# MATH 7251: High-Dimensional Probability Lecture Notes 

Galyna V. Livshyts

Scribes: Max Dabagia, Ruijia Cao, Dekang Meng, Xuanzhou Chen, Miao Li, Wanying Fu, Rakshit Naidu, Hanyang Jiang, Rahul Sethi, Chengyue Huang, Yuzhou Wang, William Held, Sam van Der Poel, Akash Harapanahalli, Jingyi Zhang, Chen Lin, Daqian Bao, Zichong Li


#### Abstract

Disclaimer: these lecture notes are currently under construction, and may not be fully proofread yet. If you spot a typo please let me know! Also, the references are in the process of being matched to the text.


## Contents

1 Notation and Preliminaries ..... 3
1.1 Notation ..... 3
1.2 Preliminaries from geometry in High dimension and convexity ..... 4
1.3 Preliminaries from linear algebra ..... 4
1.4 Preliminaries from Functional Analysis and Probability ..... 5
2 The Probabilistic Method and the concept of the high-dimensional phe- nomena ..... 6
2.1 A cool fact about the cube ..... 6
2.2 Random rounding ..... 7
2.3 Proof of the cool fact about the cube ..... 8
2.4 The Randomized Carathéodory theorem ..... 9
2.5 The standard $\varepsilon$-net argument ..... 12
2.6 A lattice net ..... 15
2.7 An application: an efficient net for matrix multiplication ..... 18
2.8 High-dimensional phenomenon: some themes ..... 21
3 Concentration Inequalities for sums of independent random variables ..... 23
3.1 What is a concentration inequality? ..... 23
3.2 Hoeffding's inequality ..... 24
3.3 Chernoff's inequality ..... 26
3.4 Application to Random Graphs ..... 27
3.5 Sub-Gaussian Random Variables ..... 29
3.6 General Hoeffding's inequality and Khinchine's inequality ..... 32
3.7 Sub-Exponential Random Variables ..... 34
3.8 Bernstein's inequality ..... 37
3.9 Concentration of the norm of a random vector with independent sub-Gaussian coordinates ..... 38
3.10 Sub-Gaussian Random vectors ..... 39
3.11 Grothendieck's inequality ..... 41
4 Random Matrices ..... 43
4.1 Norm of a sub-Gaussian random matrix ..... 44
4.2 Two-sided bounds for intermediate singular values of tall enough random ma- trices ..... 46
4.3 Matrix Bernstein Inequality ..... 48
4.4 Non-asymptotic bounds for the smallest singular value of random matrices ..... 50
4.4.1 General discussion about the smallest singular value of a random matrix ..... 50
4.4.2 Small ball (or anti-concentration) assumption and the tensorization lemma ..... 52
4.4.3 The smallest singular value of tall random matrices ..... 55
4.4.4 A net construction which works with high probability for matrices with independent columns ..... 56
4.5 Proof of Theorem 4.25. ..... 57
4.6 The smallest singular value of square random matrices ..... 63
4.6.1 Rudelson-Vershynin decomposition of the sphere ..... 65
4.6.2 Survey of results regarding the smallest singular value of square ran- dom matrices ..... 66
4.7 Proof of Theorem 4.45 Part 1 ..... 68
4.7.1 Compressible Vectors ..... 69
4.7.2 Incompressible Vectors ..... 70
4.7.3 Distance Theorem ..... 72
4.7.4 Proof of the first part of Theorem 4.45 ..... 74
5 Gaussian Random Processes ..... 75
5.1 Basic Concepts and Examples ..... 75
5.2 Slepian's Inequality ..... 76
5.2.1 Gaussian Interpolation ..... 77
5.2.2 Proof of Slepian's Inequality ..... 79
5.2.3 The Sudakov-Fernique Inequality ..... 81
5.2.4 Application of Sudakov-Fernique Inequality to Gaussian Random Ma- trices ..... 82
6 The Semigroup method ..... 84
6.1 Basic definitions and set up ..... 85
6.2 Properties ..... 86

## 1 Notation and Preliminaries

### 1.1 Notation

- $\mathbb{R}^{n}$ is the $n$-dimensional Euclidean space
- Lebesgue measure (volume) in $\mathbb{R}^{n}$ of a measurable set $A \subset \mathbb{R}^{n}$ is denoted by $|A|$
- $\mathbb{N}$ is the set of positive integers
- $\|x\|_{p}=\left(\left|x_{1}\right|^{p}+\cdots+\left|x_{n}\right|^{p}\right)^{1 / p}$ is the $p$-norm in $\mathbb{R}^{n}$ for $p \geq 1$
- $\|x\|_{\infty}=\max _{i=1, \ldots, n}\left|x_{i}\right|$ is the $\infty$-norm
- $|x|=\|x\|_{2}$ is a shorthand for Euclidean length
- $\mathbf{B}_{p}^{n}=\left\{x \in \mathbb{R}^{n}:\|x\|_{p} \leq 1\right\}$ is $p$-ball in $\mathbb{R}^{n}$
- $\mathbb{S}^{n-1}=\partial \mathbf{B}_{2}^{n}=\left\{x \in \mathbb{R}^{n}:|x|=1\right\}$ is the $n$-dimensional hypersphere, or the boundary of $B_{2}^{n}$
- $\langle x, y\rangle=\sum_{i=1}^{n} x_{i} y_{i}$ is the standard inner product
- For $\theta \in \mathbb{S}^{n-1}$, we have the hyperplane

$$
\theta^{\perp}=\left\{x \in \mathbb{R}^{n}:\langle x, \theta\rangle=0\right\}
$$

and the affine hyperplane

$$
\theta^{\perp}+t \theta=\left\{x \in \mathbb{R}^{n}:\langle x, \theta\rangle=t\right\}
$$

for all $t \in \mathbb{R}$.

- A half-space is a set of the form $\left\{x \in \mathbb{R}^{n}:\langle x, \theta\rangle \leq t\right.$, for some given $\theta \in \mathbb{S}^{n-1}$ and $t \in \mathbb{R}$.
- A strip is a set of the form $\left\{x \in \mathbb{R}^{n}:|\langle x-y, \theta\rangle| \leq t\right.$, for some given $y \in \mathbb{R}^{n}, \theta \in \mathbb{S}^{n-1}$ and $t \geq 0$.
- Fix $x \in \mathbb{R}$. Then $[x]$ is the floor function, the largest integer which is no larger than $x$.


### 1.2 Preliminaries from geometry in High dimension and convexity

Definition 1.1 (Convex Set). A set $K \subseteq \mathbb{R}^{n}$ is called convex if for all $x, y \in K$, the line segment $[x, y]=\{\lambda x+(1-\lambda) y: \lambda \in[0,1]\}$ is contained in $K$.

Note that strips, half-spaces and $B_{p}^{n}$ for $p \geq 1$ are convex sets. On the other hand, $B_{p}^{n}$ for $p<1, \mathbb{S}^{n-1}$, sets which are not 1 -connected, are non-convex. Below see an example of a non-convex set:


Any convex set is the intersection of (possibly infinitely many) half-spaces. A convex polytope is an intersection of finitely many half-spaces.

### 1.3 Preliminaries from linear algebra

The operator norm $\|\cdot\|_{o p}$ of a matrix $A$ is defined by

$$
\begin{equation*}
\|\mathbf{A}\|_{o p}=\sup _{x \in \mathbb{R}^{n} \backslash\{0\}} \frac{|\mathbf{A} x|}{|x|}=\sup _{y \in \mathbb{S}^{n-1}}|A y| . \tag{1}
\end{equation*}
$$

Definition 1.2 (Hilbert-Schmidt norm). Given a matrix $\mathbf{A}=\left(a_{i j}\right)$,

$$
\|\mathbf{A}\|_{H S}=\sqrt{\sum_{i, j} a_{i j}^{2}}=\sqrt{\sigma_{1}^{2}+\ldots+\sigma_{n}^{2}},
$$

where $\sigma_{1} \geq \cdots \geq \sigma_{n}$ are the singular values of $\mathbf{A}$.
Recall that the smallest singular value is also defined as

$$
\sigma_{n}(A)=\inf _{x \in \mathbb{S}_{n-1}}|A x| .
$$

Remark 1.3. Note that

$$
\sqrt{n}\|\mathbf{A}\|_{o p} \geq\|\mathbf{A}\|_{H S} \geq\|\mathbf{A}\|_{o p},
$$

since $\sigma_{i} \leq \sigma_{1}$, and $\sigma_{1} \leq \sqrt{\sigma_{1}^{2}+\ldots+\sigma_{n}^{2}} \leq \sqrt{n} \sigma_{1}$.

Theorem 1.4 (Spectral Decomposition). Let $A$ be a symmetric matrix over $\mathbb{R}$ with $n$ eigenvalues $\lambda_{1}, \cdots, \lambda_{n} \in \mathbb{R}$ and corresponding eigenvectors $u_{1}, \cdots, u_{n} \in \mathbb{S}^{n-1}$. Then

$$
\begin{equation*}
A=\sum_{i=1}^{n} \lambda_{i} u_{i} \otimes u_{i} \tag{2}
\end{equation*}
$$

Note that for $\forall u \in \mathbb{R}^{n}, u \otimes u=u u^{T}=\left[\begin{array}{cccc}u_{1}^{2} & u_{1} u_{2} & \cdots & u_{1} u_{n} \\ u_{1} u_{2} & u_{2}^{2} & \cdots & u_{2} u_{n} \\ \vdots & \vdots & \vdots & \vdots \\ u_{1} u_{n} & u_{2} u_{n} & \cdots & u_{n} u_{n}\end{array}\right]$ is a rank-1 matrix. For $\forall x \in \mathbb{R}^{n}, A x=\sum_{i=1}^{n} \lambda_{i}\left\langle x, u_{i}\right\rangle u_{i}$.

Definition 1.5 (Functions on Matrices). For any function $f: \mathbb{R} \rightarrow \mathbb{R}$ and $n \times n$ symmetric matrix $X=\sum_{i=1}^{n} \lambda_{i} u_{i} \otimes u_{i}$, then $f(X):=\sum_{i=1}^{n} f\left(\lambda_{i}\right) u_{i} \otimes u_{i}$.

### 1.4 Preliminaries from Functional Analysis and Probability

Recall that for any non-negative random variable

$$
\begin{equation*}
\mathbb{E} X=\int_{0}^{\infty} \mathbb{P}(X>t) d t \tag{3}
\end{equation*}
$$

Indeed, $X=\int_{0}^{\infty} 1_{\{X>t\}} d t$, and therefore

$$
\mathbb{E} X=\mathbb{E} \int_{0}^{\infty} 1_{\{X>t\}} d t=\int_{0}^{\infty} \mathbb{E} 1_{\{X>t\}} d t=\int_{0}^{\infty} \mathbb{P}(X>t) d t
$$

Lemma 1.6 (Markov). Let $X \geq 0$ be a random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then, for $t>0$, we have

$$
\mathbb{P}(X>t) \leq \frac{\mathbb{E} X}{t}
$$

Proof. For any $t>0$, note that

$$
\mathbb{E} X=\int_{0}^{\infty} \mathbb{P}(X>s) \mathrm{d} s \geq \int_{0}^{t} \mathbb{P}(X>s) \mathrm{d} s \geq t \cdot \mathbb{P}(X>t)
$$

Definition 1.7. A function $F: \mathbb{R}^{n} \rightarrow \mathbb{R}$ is called convex if for all $x, y \in \mathbb{R}^{n}$

$$
F(\lambda x+(1-\lambda) y) \leq \lambda F(x)+(1-\lambda) F(y)
$$

Remark 1.8. Note that if $F$ is convex, then we may conclude inductively that

$$
F\left(\sum_{i=1}^{m} \lambda_{i} x_{i}\right) \leq \sum_{1 \leq i \leq m} \lambda_{i} F\left(x_{i}\right)
$$

where $\lambda_{1}+\cdots+\lambda_{m}=1$.
Theorem 1.9 (Jensen Inequality). Let $\mu$ be any probability measure and $g \in L^{1}\left(\mathbb{R}^{n}\right)$. Let $F: \mathbb{R} \rightarrow \mathbb{R}$ be a convex function. Then,

$$
F\left(\int_{\mathbb{R}^{n}} g \mathrm{~d} \mu\right) \leq \int_{\mathbb{R}^{n}} F(g) \mathrm{d} \mu
$$

Remark 1.10. Note that Jensen inequality implies Remark 1.8 by taking $\mu$ to be the discrete measure on $\mathbb{R}^{n}$ with $\operatorname{supp} \mu=\{1, \ldots, m\}$ and $\mu(\{i\})=\lambda_{i}$ for $1 \leq i \leq m$.

