) Tails:
$$P(|X| > t) \leqslant \exp\left(1 - \frac{t}{K_1}\right) \qquad \forall t \geqslant 0$$

(2) Moments:
$$(\mathbb{E}|X|^p)^{\frac{1}{p}} \leqslant K_2 \cdot p \qquad \forall p \geqslant 1$$

(3) Exponential moments:
$$\mathbb{E}\left[\exp\left(\frac{X}{K_3}\right)\right] \leqslant e$$

Proof. Special case of Exercise. ■

Definition 2.18 (Sub-exponential random variables)**.** A random variable X that satisfies one of the properties in Lemma 2.17 is called a **sub-exponential** random variable. The **sub-exponential** norm of X, denoted by $\|X\|_{\psi_1}$, is defined to be the smallest parameter K_2, i.e.

$$\|X\|_{\psi_1} = \sup_{p \geqslant 1} p^{-1} (\mathbb{E}[|X|^p])^{\frac{1}{p}}.$$

What are the relations between subgaussian and subexponential random variables?

Lemma 2.19 (Sub-exponential is sub-gaussian squared)**.** A random variable X is subgaussian if and only if X^2 is sub exponential. Moreover,

$$\|X\|_{\psi_2}^2 \leqslant \|X^2\|_{\psi_1} \leqslant 2 \cdot \|X\|_{\psi_2}^2.$$

Proof. First note that

$$\|X\|_{\psi_2}^2 = \left(\sup_{p \geqslant 1} p^{-\frac{1}{2}} (\mathbb{E}[|X|^p])^{\frac{1}{p}}\right)^2 = \sup_{p \geqslant 1} p^{-1} (\mathbb{E}[|X|^p])^{\frac{2}{p}}.$$

On one hand,

$$\begin{aligned}
\sup_{p \geqslant 1} p^{-1} (\mathbb{E}[|X|^p])^{\frac{2}{p}} &= \frac{1}{2} \sup_{p \geqslant 1} \left(\frac{p}{2}\right)^{-1} (\mathbb{E}[|X^2|^{\frac{p}{2}}])^{\frac{2}{p}} \\
&\overset{q := \frac{1}{2}p}{=} \frac{1}{2} \sup_{q \geqslant \frac{1}{2}} q^{-1} (\mathbb{E}[|X^2|^q])^{\frac{1}{q}} \\
&\leqslant \frac{1}{2} \sup_{q \geqslant 1} q^{-1} (\mathbb{E}[|X^2|^q])^{\frac{1}{q}} = \frac{1}{2} \|X^2\|_{\psi_1},
\end{aligned}$$

and thus

$$\|X^2\|_{\psi_1} \leqslant 2 \cdot \|X\|_{\psi_2}^2.$$

On the other hand,

$$\sup_{p \geqslant 1} p^{-1} (\mathbb{E}[|X|^p])^{\frac{2}{p}} \overset{\text{Jensen}}{\leqslant} \sup_{p \geqslant 1} p^{-1} (\mathbb{E}[|X^2|^p])^{\frac{1}{p}} = \|X^2\|_{\psi_1},$$

thereby establishing the claim. ■

Recall that there are four equivalent defining properties of a *centered* subgaussian random variable, but only three of a subexponential. The moment generating function property is missing.

Problem: Even for the prototype exponential distribution, the moment generating function is not finite for $t \geqslant 1$. We thus need a "local" version.

Lemma 2.20 (Moment generating function of subexponential random variables)**.** Let X be a centered sub-exponential random variable. Then the following holds

$$\mathbb{E}[\exp(tX)] \leqslant \exp(Ct^2 \|X\|_{\psi_1}^2) \qquad \forall\, |t| \leqslant \frac{c}{\|X\|_{\psi_1}}$$

with absolute constants c, C'.