Definition 1.11. Let $X$ be a random variable. For the $p \geq 1$, the $p$-norm of $X$ is defined by

$$
\|X\|_{p}=\left(\mathbb{E}|X|^{p}\right)^{1 / p}
$$

Theorem 1.12 (Minkowski). Let $X, Y$ be two random variables and $p \geq 1$. Then,

$$
\|X+Y\|_{p} \leq\|X\|_{p}+\|Y\|_{p}
$$

Theorem 1.13 (Cauchy-Schwarz). Let $a, b \in V$, where $(V,\langle\cdot, \cdot\rangle)$ is an inner product space. Then,

$$
|\langle a, b\rangle| \leq\|a\| \cdot\|b\|,
$$

where $\|\cdot\|$ denotes the norm induced by $\langle\cdot, \cdot\rangle$.
Theorem 1.14 (Hölder's Inequality). Suppose that $p, q \geq 1$ and $1 / p+1 / q=1$. Then,

$$
\int|f g| \mathrm{d} \mu \leq\left(\int|f|^{p} \mathrm{~d} \mu\right)^{1 / p} \cdot\left(\int|g|^{q} \mathrm{~d} \mu\right)^{1 / q}
$$

## 2 The Probabilistic Method and the concept of the high-dimensional phenomena

### 2.1 A cool fact about the cube

We will mostly study $\mathbb{R}^{n}$ - the $n$-dimensional Euclidean space - where $n$ is a large, positive integer (or, in other words, the dimension $n$ of our space is high). We will often think that $n \rightarrow \infty$ and analyze things asymptotically. We know intuitively that functions on $\mathbb{R}^{2}$ are more complex than those on $\mathbb{R}$, and functions on $\mathbb{R}^{3}$ are more complex than on $\mathbb{R}^{2}$. However, in some sense, objects in high-dimensional spaces actually become simpler, and more predictable. We will soon see some examples of this phenomenon, but for the time being we concentrate on the so-called probabilistic method in High-Dimensional Geometry.

We now state the following cool fact:

Fact 2.1. Consider the cube $[0,1]^{n}$ (with side-length 1) and pick $x \in[0,1]^{n}$. Take any (arbitrary!) $\theta \in \mathbb{S}^{n-1}$, and consider the strip centered at $x$ orthogonal to $\theta$ of width 1 , i.e.

$$
S=\left\{y \in \mathbb{R}^{n}:|\langle y-x, \theta\rangle| \leq \frac{1}{2}\right\}
$$

Then at least one vertex of $[0,1]^{n}$ belongs to $S$.
The proof will be based on the probabilistic method. Imagine that someone shows you a non-transparent box full of balls and tells you that if you draw a ball from it then with probability 0.3 you get a red ball from it. Then you can conclude that there exists at least one red ball in the box.


In our case, the box will be the vertices of the cube, and the red ball will be the vertex with the desired property of falling into the specific strip.

### 2.2 Random rounding

We will first define the concept of randomized rounding, which was introduced in one dimension by Raghavan, Thompson [31], and later extended and studied by many authors including Kannan, Vempala [16], Alon, Klartag [2], Klartag, Livshyts [18], see a survey by Srinivasan [42]. This object is very useful in Computer Science, as well as in High-Dimensional Probability and related areas.

Definition 2.2 (Random Rounding). We outline the definition in two steps.
Step 1 (dimension 1). For $x \in \mathbb{R}$, define $\eta_{x}$ to be a random variable such that

$$
\eta_{x}= \begin{cases}{[x]} & \text { w.p. } 1-p \\ {[x]+1} & \text { w.p. } p\end{cases}
$$

where $p=p(x)=x-[x]$. Note that $p$ is chosen so that $\eta_{x}$ is centered at $x$ :

$$
\mathbb{E} \eta_{x}=[x](1-(x-[x]))+([x]+1)(x-[x])=[x]+x-[x]=x .
$$

Step 2 (dimension $n$ ). For $x \in \mathbb{R}^{n}$, define $\eta_{x}$ to be a random vector taking values in the vertex set of the lattice cube which $x$ falls into, so that the coordinates of $\eta_{x}$ are independent
and $\mathbb{E} \eta_{x}=x$. In other words, each coordinate of $x$ is independently randomly rounded using the 1-dimensional definition. Namely, we let $\eta_{x}=\left(\left(\eta_{x}\right)_{1}, \ldots,\left(\eta_{x}\right)_{n}\right)$, where the random variables $\left(\eta_{x}\right)_{i}$ are independent, and distributed as follows:

$$
\left(\eta_{x}\right)_{i}= \begin{cases}{\left[x_{i}\right]} & \text { w.p. } 1-\mathrm{p} \\ {\left[x_{i}\right]+1} & \text { w.p. p }\end{cases}
$$

where $p=p(i, x)=x_{i}-\left[x_{i}\right]$,


### 2.3 Proof of the cool fact about the cube

First, recall the following basic fact:
Lemma 2.3. For independent random variables $X_{1}, \ldots, X_{n}$,

$$
\operatorname{Var}\left(X_{1}+\cdots+X_{n}\right)=\operatorname{Var}\left(X_{1}\right)+\cdots+\operatorname{Var}\left(X_{n}\right)
$$

Proof. First, consider

$$
\mathbb{E}\left(X_{1}+\ldots+X_{n}\right)^{2}=\sum_{i=1}^{n} \mathbb{E} X_{i}^{2}+\sum_{i \neq j} 2 \mathbb{E} X_{i} X_{j}=\sum_{i=1}^{n} \mathbb{E} X_{i}^{2}+\sum_{i \neq j} 2 \mathbb{E} X_{i} \mathbb{E} X_{j} .
$$

where in the last step we used that by independence, $\mathbb{E} X_{i} X_{j}=\mathbb{E} X_{i} \mathbb{E} X_{j}$. On the other hand,

$$
\left(\mathbb{E} X_{1}+\ldots+\mathbb{E} X_{n}\right)^{2}=\sum_{i=1}^{n}\left(\mathbb{E} X_{i}\right)^{2}+\sum_{i \neq j} 2 \mathbb{E} X_{i} \mathbb{E} X_{j}
$$

Therefore,

$$
\begin{gathered}
\operatorname{Var}\left(X_{1}+\cdots+X_{n}\right)=\mathbb{E}\left(X_{1}+\ldots+X_{n}\right)^{2}-\left(\mathbb{E}\left(X_{1}+\ldots+X_{n}\right)\right)^{2}=\sum_{i=1}^{n} \mathbb{E} X_{i}^{2}-\sum_{i=1}^{n}\left(\mathbb{E} X_{i}\right)^{2}= \\
\operatorname{Var}\left(X_{1}\right)+\cdots+\operatorname{Var}\left(X_{n}\right) .
\end{gathered}
$$

We may now proceed with the proof of the cool fact.
Proof of Fact 2.1. For the vector $x \in[0,1]^{n}$ consider the random rounding $\eta_{x}$. Note that $\mathbb{E}\left\langle\eta_{x}-x, \theta\right\rangle=0$, so

$$
\mathbb{E}\left\langle\eta_{x}-x, \theta\right\rangle^{2}=\operatorname{Var}\left\langle\eta_{x}-x, \theta\right\rangle=\sum_{i=1}^{n} \mathbb{E}\left(\left(\eta_{x}\right)_{i}-x_{i}\right)^{2} \theta_{i}^{2}
$$

by Lemma 2.3. Hence,

$$
\mathbb{E}\left\langle\eta_{x}-x, \theta\right\rangle^{2}=\sum_{i=1}^{n} \theta_{i}^{2}\left(\mathbb{E}\left(\eta_{x}\right)_{i}^{2}-x_{i}^{2}\right)
$$

Note that (by definition of random rounding),

$$
\mathbb{E}\left(\eta_{x}\right)_{i}^{2}=0 \cdot\left(1-x_{i}\right)+1 \cdot x_{i}=x_{i}
$$

so

$$
\mathbb{E}\left(\eta_{x}\right)_{i}^{2}-x_{i}^{2}=x_{i}-x_{i}^{2} \leq \frac{1}{4}
$$

where we use the fact that for any $a \in[0,1]$ one has $a-a^{2} \leq \frac{1}{4}$. We conclude that

$$
\begin{equation*}
\mathbb{E}\left\langle\eta_{x}-x, \theta\right\rangle^{2}=\sum_{i=1}^{n} \theta_{i}^{2} x_{i}\left(1-x_{i}\right) \leq \frac{|\theta|^{2}}{4}=\frac{1}{4} . \tag{4}
\end{equation*}
$$

Therefore, there exists some realization of $\eta_{x}$, a vertex $y \in\{0,1\}^{n}$ such that $\langle y-x, \theta\rangle^{2} \leq 1 / 4$. Hence, $|\langle y-x, \theta\rangle| \leq 1 / 2$.

### 2.4 The Randomized Carathéodory theorem

We will now state a second cool fact, which is an approximate version of Carathéodory's theorem. We follow Vershynin [53] in this subsection.
Definition 2.4 (Convex hull). Let $A \subseteq \mathbb{R}^{n}$. The convex hull of $A$ is

$$
\operatorname{conv}(A)=\left\{\sum_{i=1}^{m} \lambda_{i} x_{i}: m \in \mathbb{N}, x_{1}, \ldots, x_{m} \in A, \lambda_{i} \geq 0, \sum_{i=1}^{m} \lambda_{i}=1\right\}
$$

The expression $\sum_{i=1}^{m} \lambda_{i} x_{i}$ is called a convex combination of $x_{1}, \ldots, x_{m}$.


For example, the convex hull of $n+1$ linearly independent points in $\mathbb{R}^{n}$ is called a simplex.


The next classical result tells us that every point in a convex hull of some set in $\mathbb{R}^{n}$ can be represented as a convex combination of at most $n+1$ points from this set, i.e. $m$ in the definition of convex hull need not be larger than $n+1$.

Theorem 2.5 (Carathéodory's theorem). Let $A \subseteq \mathbb{R}^{n}$ and $x \in \operatorname{conv}(A)$. Then there exist $x_{1}, \ldots, x_{n+1} \in A$ and $\lambda_{1}, \ldots, \lambda_{n+1} \geq 0, \sum_{i=1}^{n+1} \lambda_{i}=1$ such that $x=\sum_{i=1}^{n+1} \lambda_{i} x_{i}$. In other words, $x$ belongs to a simplex with vertices spanned by $x_{1}, \ldots, x_{n+1} \in A$.

Proof. Home work!
Carathéodory's Theorem could be useful for constructing various algorithms. However, having to operate with $n+1$ points could still be too difficult if $n=100000000$, say. Would it be possible to represent $x$ with less than $n+1$ points from $A$ ? In general, of course, not. However, if we were willing to represent $x$ approximately, with some small error, then we could get away with using a lot less points, potentially:

Fact 2.6 (Randomized/Approximate Carathéodory). Suppose $A \subseteq \mathbb{R}^{n}$ with diameter

$$
\operatorname{diam}(A)=\sup _{x, y \in A}|x-y| \leq 1
$$

Then for all $x \in \operatorname{conv}(A), k \in \mathbb{N}$, there exist $x_{1}, \ldots, x_{k} \in A$, such that

$$
\left\|x-\frac{1}{k} \sum_{i=1}^{k} x_{i}\right\| \leq \frac{1}{\sqrt{k}} .
$$

Note that we allow repetitions among the points.
While $\frac{1}{k} \sum_{i=1}^{k} x_{i}$ does not equal to $x$, but rather approximates it with error $\frac{1}{\sqrt{k}}$, the advantage is that we only need $k$ points, and potentially $k \ll n+1$. Another interesting feature here is that we can take $\lambda_{1}=\lambda_{2}=\ldots=\lambda_{k}=\frac{1}{k}$, rather than deal with different weights.


Proof of Fact 2.6. Choose any $x \in \operatorname{conv}(A)$. By Carathéodory's theorem, there exist $z_{1}, \ldots, z_{n+1} \in$ $A$ and $\lambda_{1}, \ldots, \lambda_{n+1} \geq 0$ such that $\sum_{i=1}^{n+1} \lambda_{i}=1$ and

$$
x=\sum_{1 \leq i \leq n+1} \lambda_{i} z_{i} .
$$

Consider a random vector $Z$ which takes value $z_{1}$ with probability $\lambda_{1}$, value $z_{2}$ with probability $\lambda_{2}$, and so on (overall, this random vector takes $n+1$ values $z_{1}, \ldots, z_{n} \in A$ ). Note that

$$
\mathbb{E} Z=\sum_{1 \leq i \leq n+1} \lambda_{i} z_{i}=x
$$

Now, given an integer $k$, consider $k$ i.i.d. copies of $Z$ denoted by $Z_{1}, \ldots, Z_{k}$ (the existence of these random vectors is guaranteed by Kolmogorov's extension theorem, see Theorem 1 in section 2.9 of [57]). Note that by Lemma 2.3,

$$
\mathbb{E}\left|x-\frac{1}{k} \sum_{1 \leq i \leq k} Z_{i}\right|^{2}=\frac{1}{k^{2}} \sum_{1 \leq i \leq k} \mathbb{E}\left|x-Z_{i}\right|^{2} \leq \frac{1}{k^{2}} \cdot k=\frac{1}{k},
$$

where in the last passage we used the fact that $x \in \operatorname{conv}(A)$, and $\frac{1}{k} \sum_{1 \leq i \leq k} Z_{i} \in A \subset \operatorname{conv}(A)$ with probability 1 , and the diameter of the convex hull of $A$ is bound from above by the diameter of $A$ (homework!), which in turn is bounded by 1 .

We conclude that

$$
\mathbb{E}\left|x-\frac{1}{k} \sum_{1 \leq i \leq k} Z_{i}\right|^{2} \leq \frac{1}{k}
$$

Hence, there exist $y_{1}, \ldots, y_{k} \in A$, the realizations of $Z_{1}, \ldots, Z_{n}$, such that

$$
\left|x-\frac{1}{k} \sum_{1 \leq j \leq k} y_{i}\right| \leq \frac{1}{\sqrt{k}},
$$

which proves our Fact.
The Approximate Carathéodory Theorem has, for example, the following useful consequence:

Corollary 2.7. Let $P$ be a polytope in $\mathbb{R}^{n}$ with at most $N$ vertices, i.e.,

$$
P=\operatorname{conv}\left\{x_{1}, \ldots, x_{N}\right\}
$$

where each $x_{i} \in \mathbb{R}^{n}$. Suppose that $\operatorname{diam}(P) \leq 1$. Fix any $\varepsilon>0$. Then, $P$ can be covered by at most $N^{\left\lfloor 1 / \varepsilon^{2}\right\rfloor}$ euclidean balls of radius $\varepsilon$.

Proof. Home work!
While we leave this fact as a home work, in the next sub-section we prove a related and very useful result.

### 2.5 The standard $\varepsilon$-net argument

Suppose we would like to view the sphere or the unit ball as a discrete set, in which each point is represented approximately. What would be a good way to do that? We start by presenting the most classical result.

Theorem 2.8 (Classical $\varepsilon$-net Construction). Let $n \in \mathbb{N}$ and $\varepsilon>0$. Then, there exists a positive integer $m \leq\left(\frac{2+\varepsilon}{\varepsilon}\right)^{n}$, and $y_{1}, \ldots, y_{m} \in \mathbf{B}_{2}^{n}$, such that the unit ball $\mathbf{B}_{2}^{n}$ is covered by balls of radius $\varepsilon$ with the centers at $y_{i}$, that is,

$$
\mathbf{B}_{2}^{n} \subseteq \bigcup_{1 \leq i \leq m}\left(y_{i}+\varepsilon \mathbf{B}_{2}^{n}\right) .
$$



Proof. Without loss of generality, we may assume that $\varepsilon<1$ (otherwise, the statement is straightforward). Consider a maximal packing of balls of radius $\varepsilon / 2$ in $(1+\varepsilon / 2) \mathbf{B}_{2}^{n}$, i.e., pick $y_{1}, \ldots, y_{m}$ such that $y_{i}+(\varepsilon / 2) \mathbf{B}_{2}^{n}$ are disjoint, $y_{i}+(\varepsilon / 2) \mathbf{B}_{2}^{n} \subseteq(1+\varepsilon / 2) \mathbf{B}_{2}^{n}$, and $m$ is the largest possible number of balls with such properties (see Remark below for a justification for the existence of such a packing).


First, we show that

$$
\begin{equation*}
B_{2}^{n} \subseteq \bigcup_{1 \leq i \leq m}\left(y_{i}+\varepsilon B_{2}^{n}\right) \tag{5}
\end{equation*}
$$

Suppose not. Then, there exists $x \in B_{2}^{n}$ such that $\left|x-y_{i}\right|>\varepsilon / 2$ for all $1 \leq i \leq m$, and $x+\frac{\varepsilon}{2} \mathbf{B}_{2}^{n} \subset\left(1+\frac{\varepsilon}{2}\right) \mathbf{B}_{2}^{n}$. But then $x+(\varepsilon / 2) \mathbf{B}_{2}^{n} \cap y_{i}+(\varepsilon / 2) \mathbf{B}_{2}^{n}=\emptyset$, contradicting the maximality of $m$. Hence (5) is confirmed. That is, we found a construction of some covering of $B_{2}^{n}$ with balls of radius $\varepsilon$; our next task is to estimate its size.

Observe that

$$
(1+\varepsilon / 2) \mathbf{B}_{2}^{n} \supseteq \bigsqcup_{1 \leq i \leq m}\left(y_{i}+\varepsilon / 2 \mathbf{B}_{2}^{n}\right)
$$

Therefore, by the additive property of the Lebesgue measure, we have

$$
\sum_{1 \leq i \leq m}\left|y_{i}+\frac{\varepsilon}{2} \mathbf{B}_{2}^{n}\right| \leq\left|(1+\varepsilon / 2) \mathbf{B}_{2}^{n}\right|
$$

where $|\cdot|$ denotes the $n$-dimensional Lebesgue measure. Therefore, we have

$$
m \cdot\left(\frac{\varepsilon}{2}\right)^{n} \leq\left(1+\frac{\varepsilon}{2}\right)^{n}
$$

Hence, $m \leq(1+2 / \varepsilon)^{n}$, as desired.
Remark 2.9 (Ideas by Ruijia Cao, not something Galyna suggested). To show that a maximal packing used in the proof above exists, we can use the following argument. Note that for $t>0$ and any packing of $m$ disjoint balls, we have

$$
\bigsqcup_{i=1}^{m}\left(y_{i}+t \mathbf{B}_{2}^{n}\right) \subseteq \mathbf{B}_{2}^{n},
$$

Then, $m \leq 1 / t^{n}$. Let $\varepsilon>0$. We will use the following algorithm to explicitly construct a maximal packing of $\mathbf{B}_{2}^{n}$ using balls with radii $\varepsilon / 2$ :

1. First, pick any $x_{1} \in \mathbf{B}_{2}^{n}$.
2. In the second step, choose $x_{2} \in \mathbf{B}_{2}^{n}$ such that $\left\|x_{2}-x_{1}\right\| \geq \varepsilon$.
3. At the $k$-th step, suppose that $x_{1}, \ldots, x_{k-1}$ have been chosen such that $x_{i}+(\varepsilon / 2) \mathbf{B}_{2}^{n}$ are pairwise disjoint. Choose $x_{k}$ such that $\left\|x_{k}-x_{i}\right\| \geq \varepsilon$ for all $1 \leq i \leq k-1$.
4. If no such $x_{k}$ exists, then the algorithm terminates.

Since the number of balls in any packing of disjoint ball is finite, the above algorithm will terminate in finitely many steps.


Remark 2.10 (a small addition by Galyna). Arguing along the lines of the previous Remark (by Ruijia), one can consider the collection of all such "locally optimal" packings. Since the sizes of all of them are bounded by $t^{-n}$ (as was explained using the volumetric argument), and since the supremum of a bounded from above sequence is attained, we conclude that at least one of such packings is maximal. For details, see e.g. Vershynin [53].

Remark 2.11. Note that the size $m$ of the covering of the unit ball by balls of radius $\varepsilon$ is necessarily at least $\frac{1}{\varepsilon^{n}}$, since $m \cdot\left|\varepsilon \mathbf{B}_{2}^{n}\right| \geq\left|\mathbf{B}_{2}^{n}\right|$. Therefore, the bound in Theorem 2.8 is sharp up to $3^{n}$, when $\varepsilon<1$.

Therefore, one may conclude that the optimal $\varepsilon$-covering is of the size $\left(\frac{c}{\varepsilon}\right)^{n}$ where $c \in[1,3]$ (technically, the sharp c might depend on $\varepsilon$ but we could discuss the "limiting value" in some sense). There is vast literature on the subject and tighter bounds are known, however the only dimensions in which the limiting value of $c$ is known (when $\varepsilon \rightarrow 0$ ) are 2, 8 and 24 . Dimension 2 is assigned as a (difficult!) home work, while in dimensions 8 and 24 this result was established by a Fields medalist Viazovska [55], and Cohn, Kumar, Miller, Radchenko, Viazovska [56].


Remark 2.12. Since $\mathbb{S}^{n-1} \subset \mathbf{B}_{2}^{n}$, Theorem 2.8 implies that there exists a covering of size $\left(\frac{3}{\varepsilon}\right)^{n}$ of the unit sphere by $\varepsilon$-balls. In fact, a stronger result is true: there is such a cover of size at most $\left(\frac{c}{\varepsilon}\right)^{n-1}$. This important fact is left as a home work.

Theorem 2.8 is very classical and it has many powerful applications, some of which we will see in this course. However, there are situations when this result is, in fact, insufficiently sharp. Another disadvantage of the argument we presented is that it is not constructive (as one of the students cleverly pointed out), because an optimal packing is not constructed explicitly.