Proof. As in the subgaussian case, w.l.o.g. we assume $\|X\|_{\psi_1} = 1$. Taylor expansion for centered variables implies

$$\begin{aligned}
\mathbb{E}[\exp(tX)] &= 1 + \underbrace{\mathbb{E}(tX)}_{=0} + \sum_{p=2}^{\infty} \frac{|t|^p \mathbb{E}[|X|^p]}{p!} \\
&\overset{X \text{ subexp}}{\leqslant} 1 + \sum_{p=2}^{\infty} \frac{|t|^p p^p}{p!} \\
&\overset{p! \geqslant \frac{p^p}{e^p}}{\leqslant} 1 + \sum_{p=2}^{\infty} |t|^p e^p \\
&= 1 + e^2 t^2 \sum_{p=0}^{\infty} |t|^p e^p \\
&\overset{\text{if } e|t|<1}{=} 1 + e^2 t^2 \cdot \frac{1}{1 - e|t|} \\
&\overset{\text{if } e|t| \leqslant \frac{1}{2}}{\leqslant} 1 + 2e^2 t^2 \leqslant \exp(2e^2 t^2)
\end{aligned}$$

thereby the claim. ■

By the central limit theorem, a sum of independent subexponential variables will be sub-gaussian in the limit. For non-asymptotic regime, we have a combination of subgaussian (from the limit behaviour) and subexponential (behaviour of variables)

Proposition 2.21 (Bernstein-type inequality)**.** Let $X_1, ..., X_N$ be independent centered sub-exponential random variables and let $K = \max_i \|X\|_{\psi_1}$. Then for any $a = (a_1, ..., a_N) \in \mathbb{R}^N$ it holds

$$P\left(\left|\sum_{i=1}^N a_i X_i\right| \geqslant t\right) \leqslant 2\exp\left(-c\min\left\{\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right\}\right) \quad \forall\, t \geqslant 0$$

where $c > 0$ is an absolute constant.

Proof. Using homogenity, we assume $K = 1$ and set $S := \sum_i a_i X_i$. We have

$$\begin{aligned} P(S > t) =& P(\exp(\lambda S) > \exp(\lambda t)) \\ &\overset{\text{Mkv}}{\leqslant} \exp(-\lambda t)\mathbb{E}[\exp(\lambda S)] \\ &\overset{\text{indep}}{=} \exp(-\lambda t)\prod_i \mathbb{E}[\exp(\lambda a_i X_i)] \quad \forall\, \lambda > 0. \end{aligned}$$

We want to use Lemma 2.20 which requires $|\lambda a_i| \leqslant c$ (we assumed $K = 1$). Thus, we choose $\lambda \leqslant \frac{c}{\|a\|_\infty}$. We have then

$$\begin{aligned} P(S \geqslant t) &\overset{\text{Lem 2.20}}{\leqslant} \exp(-\lambda t)\prod_i \exp(C\lambda^2 a_i^2) \\ &= \exp\left(-\lambda t + C\lambda^2\|a\|_2^2\right) \qquad \forall\, \lambda \leqslant \frac{c}{\|a\|_\infty} \end{aligned}$$

To control the exp here, we need

$$\lambda \leqslant \frac{t}{2C\|a\|_2^2}.$$

Hence, we choose

$$\lambda = \min\left\{\frac{t}{2C\|a\|_2^2}, \frac{c}{\|a\|_\infty}\right\}.$$

Then, we have

$$P(S > t) \leqslant \exp\left(-\frac{\lambda}{2}t\right) = \exp\left(-\min\left\{\frac{t^2}{4C\|a\|_2^2}, \frac{ct}{2\|a\|_\infty}\right\}\right)$$

Similarity, we can show

$$P(-S > t) \leqslant \exp\left(-\min\left\{\frac{t^2}{4C\|a\|_2^2}, \frac{ct}{2\|a\|_\infty}\right\}\right).$$

Consequently,

$$P(|S| > t) \leqslant 2\exp\left(-\min\left\{\frac{t^2}{4C\|a\|_2^2}, \frac{ct}{2\|a\|_\infty}\right\}\right)$$

and hereby the claim. ∎