### 2.6 A lattice net

In this subsection, we discuss a refinement of Theorem 2.8 which is furthermore constructive (although of course one cannot only gain and not lose, so some features of Theorem 2.8 might not be kept). First, we note the following classical
Lemma 2.13. Let $n, N \in \mathbb{Z}$ be positive integers. Then, there are $\binom{N+n-1}{n-1}$ number of solutions to the following equation

$$
x_{1}+\cdots+x_{n}=N
$$

where $x_{i} \in \mathbb{Z}$ and $x_{i} \geq 0$.
Proof. Note that the number of ways to represent $N$ as an ordered sum of $n$ non-negative integers is the same as the number of ways to distribute $N$ identical balls into $n$ different boxes (any box can contain between 0 and $n$ balls). The answer to this classical combinatiorial riddle is $\binom{N+n-1}{n-1}$. Indeed, this corresponds to the number of ways to arrange a sequence of $N$ zeroes (corresponding to the balls) and $n-1$ ones (corresponding to the sides of the boxes, except the two on the edges). And this is the same as selecting $n-1$ spots to place ones in the collection of $N+n-1$ spots. See Figure 1.
Theorem 2.14. Let $R>0$. Then, there exists a constant $C>0$ such that

$$
\#\left\{x \in \mathbb{Z}^{n}:\|x\|_{1} \leq R\right\} \leq\left(\frac{C R}{n}\right)^{n}
$$



Figure 1: $(n-1)$ dividers and $N$ balls

Proof. Applying Lemma 2.13, we have

$$
\begin{aligned}
\#\left\{x \in \mathbb{Z}^{n}:\|x\|_{1} \leq R\right\} & \leq 2^{n} \cdot \#\left\{x \in \mathbb{Z}^{n}: x_{i} \geq 0, \sum_{1 \leq i \leq n} x_{i} \leq R\right\} \\
& =2^{n} \cdot \sum_{N=1}^{\lfloor R\rfloor}\binom{N+n-1}{n-1}
\end{aligned}
$$

Now, Stirling's formula states that for any $k \geq 1$

$$
k!=(1+o(1)) \cdot \sqrt{2 \pi k}\left(\frac{k}{e}\right)^{k} .
$$

Therefore, with a bit of arithmetic, one can show (Home work!) that there exists a constant $C>0$ such that

$$
2^{n} \cdot \sum_{N=1}^{\lfloor R\rfloor}\binom{N+n-1}{n-1} \leq\left(\frac{C R}{n}\right)^{n}
$$

Corollary 2.15. Fix any $\varepsilon>0$. Then, there exist $y_{1}, \ldots, y_{m} \in \mathbf{B}_{2}^{n}$ and a constant $c>0$ such that

$$
\bigcup_{1 \leq i \leq m}\left(y_{i}+\frac{\varepsilon}{\sqrt{n}} \mathbf{B}_{\infty}^{n}\right) \supseteq \mathbf{B}_{2}^{n}
$$

where $m \leq(C / \varepsilon)^{n}$.


Proof. Let $R=n / \varepsilon$. Then, by Theorem 2.14, there exist $c>0$ and $y_{1}, \ldots, y_{m} \in \mathbb{Z}^{n}$ with $\left\|y_{i}\right\|_{1} \leq R$, where $m \leq(C / \varepsilon)^{n}$. Therefore,

$$
\frac{R}{\sqrt{n}} \cdot \mathbf{B}_{2}^{n} \subseteq \bigcup_{1 \leq i \leq m}\left(y_{i}+\mathbf{B}_{\infty}^{n}\right)
$$

Now, by excluding those $y_{i}$ that lie outside a ball of radius $\frac{3}{2} \frac{R}{\sqrt{n}}$ and by re-indexing if necessary, we may choose $y_{1}, \ldots, y_{m^{\prime}} \in \frac{3}{2} \frac{R}{\sqrt{n}} \mathbf{B}_{2}^{n} \cap\left\{y_{1}, \ldots, y_{m}\right\}$ such that

$$
\frac{R}{\sqrt{n}} \cdot \mathbf{B}_{2}^{n} \subseteq \bigcup_{1 \leq i \leq m^{\prime}}\left(y_{i}+\mathbf{B}_{\infty}^{n}\right)
$$

where $m^{\prime} \leq m$. Therefore,

$$
\mathbf{B}_{2}^{n} \subseteq \bigcup_{1 \leq i \leq m}\left(\frac{\varepsilon}{\sqrt{n}} y_{i}+\frac{\varepsilon}{\sqrt{n}} \mathbf{B}_{\infty}^{n}\right)
$$

Note that each $\frac{\varepsilon}{\sqrt{n}} y_{i} \in \frac{3}{2} \mathbf{B}_{2}^{n}$ for $1 \leq i \leq m^{\prime}$ because $\left\|y_{i}\right\|_{1} \leq R$, and because

$$
\sqrt{n}|x| \geq\|x\|_{1} \geq|x|
$$

for any $x \in \mathbb{R}^{n}$.
Remark 2.16. Here, we give a comparison between Theorem 2.8 and Corollary 2.15. By Theorem 2.14 and Corollary 2.15,

$$
\begin{equation*}
\mathbf{B}_{2}^{n} \subseteq \bigcup_{1 \leq i \leq m}\left(y_{i}+\frac{\varepsilon}{\sqrt{n}} \mathbf{B}_{\infty}^{n}\right) \tag{6}
\end{equation*}
$$

where $m \leq\left(\frac{c}{\varepsilon}\right)^{n}$. On the other hand, Theorem 2.8 implies that

$$
\begin{equation*}
\mathbf{B}_{2}^{n} \subseteq \bigcup_{1 \leq i \leq m}\left(y_{i}+\varepsilon \mathbf{B}_{2}^{n}\right) \tag{7}
\end{equation*}
$$

where $m \leq\left(\frac{3}{\varepsilon}\right)^{n}$.
Note that in some sense, Corollary 2.15 is stronger than Theorem 2.8 since $\mathbf{B}_{\infty}^{n} \subseteq \sqrt{n} \mathbf{B}_{2}^{n}$ (see Figure 2), and therefore (6) implies (7). However, Corollary 2.15 does have a minor disadvantage comparing to Theorem 2.8: the constant $C>3$ is unspecified in our computation. But this usually doesn't play a role in applications.

Since $\mathbb{S}^{n-1} \subset \mathbf{B}_{2}^{n}$, and by removing some unused points of small Euclidean norm, we get the following result about sphere covering from Corollary 2.15:

Corollary 2.17 (Home Work). For any $\varepsilon>0$, there there exist $y_{1}, \ldots, y_{m} \in B_{2}^{n}$ such that for any $x \in \mathbb{S}^{n-1}$ there exists $i \in[m]$ with $\left\|x-y_{i}\right\|_{\infty}<\varepsilon / \sqrt{n}$ and $m \leq(C / \varepsilon)^{n-1}$.


Figure 2: $L_{2}$ and $L_{\infty}$ norms comparison
Remark 2.18. Note that not all $y_{i} \in \mathbb{S}^{n-1}$, as in our previous construction of the standard $\epsilon$-net argument.

We can also restate Theorem 2.14 and Corollary 2.17 in the following manner:
Theorem 2.19. 1. For all $\varepsilon>0$, there exists a collection of points $y_{1}, \ldots, y_{m} \in \mathbb{S}^{n-1}$ with $m \leq\left(\frac{C}{\varepsilon}\right)^{n}$, such that for all $x \in \mathbb{S}^{n-1}$, there exists $i \in\{1, \ldots, m\}$ such that $\left|x-y_{i}\right|<\varepsilon$.
2. For all $\varepsilon>0$, there exists $y_{1}, \ldots, y_{m} \subset \frac{3}{2} \mathbf{B}_{2}^{n} \backslash \frac{1}{2} \mathbf{B}_{2}^{n}$ with $m \leq\left(\frac{C^{\prime}}{\varepsilon}\right)^{n}$, such that for all $x \in \mathbb{S}^{n-1}$, there exists $i \in\{1, \ldots, m\}$ such that $\left\|x-y_{i}\right\|_{\infty} \leq \frac{\varepsilon}{\sqrt{n}}$.


Proof. The first part follows from By Theorem 2.8 and the standard net argument, while the second part follows from Corollary 2.17.
Remark 2.20. Note that $\frac{1}{\sqrt{n}}|a| \leq\|a\|_{\infty} \leq|a|$, where $|\cdot|=\|\cdot\|_{2}$, and therefore the second part of Theorem 2.19 is stronger than the first part, in the essential sense.

Remark 2.21. However note that in the second part of Theorem 2.19, unlike in the first part, one needs to assume that the points of the net are in $\frac{3}{2} \mathbf{B}_{2}^{n} \backslash \frac{1}{2} \mathbf{B}_{2}^{n}$, rather than simply in $B_{2}^{n}$. This is because the collection of our approximating points in the second part consists of the vertices of the lattice cubes, rather than centers! This will not make a difference for our applications, however.

### 2.7 An application: an efficient net for matrix multiplication

From now on, we will use the word net to mean a collection of points which has some good approximating properties for a given set (usually, the sphere). We will use notation $\mathcal{N}$ for a finite set of points, in place of just listing the points $\left\{y_{1}, \ldots, y_{m}\right\}$ as before, to save time. First, we outline the following

Theorem 2.22. For all $\varepsilon>0$, there exists a set $\mathcal{N} \subseteq \mathbb{S}^{n-1}$ such that $\# \mathcal{N} \leq\left(\frac{c}{\varepsilon}\right)^{n}$, and for all $x \in \mathbb{S}^{n-1}$, there exists $y \in \mathcal{N}$ such that for any integer $N \geq 0$, and for any $N \times n$ matrix A,

$$
\|\mathbf{A}(x-y)\| \leq\|\mathbf{A}\|_{o p} \cdot \varepsilon
$$

where we recall that the operator norm of a matrix was defined in (1).
Proof. By the usual net argument, for all $x \in \mathbb{S}^{n-1}$, there exists some $y \in \mathcal{N}$ such that $|x-y|<\varepsilon$. Take any matrix $\mathbf{A}$, we have $|\mathbf{A}(x-y)| \leq\|\mathbf{A}\|_{o p} \cdot|x-y| \leq\|\mathbf{A}\|_{o p} \cdot \varepsilon$.

Next, we now explain a more precise way to approximate points on the sphere in order to compare the functional $|\mathbf{A} x|$ at those points. This net construction combines the ideas we already explained in Theorem 2.19 (part 2), with the idea of random rounding which we discussed in subsection 2.2. This net construction first appears in Klartag, Livshyts [18], and was later further developed in Livshyts [25].

Theorem 2.23. Fix $n \in \mathbb{N}$. For all $\varepsilon>0$, there exists $\mathcal{N} \subset \frac{3}{2} \mathbf{B}_{2}^{n} \backslash \frac{1}{2} \mathbf{B}_{2}^{n}$ such that $\# \mathcal{N} \leq\left(\frac{C}{\varepsilon}\right)^{n}$, and so that for all $x \in \mathbb{S}^{n-1}$ and for all $N \in \mathbb{N}$, for any $N \times n$ matrix $\mathbf{A}$, there exists $y \in \mathcal{N}$ such that

$$
|\mathbf{A}(x-y)| \leq \varepsilon \cdot \frac{\|\mathbf{A}\|_{H S}}{\sqrt{n}}
$$

where the Hilbert-Schmidt norm of a matrix was defined in Definition 1.2.
Remark 2.24. As we noted before, $\frac{\|\mathbf{A}\|_{H S}}{\sqrt{n}} \cdot \varepsilon \leq\|\mathbf{A}\|_{o p} \cdot \varepsilon$, and thus the net in Theorem 2.23 is more precise than the net in Theorem 2.22. Apart from some minor differences in the order of quantifiers (which we discuss below), Theorem 2.23 is stronger than Theorem 2.22, in the essential sense (but not literally).

Proof of 2.23. Consider $\mathcal{N}$ to be the net given in Corollary 2.17. Then indeed $\mathcal{N} \subset \frac{3}{2} \mathbf{B}_{2}^{n} \backslash \frac{1}{2} \mathbf{B}_{2}^{n}$ such that $\# \mathcal{N} \leq\left(\frac{C}{\varepsilon}\right)^{n}$, and for every $x \in \mathbb{S}^{n-1}$ one may find a lattice cube

$$
Q=\frac{\epsilon}{\sqrt{n}} \prod_{i=1}^{n}\left[t_{i}, t_{i}+1\right]
$$

for some integers $t_{i}$, such that $x \in Q$, and the vertices of $Q$ all belong to the point set $\mathcal{N}$.
Consider the (scaled) random rounding $\eta_{x}$ of $x \in \mathbb{S}^{n-1}$ to the vertices of the cube $Q$ (as defined earlier).


Namely, $\eta_{x}$ is a random vector which takes values in the vertices of $Q$, has independent coordinates, and $\mathbb{E} \eta_{x}=x$. Recall from (4), after the appropriate re-scaling by $\frac{2 \epsilon}{\sqrt{n}}$, that

$$
\begin{equation*}
\mathbb{E}\left\langle x-\eta_{x}, \theta\right\rangle^{2} \leq \frac{\epsilon^{2}|\theta|^{2}}{n} \tag{8}
\end{equation*}
$$

Recall that $|\mathbf{A} y|^{2}=\sum_{i=1}^{N}\left\langle\mathbf{A}^{T} e_{i}, y\right\rangle^{2}$ (here $A^{T} e_{i}$ are the rows of $A$ ). Using the inequality (8) with $\theta=A^{T} e_{i}$, we get

$$
\begin{aligned}
\mathbb{E}\left|\mathbf{A}\left(x-\eta_{x}\right)\right|^{2} & =\mathbb{E} \sum_{i=1}^{N}\left\langle\mathbf{A}^{T} e_{i}, x-\eta_{x}\right\rangle^{2} \\
& =\sum_{i=1}^{N} \mathbb{E}\left\langle\mathbf{A}^{T} e_{i}, x-\eta_{x}\right\rangle^{2} \\
& \leq \frac{\varepsilon^{2}}{n} \sum_{i=1}^{N}\left|\mathbf{A}^{T} e_{i}\right|^{2} \\
& =\|\mathbf{A}\|_{H S}^{2} \cdot \frac{\varepsilon^{2}}{n}
\end{aligned}
$$

Since $\mathbb{E}\left|\mathbf{A}\left(x-\eta_{x}\right)\right|^{2} \leq\|\mathbf{A}\|_{H S}^{2} \cdot \frac{\varepsilon^{2}}{n}$, there exists a realization of random vector $y$ of $\eta_{x}$ such that

$$
|\mathbf{A}(x-y)|^{2} \leq \varepsilon^{2} \cdot \frac{\|\mathbf{A}\|_{H S}^{2}}{n},
$$

or equivalently,

$$
|\mathbf{A}(x-y)| \leq \varepsilon \cdot \frac{\|\mathbf{A}\|_{H S}}{\sqrt{n}}
$$

Note that any realization of $\eta_{x}$ is a vertex of $Q$ and in particular, $y \in \mathcal{N}$.
Remark 2.25. Note that:

- For both Theorem 2.22 and Theorem 2.23, the net does not depend on the matrix A.
- For Theorem 2.22, the approximation point $y$ depends on both $\mathbf{A}$ and $x$, which is not the case for Theorem 2.23. But we will see that in applications that we consider, this does not make a difference.

Example 2.26. Let us consider two concrete examples of matrices.

- $\mathbf{A}=\mathrm{Id}_{\mathrm{n}}$. Then $\|\mathbf{A}\|_{o p}=1$, and $\|\mathbf{A}\|_{H S}=\sqrt{1+\cdots+1}=\sqrt{n}$. Therefore, $\frac{\|A\|_{H S}}{\sqrt{n}}=$ $\|A\|_{o p}$, and therefore Theorems 2.22 and 2.23 provide the same precision.
- $\mathbf{A}=\left(\begin{array}{cccc}1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0\end{array}\right)$

Then $\frac{1}{\sqrt{n}}\|\mathbf{A}\|_{H S}=\frac{1}{\sqrt{n}} \ll 1=\|\mathbf{A}\|_{o p}$, and therefore, Theorem 2.23 is $\sqrt{n}$ more precise when approximating the functional $A x=x_{1}$ on the sphere, or, equivalently, any one-dimensional projection $\langle x, \xi\rangle$ ! For instance, this was crucial in [18], and the necessity to have a net with such a property motivated this construction.

### 2.8 High-dimensional phenomenon: some themes

## Weirdness of high dimension

Consider the cube $B_{\infty}^{n}$ and place a copy of $B_{2}^{n}$ centered at each vertex of $B_{\infty}^{n}$ (see the Figure below). Note that the largest ball you can place at the center of $B_{\infty}$ without intersecting any of the copies of $B_{2}^{n}$ will have radius $\sqrt{n}-1$. This might be somewhat surprising, since in many directions the ball extends out much further from the origin than the cube does, when $n$ is large.


## Central Limit Theorem and its geometric meaning

The following result is a very classical fact in Probability, see e.g. Durrett [7].
Theorem 2.27 (Central Limit Theorem). Let $X_{1}, X_{2}, \ldots$ be i.i.d. random variables with finite second moment. Let $\mu=\mathbb{E} X_{1}$ and $\sigma^{2}=\operatorname{Var}\left(X_{1}\right)$. Let $S_{n}=\sum_{1 \leq k \leq n} X_{k}$. Then,

$$
\frac{S_{n}-n \mu}{\sqrt{n \sigma^{2}}} \rightarrow \mathcal{N}(0,1)
$$

where the convergence is in distribution as $n \rightarrow \infty$.
To get a geometric interpretation of this, note that the vector $X:=\left(X_{1}, \cdots, X_{n}\right)$ is distributed uniformly over the unit cube $\mathbf{B}_{n}^{\infty}:=[-1,1]^{n}$. Given the vector $\theta=(1 / \sqrt{n}, \cdots, 1 / \sqrt{n})$ the above example says that the random variable $\langle X, \theta\rangle$ is distributed roughly as a normal random variable.

What is the geometric meaning of the density $f(t)$ of $\langle X, \theta\rangle$ ? After thinking a little, we see that

$$
f(t)=\left|\mathbf{B}_{n}^{\infty} \cap\{\langle x, \theta\rangle=t\}\right|_{n-1} .
$$

Thus $f(t)$ the $n-1$ dimensional area of the hyperplane section of the cube perpendicular to $\theta$, distance $t$ from the origin. Although sections of cubes are hard to compute exactly, as the dimension goes to infinity they resemble a normal random variable, which is a simple object.


This phenomenon appears to stem from independence. However, this fact is true not just about cubes! For any convex body there exists a direction $\theta$ (in fact, many of them) for which $\langle X, \theta\rangle$ behaves a bit like a Gaussian (in the appropriate sense). This is the content of the Central Limit Theorem for convex bodies from 2007 due to Klartag [17].

## About the thin shell type concentration

Suppose that $X$ is a random vector that is distributed uniformly over the sphere. Then with high probability, its Euclidean norm is very close to 1 when $n$ is large!

Indeed, note that

$$
\begin{aligned}
\mathbb{P}\left(|X| \geq 1-\frac{10}{n}\right)=\frac{\left|\mathbf{B}_{2}^{n} \backslash(1-10 / n) \mathbf{B}_{2}^{n}\right|}{\left|\mathbf{B}_{2}^{n}\right|} & =1-\left(1-\frac{10}{n}\right)^{n} \\
& \geq 1-e^{-10} \\
& \geq 0.99
\end{aligned}
$$

## 3 Concentration Inequalities for sums of independent random variables

### 3.1 What is a concentration inequality?

In this section, we closely follow Vershynin [53]; see also the references therein. Recall the Law of Large Numbers (see Theorem 2.4.1 in Durrett [7]).
Theorem 3.1 (LLN). Let $\left(X_{n}\right)_{n \geq 1}$ be iid random variables with $\mathbb{E}\left|X_{1}\right|<\infty$. Then,

$$
\frac{X_{1}+\ldots+X_{N}}{N} \xrightarrow{\text { a.s. }} \mathbb{E} X_{1}
$$

as $N \rightarrow \infty$.
Definition 3.2. Let $X$ be a random variable. We say that $X$ satisfies a concentration inequality if

$$
\mathbb{P}(|X-\mathbb{E} X|>t) \leq \odot(t)
$$

We usually think of $\odot(t)$ as a function that decays to 0 fast.
In some sense, a concentration inequality is a quantitative version of the Law of Large Numbers. A classical example is the Chebychev inequality:

Theorem 3.3 (Chebychev). Let $X$ be a random variable with $\mathbb{E}|X|^{2}<\infty$ and $t \geq 0$. Then,

$$
\mathbb{P}(|X-\mathbb{E} X| \geq t) \leq \frac{\operatorname{Var}(X)}{t^{2}}
$$

Proof. First, note that

$$
\mathbb{E}|X-\mathbb{E} X|^{2}=\mathbb{E} X^{2}-(\mathbb{E} X)^{2} \leq \mathbb{E} X^{2}<\infty
$$

Applying Markov's inequality to the random variable $|X-\mathbb{E} X|$, we see that

$$
\begin{aligned}
\mathbb{P}(|X-\mathbb{E} X| \geq t) & =\mathbb{P}\left(|X-\mathbb{E} X|^{2} \geq t^{2}\right) \\
& =\frac{1}{t^{2}} \operatorname{Var}(X)
\end{aligned}
$$

### 3.2 Hoeffding's inequality

However, Chebyshev's inequality provides a weak bound when it comes to random variables which are sums of independent random variables with certain properties. Below, we present the first example of a much stronger result.

Theorem 3.4 (Hoeffding's Inequality). Let $X_{1}, \ldots, X_{n}$ be independent symmetric Bernoulli random variables (taking values 1 and -1 with probability 0.5 each). Let $a=\left(a_{1}, \ldots, a_{n}\right) \in$ $\mathbb{R}^{n}$. Then, for all $t \geq 0$, we have

$$
\mathbb{P}\left(\sum_{1 \leq i \leq n} a_{i} X_{i} \geq t\right) \leq e^{-\frac{t^{2}}{2|a|^{2}}}
$$

Proof of Theorem 3.4. Without loss of generality, we may assume that $a \in \mathbb{S}^{n-1}$ (otherwise, replace $a$ by $a /|a|$ in the proof below).

Then, using the so-called Chernoff's trick, we have

$$
\begin{aligned}
\mathbb{P}\left(\sum_{i} a_{i} X_{i} \geq t\right) & =\mathbb{P}\left(e^{\lambda \sum a_{i} X_{i}} \geq e^{\lambda t}\right) \\
& \leq e^{-\lambda t} \prod_{1 \leq i \leq n} \mathbb{E}\left(e^{\lambda a_{i} X_{i}}\right)
\end{aligned}
$$

Recall that $X_{1}, \ldots, X_{n}$ are symmetric Bernoulli random variables, i.e.,

$$
X_{i}= \begin{cases}1, & \text { w.p. } 1 / 2 \\ -1, & \text { w.p. } 1 / 2\end{cases}
$$

Hence, we have

$$
\begin{aligned}
\mathbb{E}\left(e^{\lambda a_{i} X_{i}}\right) & =\frac{1}{2} e^{\lambda a_{i}}+\frac{1}{2} e^{-\lambda a_{i}} \\
& =\cosh \left(\lambda a_{i}\right)
\end{aligned}
$$

Hence, for any $\lambda>0$,

$$
\left.\begin{array}{rl}
\mathbb{P}\left(\sum_{1 \leq i \leq n} a_{i} X_{i} \geq t\right) & \leq e^{-\lambda t} \prod_{1 \leq i \leq n} \cosh \left(\lambda a_{i}\right) \\
& \leq e^{-\lambda t} \prod_{1 \leq i \leq n} e^{\lambda^{2} a_{i}^{2} / 2} \\
& =e^{-\lambda t} e^{\lambda^{2} / 2 \sum_{1 \leq i \leq n} a_{i}^{2}}  \tag{10}\\
& =e^{-\lambda t+\lambda^{2} / 2}
\end{array} \quad\left(\text { prove } \cosh (x) \leq e^{x^{2} / 2} \text { in } H W\right) ~(10) ~ 子 1, \forall \lambda \geq 0\right) ~ ? ~(|a|=1, \forall \lambda)
$$

By picking the optimal $\lambda=t$, the upper bound becomes $e^{-t^{2} / 2}$.

Remark 3.5. There is a nice geometric interpretation for this inequality, Let $X \in\{-1,1\}^{n}$, $t>0$ and $a \in \mathbb{S}^{n-1}$.

$$
\mathbb{P}(\langle X, a\rangle \geq t) \leq e^{-t^{2} / 2}
$$

and therefore, the relative number of the points of the hypercube located in a half-space orthogonal to a and distance $t$ from the origin behaves as (or, rather, better than) the Gaussian function.

Theorem 3.6 (Hoeffding's inequality for Bounded R.V.). Let $X_{1}, \ldots, X_{n}$ be bounded independent random variables, i.e., $X_{i} \in\left[m_{i}, M_{i}\right]$ for some $m_{i}, M_{i} \in \mathbb{R}$. Then, for all $\beta>0$, we have

$$
\mathbb{P}\left(\left|\sum_{i} X_{i}-\mathbb{E} X_{i}\right| \geq \beta\right) \leq 2 e^{-\frac{c \beta^{2}}{\sum_{i}\left(M_{i}-m_{i}\right)^{2}}}
$$

Proof of Theorem 3.6. Homework.
Hint:

$$
\mathbb{P}\left(\left|\sum_{i} X_{i}-\mathbb{E} X_{i}\right| \geq \beta\right)=\mathbb{P}\left(\sum_{1 \leq i \leq n} X_{i}-\mathbb{E} X_{i} \geq \beta\right)+\mathbb{P}\left(\sum_{1 \leq i \leq n} X_{i}-\mathbb{E} X_{i} \leq-\beta\right)
$$

and estimate both parts.
Remark 3.7. (Hoeffding vs. Chebyshev) Theorem 3.6 tells us that

$$
\mathbb{P}\left(\left|\sum_{1 \leq i \leq n} X_{i}-\mathbb{E} X_{i}\right| \geq t\right) \leq 2 e^{-c t^{2}}
$$

On the other hand, Chebyshev's inequality yields that

$$
\mathbb{P}\left(\left|\sum_{i} X_{i}-\mathbb{E} X_{i}\right| \geq t\right) \leq \frac{\operatorname{Var}\left(\sum_{i} X_{i}\right)}{t^{2}}=\frac{\sum_{i} \operatorname{Var}\left(X_{i}\right)}{t^{2}}
$$

Since $\frac{c_{1}}{t} \geq e^{-c t^{2}+\log 2}$ for large enough $t$ and some positive constants $c_{1}$ and $c_{2}$, we see that Hoeffding's inequality gives a tighter bound than Chebyshev's inequality.

Remark 3.8. (Hoeffding vs. Central Limit Theorem) According to CLT, we know $\frac{X_{1}+\cdots+X_{n}}{\sqrt{n}} \xrightarrow{d}$ $Z \sim \mathcal{N}(0,1)$. Let $X_{i}$ be independent Bernoulli random variables, i.e.,

$$
X_{i}= \begin{cases}1, & \text { w.p. } p \\ -1, & \text { w.p. } 1-p\end{cases}
$$

For $t>0$, observe that

$$
\begin{equation*}
\frac{1}{\sqrt{2 \pi}}\left(\frac{1}{t}-\frac{1}{t^{3}}\right) e^{-\frac{t^{2}}{2}} \leq \mathbb{P}(Z>t) \leq \frac{1}{\sqrt{2 \pi}} \frac{1}{t} e^{-\frac{t^{2}}{2}} \tag{11}
\end{equation*}
$$

The lower bound is for $t \geq 1$ is left as homework. For the upper bound, note that

$$
\begin{aligned}
P(Z>t) & =\int_{t}^{\infty} \frac{1}{\sqrt{2 \pi}} e^{-\frac{s^{2}}{2}} \mathrm{~d} s \\
& =\frac{1}{\sqrt{2 \pi}} \int_{t}^{\infty} s \cdot \frac{1}{s} \cdot e^{-\frac{s^{2}}{2}} \mathrm{~d} s \\
& \leq \frac{1}{t} \frac{1}{\sqrt{2 \pi}} \int_{t}^{\infty} s \cdot e^{-\frac{s^{2}}{2}} \mathrm{~d} s \\
& =\frac{1}{t} \frac{1}{\sqrt{2 \pi}} \int_{t^{2} / 2}^{\infty} e^{-s} \mathrm{~d} s \\
& =\frac{1}{t} \frac{1}{\sqrt{2 \pi}} e^{-\frac{t^{2}}{2}}
\end{aligned}
$$

But we cannot directly deduce Hoeffding from CLT: although

$$
\left|\mathbb{P}\left(\frac{X_{1}+\cdots+X_{n}}{\sqrt{n}} \geq t\right)-\mathbb{P}(Z \geq t)\right| \rightarrow_{n \rightarrow \infty} 0
$$

For fixed $n$, it can be polynomial in $t$. Hoeffding is not a stronger theorem than CLT, but it implies the behaviour predicted by the CLT when it comes to tail probabilities.

Remark 3.9. (HW) In fact, Hoeffding is not tight for non-symmetric Bernoulli random variables

$$
X_{i}= \begin{cases}1 & \text { w.p. } p_{i} \\ -1 & \text { w.p. } 1-p_{i}\end{cases}
$$

where $p_{i} \approx 0.99$. Chernoff's Inequality below is better.

### 3.3 Chernoff's inequality

Theorem 3.10. (Chernoff's Inequality) Let

$$
X_{i}= \begin{cases}1, & w \cdot p \cdot p_{i} \\ 0, & w \cdot p \cdot 1-p_{i}\end{cases}
$$

be independent Bernoulli random variables with parameters $p_{i}$. Denote $S_{N}=\sum_{i=1}^{N} X_{i}$ and

$$
\mathbb{E} S_{N}=\sum_{i=1}^{N} \mathbb{E} X_{i}=\sum_{i=1}^{N} p_{i}=: \mu
$$

Then for all $t>\mu$,

$$
\mathbb{P}\left(S_{N}>t\right) \leq e^{-\mu}\left(\frac{e \mu}{t}\right)^{t}
$$

Proof. For any $\lambda \in \mathbb{R}$,

$$
\begin{array}{rlr}
\mathbb{P}\left(\sum X_{i}>t\right)=\mathbb{P}\left(e^{\lambda \sum X_{i}}>e^{\lambda t}\right) & \leq e^{-\lambda t} \mathbb{E} e^{\lambda \sum X_{i}} & \\
& =e^{-\lambda t} \prod_{i=1}^{N} \mathbb{E} e^{\lambda X_{i}} & \text { (by independence) } \\
& =e^{-\lambda t} \prod_{i=1}^{N}\left(p_{i} e^{\lambda}+1-p_{i}\right) & \\
& =e^{-\lambda t} \prod_{i=1}^{N}\left(1+p_{i}\left(e^{\lambda}-1\right)\right) & \\
& \leq e^{-\lambda t} \prod_{i=1}^{N} e^{p_{i}\left(e^{\lambda}-1\right)} & \left(1+x \leq e^{x}, \forall x \in \mathbb{R}\right) \\
& =e^{-\lambda t} e^{\left(e^{\lambda}-1\right) \sum p_{i}} \\
& =e^{-\lambda t} e^{\mu\left(e^{\lambda}-1\right)} &
\end{array}
$$

By choosing the optimal $\lambda=\log \frac{t}{\mu}$ (HW: show that this choice is indeed optimal), the upper bound becomes $e^{-\mu}\left(\frac{e \mu}{t}\right)^{t}$.

Corollary 3.11 (HW, "small deviation Chernoff"). Suppose

$$
X_{i}=\left\{\begin{array}{ll}
1, & w \cdot p \cdot p_{i} \\
0, & w \cdot p \cdot 1-p_{i}
\end{array},\right.
$$

and $X_{1}, \cdots, X_{N}$ are independent. Let $S_{N}=\sum_{i=1}^{N} X_{i}$ and $\mathbb{E} S_{N}=\mu$. Then, for all $\delta \in[0,1]$,

$$
\mathbb{P}\left(\left|S_{N}-\mu\right| \geq \delta \mu\right) \leq 2 e^{-c \mu \delta^{2}}
$$

### 3.4 Application to Random Graphs

A graph is a pair $(V, E)$ where $V$ (vertices) is some finite set and $E$ (edges) is a subset of $V \times V$. Schematically, if a pair $\left(v_{i}, v_{j}\right)$ (where $\left.v_{i}, v_{j} \in V\right)$ belongs to $E$ then there is an edge between $v_{i}$ and $v_{j}$.


A random graph is a graph selected in some random way.
Definition 3.12 (Erdös-Rényi Random Graph). Given a positive integer $n$ and $p \in[0,1]$, a random graph $\mathbb{G}(n, p)$ satisfies the properties:

1. the set of vertices is $\{1, \ldots, n\}$.
2. For a given pair of vertices, there is an edge with probability $p \in(0,1)$.
3. The edge events are independent.


Recall that the degree of a vertex $v$, denoted by $\operatorname{deg}(v)$, is the number of edges that terminate at $v$.

Theorem 3.13 (The degree of each vertex of $G(n, p)$ is well concentrated around its mean). Assume that there exists $c>0$ such that for any vertex $v$ of $G \sim \mathbb{G}(n, p)$ one has

$$
\mathbb{E}[\operatorname{deg}(v)]=d \geq c \log n
$$

Then, for all $v \in G$,

$$
\mathbb{P}(\operatorname{deg}(v) \in[0.9,1.1 d]) \geq 0.9
$$

Proof of Theorem 3.13. For $i \neq j$, let

$$
X_{i j}= \begin{cases}1, & \text { w.p. } p \\ 0, & \text { w.p. } 1-p\end{cases}
$$

denote the existence of an edge between $i$-th and $j$-th vertex. Then $\operatorname{deg}(i)=\sum_{j \neq i} X_{i j}$. By Chernoff's inequality Corollary 3.11,

$$
\mathbb{P}(|\operatorname{deg}(i)-d| \geq 0.1 d)=\mathbb{P}\left(\left|\sum_{j \neq i} X_{i j}-d\right| \geq 0.1 d\right) \leq 2 e^{-c d}=0.9
$$

for right $c, c>0$.
Remark 3.14. Since the $X_{i}$ are i.i.d, we see that $\mathbb{E}[\operatorname{deg}(v)]=p(n-1)$, and thus $d \geq c \log n$ is equivalent to $p \geq \frac{c^{\prime} \log n}{n}$.

### 3.5 Sub-Gaussian Random Variables

Next, we busy ourselves with the question:
Could we prove Hoeffding's inequality in a more general situation than bounded random variables? What sort of generality could we hope for?

Note that if the inequality

$$
\mathbb{P}\left(\left|\sum_{i}^{n} a_{i} X_{i}\right| \geq t\right) \leq 2 e^{-c t^{2}}
$$

holds for all $a \in \mathbb{S}^{n-1}$, then by taking $a=e_{i}$, we see that necessarily

$$
\begin{equation*}
\mathbb{P}\left(\left|X_{i}\right| \geq t\right) \leq 2 e^{-c t^{2}} \tag{12}
\end{equation*}
$$

We will show that (12) is in fact also a sufficient condition for Hoeffding's inequality to hold when $X_{1}, \cdots, X_{n}$ are independent. As a first step, we show the following (see Proposition 2.5.2 of [53]):

Proposition 3.15. Let $X$ be a random variable. Then, the following are equivalent (the parameters $K_{1}, K_{2}, K_{3}, K_{4}, K_{5} \geq 0$ differ from one another at most by an absolute nonnegative constant multiple):

1. $\mathbb{P}(|X| \geq t) \leq 2 e^{-t^{2} / K_{1}^{2}}$, for all $t \geq 0$ and a fixed $K_{1}>0$;
2. $\|X\|_{p}=\left(\mathbb{E}|X|^{p}\right)^{1 / p} \leq K_{2} \sqrt{p}$ for all $p \geq 1$;
3. $\mathbb{E} e^{\lambda^{2} X^{2}} \leq e^{K_{3}^{2} \lambda^{2}}$ for all $\lambda \in\left[-\frac{1}{K_{3}}, \frac{1}{K_{3}}\right]$;
4. $\mathbb{E} e^{\frac{X^{2}}{K_{4}^{2}}} \leq 2$.

Moreover, if $\mathbb{E} X=0$, then conditions (1) - (4) are also equivalent to
5. $\mathbb{E} e^{\lambda X} \leq e^{K_{5}^{2} \lambda^{2}}$ for all $\lambda \in \mathbb{R}$.

Proof. Firstly show $(1) \Rightarrow(2)$. Suppose that $\mathbb{P}(|X| \geq t) \leq 2 e^{-t^{2}}$. We would like to show that in this case, $\mathbb{E}|X|^{p} \leq(C \sqrt{p})^{p}$. Indeed, by (3), and using a change of variables, we get

$$
\begin{aligned}
\mathbb{E}|X|^{p} & =\int_{0}^{\infty} \mathbb{P}\left(|X|^{p}>t\right) \mathrm{d} t \\
& =p \int_{0}^{\infty} \mathbb{P}(|X|>s) s^{p-1} \mathrm{~d} s \\
& \leq 2 p \int_{0}^{\infty} s^{p-1} e^{-s^{2}} \mathrm{~d} s=p \Gamma\left(\frac{p}{2}\right) \cdot C
\end{aligned}
$$

In the last passage we used the assumption. It remains to recall that

$$
\Gamma(m)=\int_{0}^{\infty} t^{m-1} e^{-t} d t \quad \Gamma\left(\frac{p}{2}\right)=(C \sqrt{p})^{p}
$$

Next we show $(2) \Rightarrow(3)$. Suppose $\left(\mathbb{E}|X|^{p}\right)^{\frac{1}{p}} \leq \sqrt{p}$. By Taylor's formula, $e^{a}=\sum_{k=0}^{\infty} \frac{a^{k}}{k!}$ for any $a \in[0,1]$, and therefore

$$
\begin{aligned}
\mathbb{E} e^{\lambda^{2} X^{2}} & =\mathbb{E} \sum_{k=0}^{\infty} \frac{\left(\lambda^{2} X^{2}\right)^{k}}{k!}=\sum_{k=0}^{\infty} \frac{\lambda^{2 k} \mathbb{E} X^{2 k}}{k!} \\
& \leq \sum_{k=0}^{\infty} \frac{(\lambda)^{2 k}(2 k)^{k}}{k!} \\
& \left.\leq \sum_{k=0}^{\infty}(C \lambda)^{2 k} \quad \quad \quad \text { by Stirling's formula } k!=\sqrt{2 \pi k}\left(\frac{k}{e}\right)^{k}\left(1+O\left(\frac{1}{k}\right)\right)\right) \\
& =\frac{1}{1-(C \lambda)^{2}} \\
& \leq e^{C \lambda^{2}} \quad \text { for } \quad C|\lambda| \in[0,1) .
\end{aligned}
$$

In the last passage we used the inequality $1-x \leq e^{-x}$ (see home work).
It is straightforward to show $(3) \Rightarrow(4):$ take $s=\frac{c}{K_{3}}$.
To verify $(4) \Rightarrow(1)$, note that

$$
\mathbb{P}(X \geq t)=\mathbb{P}\left(e^{x^{2}} \geq e^{t^{2}}\right) \leq \mathbb{E} e^{x^{2}} e^{-t^{2}} \leq 2 e^{-t^{2}}
$$

where the last inequality uses the assumption $\mathbb{E} e^{x^{2}} \leq 2$.
Now check $(3) \Rightarrow(5)$. Suppose $\mathbb{E} X=0$ and $\mathbb{E} e^{\lambda^{2} x^{2}} \leq e^{\lambda^{2}}, \forall \lambda \in[-1,1]$. We use the inequality $e^{x} \leq x+e^{x^{2}}, \forall x \in \mathbb{R}$ (which we leave as a home work).

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

$$
=\mathbb{E} e^{\lambda^{2} x^{2}} \quad(\text { Since we assumed } \mathbb{E} x=0)
$$

$$
(\text { if }|\lambda|<1, \text { apply }(3))
$$

Now assume $|\lambda| \geq 1$. Then we use the inequality $2 \lambda x \leq x^{2}+\lambda^{2}, \forall \lambda, x \in \mathbb{R}$, which is also left as a home work.

$$
\begin{align*}
\mathbb{E} e^{\lambda x} & \leq e^{\lambda^{2} / 2} \mathbb{E} e^{x^{2} / 2}  \tag{3}\\
& \leq e^{\lambda^{2} / 2} \sqrt{e} \leq e^{\lambda^{2}}
\end{align*}
$$

Lastly, we show $(5) \Rightarrow(1)$. Suppose $\mathbb{E} X=0$, and $\mathbb{E} e^{\lambda x} \leq e^{\lambda^{2}}$, for all $\lambda \in \mathbb{R}$. Then

$$
\mathbb{P}(X \geq t)=\mathbb{P}\left(e^{\lambda X} \geq e^{\lambda t}\right) \leq e^{-\lambda t} \mathbb{E} e^{\lambda X} \leq e^{-\lambda t} e^{\lambda^{2}}
$$

We plug the optimal value of $\lambda=\frac{t}{2}$, which gives $\mathbb{P}(X \geq t) \leq e^{-t^{2} / 4}$. Applying the same argument to the lower tail, we get

$$
\mathbb{P}(|X| \geq t)=\mathbb{P}(X \geq t)+\mathbb{P}(X \leq-t) \leq 2 \cdot e^{-t^{2} / 4}
$$

This brings us to a
Definition 3.16. (Sub-Gaussian Random Variables) If a random variable $X$ satisfies (either of the) properties (1) - (4), it is called a sub-Gaussian random variable. The sub-Gaussian norm of $X$, denoted $\|X\|_{\psi_{2}}$, is defined to be the smallest $K_{4}$ in property 4 . In other words, we define

$$
\|X\|_{\psi_{2}}=\inf \left\{t>0: \mathbb{E} \exp \left(X^{2} / t^{2}\right) \leq 2\right\}
$$

The sub-Gaussian norm is indeed a norm, which is left as a home work.
Example 3.17 (Examples of sub-Gaussian variables). The following random variables are sub-Gaussian:

- Gaussian: $X \sim N(0,1)$ is a sub-gaussian random variable with $\|X\|_{\psi_{2}} \leq C$, where $C$ is an absolute constant.
- Bernoulli: Let $X$ be symmetric Bernoulli random variable. Since $|X|=1$, it follows that $X$ is a sub-Gaussian random variable with

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

- Bounded: More generally, any bounded random variable $X$ with $|X| \leq M$ almost surely, is sub-Gaussian with

$$
\|X\|_{\psi_{2}} \leq C\|X\|_{\infty}
$$

where $C=\frac{1}{\sqrt{\ln 2}}$.

Next, we present
Example 3.18 (Examples of NON sub-Gaussian random variables). The following random variables are NON sub-Gaussian:

- Exponential: An exponential random variable is not sub-Gaussian, since

$$
\mathbb{P}(|X| \geq t) \not \leq e^{-C t^{2}}
$$

The probability density function ( $p d f$ ) of an exponential random variable is:

$$
f_{X}(x)=\left\{\begin{array}{ll}
0 & \text { if } x<0 \\
e^{-x} & \text { if } x \geq 0
\end{array} .\right.
$$

And the tail probability is given by:

$$
\mathbb{P}(X \geq t)=\int_{t}^{\infty} e^{-s} \mathrm{~d} s=e^{-t}
$$

Thus we could see that it violates the definition of sub-Gaussian random variables.

- Poisson: Poisson random variables are not sub-Gaussian.
- Cauchy: Cauchy random variables are not sub-Gaussian.


### 3.6 General Hoeffding's inequality and Khinchine's inequality

Recall the fact that a sum of independent normal random variables $X_{i}$ is normal. More precisely, if $X_{i} \sim N\left(0, \sigma_{i}^{2}\right)$ are independent then

$$
\begin{equation*}
\sum_{i=1}^{N} X_{i} \sim N\left(0, \sum_{i=1}^{N} \sigma_{i}^{2}\right) \tag{13}
\end{equation*}
$$

This property of the normal distribution extends to general sub-Gaussian distributions (see also Proposition 2.6.1 of [53]):

Proposition 3.19. (Sums of independent sub-Gaussians) Let $X_{1}, \ldots, X_{N}$ be independent, mean zero, sub-Gaussian random variables. Then $\sum_{i=1}^{N} X_{i}$ is also a sub-Gaussian random variable, and

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

Proof.

$$
\begin{aligned}
\mathbb{E} \exp \left(\lambda \sum_{i=1}^{N} X_{i}\right) & =\prod_{i=1}^{N} \mathbb{E} \exp \left(\lambda X_{i}\right) \quad \text { (by independence) } \\
& \leq \prod_{i=1}^{N} \exp \left(C^{2}\left\|\lambda X_{i}\right\|_{\psi_{2}}^{2}\right)=\exp \left(\lambda^{2} K^{2}\right)
\end{aligned}
$$

where $K^{2}:=C^{2} \sum_{i=1}^{N}\left\|X_{i}\right\|_{\psi_{2}}^{2}$. By equivalence of properties (4) and (5) in Proposition 3.15 and Definition 3.16, we see that the sum $\sum_{i=1}^{N} X_{i}$ is sub-Gaussian, and

$$
\left\|\sum_{i=1}^{N} X_{i}\right\|_{\psi_{2}} \leq C_{1} K
$$

Remark 3.20. If $X_{i} \sim N\left(0, \sigma_{i}^{2}\right)$ then $\left\|X_{i}\right\|_{\psi_{2}}^{2}=\sigma_{i}^{2}$, and an equality holds in place of the inequality above.

Proposition 3.19 is, in fact, nothing but the generalized form of the Hoeffding inequality!
Corollary 3.21. (General Hoeffding's inequality) Let $X_{1}, \ldots, X_{N}$ be independent, mean zero, sub-Gaussian random variables, and $a=\left(a_{1}, \ldots, a_{N}\right) \in \mathbb{R}^{N}$. Then, for every $t \geq 0$, we have

$$
\mathbb{P}\left(\left|\sum_{i=1}^{N} a_{i} X_{i}\right| \geq t\right) \leq 2 \exp \left(-\frac{c t^{2}}{K^{2}|a|^{2}}\right),
$$

where $K=\max _{i}\left\|X_{i}\right\|_{\psi_{2}}$.
Proof.

$$
\mathbb{P}\left(\left|\sum_{i=1}^{N} a_{i} X_{i}\right| \geq t\right) \leq 2 \exp \left(-\frac{c t^{2}}{K^{2}|a|^{2}}\right)
$$

is equivalent to (1) in Proposition 3.15. By Proposition 3.19,

$$
\left\|\sum_{1}^{N} a_{i} X_{i}\right\|_{\psi_{2}}^{2} \leq C \sum_{1}^{N}\left\|a_{i} X_{i}\right\|_{\psi_{2}}^{2}=C \sum_{1}^{N} a_{i}^{2}\left\|X_{i}\right\|_{\psi_{2}}^{2} \leq C K^{2}|a|^{2}
$$

Remark 3.22. Note that in some cases, a tighter version of Hoeffding's inequality, which is evident from the argument above, may be useful: if $X_{1}, \ldots, X_{N}$ be independent, mean zero, sub-Gaussian random variables, and $a=\left(a_{1}, \ldots, a_{N}\right) \in \mathbb{R}^{N}$, then, for every $t \geq 0$, we have

$$
\mathbb{P}\left(\left|\sum_{i=1}^{N} a_{i} X_{i}\right| \geq t\right) \leq 2 \exp \left(-\frac{c t^{2}}{\sum_{1}^{N} a_{i}^{2}\left\|X_{i}\right\|_{\psi_{2}}^{2}}\right)
$$

As an immediate corollary of Hoeffding's inequality Proposition 3.19 and the equivalence of the sub-Gaussian properties (2) and (4) in Proposition 3.15, we get:

Theorem 3.23. (Khintchine's inequality). Let $X_{1}, \ldots, X_{N}$ be independent sub-gaussian random variables with zero means and unit variances, and let $a=\left(a_{1}, \ldots, a_{N}\right) \in \mathbb{R}^{N}$. Prove that for every $p \in[2, \infty)$ we have

$$
\left(\sum_{i=1}^{N} a_{i}^{2}\right)^{1 / 2} \leq\left\|\sum_{i=1}^{N} a_{i} X_{i}\right\|_{L_{p}} \leq C K \sqrt{p}\left(\sum_{i=1}^{N} a_{i}^{2}\right)^{1 / 2}
$$

where $K=\max _{i}\left\|X_{i}\right\|_{\psi_{2}}$ and $C$ is an absolute constant.
Remark 3.24 (centering). In many results above we impose the assumption $\mathbb{E} X=0$ for convenience. But note that this assumption is often non-essential. Indeed, in the case where $X_{1}, X_{2}, \ldots, X_{n}$ are sub-Gaussian and $\mathbb{E} X_{i} \neq 0$, we may center $X_{i}$ by taking $Y_{i}=X_{i}-\mathbb{E} X_{i}$. It can be verified (see Lemma 2.6.8 of [53]) that $Y_{i}$ 's are also sub-Gaussian, and

$$
\left\|Y_{i}\right\|_{\psi_{2}}=\left\|X_{i}-\mathbb{E}\left(X_{i}\right)\right\|_{\psi_{2}} \leq\left\|X_{i}\right\|_{\phi_{2}}+\left\|\mathbb{E}\left(X_{i}\right)\right\|_{\psi_{2}} \leq C\left\|X_{i}\right\|_{\psi_{2}} .
$$

### 3.7 Sub-Exponential Random Variables

When independent random variables are not sub-Gaussian, we can no longer apply Hoeffding's inequality to bound the tail probability for their sum. However, there are important situations when this needs to be done, and is also (as we will see) possible. For example, for a standard Gaussian random vector $X \in \mathbb{R}^{n}$, that is, a random vector with i.i.d. coordinates $X_{i} \sim \mathcal{N}(0,1)$, consider the random variable $|X|$ - the length of the standard Gaussian random vector. We could study $|X|^{2}=\sum_{i=1}^{n} X_{i}^{2}$, which is the sum of i.i.d. random variables, but unfortunately we cannot apply Hoeffding's inequality: $X_{i}^{2}$ are not sub-Gaussian. Indeed, the tails of $X_{i}^{2}$ decrease only exponentially fast:

$$
\mathbb{P}\left(X_{i}^{2} \geq t\right)=\mathbb{P}\left(\left|X_{i}\right| \leq \sqrt{t}\right) \approx 2 e^{-\frac{(\sqrt{t})^{2}}{2}}=2 e^{-\frac{t}{2}}
$$

In this section, we are going to discuss such "sub-exponential" random variables, and study their properties, which, as it turns out, are also very nice.

Definition 3.25 (Sub-exponential random variable). a random variable $X$ is called subexponential if there exists $K>0$ such that

$$
\mathbb{P}(|X|>t) \leq 2 e^{-\frac{t}{K}}
$$

for all $t>0$.
Example 3.26 (Examples of sub-exponential random variables).

- Let $X$ be an exponential random variable, i.e., $X$ has probability density function

$$
f_{X}(t)= \begin{cases}0, & t<0 \\ e^{-t}, & t \geq 0\end{cases}
$$

Since $\mathbb{P}(X>s)=e^{-s}, X$ is sub-exponential.

- If $X \sim \mathcal{N}(0,1)$, then $X^{2}$ is sub-exponential.

Proposition 3.27 (Equivalent definitions of sub-exponential random variables). For a random variable $X$, the following are equivalent (here the non-negative constants $K_{i}$ differ from each other at most by an absolute constant factors):
(a) $\mathbb{P}(|X| \geq t) \leq e^{-\frac{t}{K_{1}}}$, for all $t>0$;
(b) $\left(\mathbb{E}|X|^{p}\right)^{\frac{1}{p}} \leq K_{2} p$ for all $p \geq 1$;
(c) $\mathbb{E}\left(e^{\lambda|X|}\right) \leq e^{K_{3}} \lambda$, for all $\lambda \in\left[0, \frac{1}{K_{3}}\right]$;
(d) $\mathbb{E}\left(e^{\frac{|X|}{K_{4}}}\right) \leq 2$;

If we further assume $\mathbb{E} X=0$, then properties (a)-(d) are equivalent to
(e) $\mathbb{E}\left(e^{\lambda X}\right) \leq e^{K_{5}^{2} \lambda^{2}}$ if $|\lambda| \leq \frac{1}{K_{5}}$.

The proof of the proposition is left as an exercise.
Definition 3.28. ( $\psi_{1}$-norm) We define the sub-exponential norm as $\|X\|_{\psi_{1}}:=\inf \{K>0$ : $\left.\mathbb{P}(|X| \geq t) \leq 2 e^{-\frac{t}{K}}, \forall t \geq 0\right\}$. Checking that this is a norm is left as an exercise.

Note that a sub-Gaussian random variable is always sub-exponential. But more is true:
Lemma 3.29. A random variable $X$ is sub-Gaussian if and only $X^{2}$ is sub-exponential, and in fact $\left\|X^{2}\right\|_{\psi_{1}}=\|X\|_{\psi_{2}}^{2}$.

Proof. Suppose that $X$ is sub-Gaussian. Then, for all $t>0, \mathbb{P}(|X| \geq t) \leq 2 e^{-t^{2} / K}$ for some constant $K>0$. Therefore,

$$
\mathbb{P}\left(X^{2} \geq t\right) \leq \mathbb{P}(|X| \geq \sqrt{t}) \leq 2 e^{-\frac{t}{K^{2}}}
$$

Therefore, $X^{2}$ is sub-exponential. Reversing the calculation above establishes the reverse implication. The relation between the norms also follows.

More generally,
Lemma 3.30. Suppose that $X, Y$ are sub-Gaussian random variables. Then $X Y$ is subexponential, and

$$
\begin{equation*}
\|X Y\|_{\psi_{1}} \leq\|X\|_{\psi_{2}} \cdot\|Y\|_{\psi_{2}} \tag{14}
\end{equation*}
$$

Proof. WLOG, we can assume $\|X\|_{\psi_{2}}=\|Y\|_{\psi_{2}}=1$, otherwise replace $X$ by $\tilde{X}=\frac{X}{\|X\|_{\psi_{2}}}$ and $Y$ by $\tilde{Y}=\frac{Y}{\|Y\|_{\psi_{2}}}$. Note that

$$
\|\tilde{X} \tilde{Y}\|_{\psi_{1}} \leq\|\tilde{X}\|_{\psi_{2}}\|\tilde{Y}\|_{\psi_{2}}
$$

if and only if

$$
\|X Y\|_{\psi_{1}} \leq\|X\|_{\psi_{2}}\|Y\|_{\psi_{2}}
$$

Now, using property (d) in Proposition 3.27, we have $\mathbb{E}\left(e^{X^{2}}\right) \leq 2, \mathbb{E}\left(e^{Y^{2}}\right) \leq 2$. We claim that

$$
\mathbb{E}\left(e^{|X Y|}\right) \leq 2
$$

Indeed, recall Young's inequality

$$
\frac{a^{2}+b^{2}}{2} \geq|a b|
$$

for all real numbers $a, b$, as follows from the fact that $(a-b)^{2} \geq 0$ and $(a+b)^{2} \geq 0$. Using this inequality twice, we get:

$$
\begin{aligned}
\mathbb{E}\left(e^{|X Y|}\right) & \leq \mathbb{E} e^{\frac{X^{2}+Y^{2}}{2}} \\
& =\mathbb{E}\left(e^{\frac{X^{2}}{2}} \cdot e^{\frac{Y^{2}}{2}}\right) \\
& \leq \frac{1}{2} \cdot \mathbb{E}\left(e^{X^{2}}+e^{Y^{2}}\right) \\
& =\frac{1}{2}\left(\mathbb{E} e^{X^{2}}+\mathbb{E} e^{Y^{2}}\right) \\
& \leq 2 .
\end{aligned}
$$

Remark 3.31. Note that the constant 1 in Lemma 3.30 can only be there when the definitions of $\|\cdot\|_{\psi_{1}}$ and $\|\cdot\|_{\psi_{2}}$ correspond to each other. Above, we used definitions (1) from Proposition 3.15 and (a) from Proposition 3.27.

Lemma 3.30 can also be derived using other equivalent definition of sub-Gaussian and sub-exponential random variables correspondingly, so long as they match. This is left as a home work.

Remark 3.32. Analogously to the centering trick of sub-Gaussians, if $X$ is sub-exponential, then $X-\mathbb{E} X$ is also sub-exponential, and

$$
\|X-\mathbb{E}(X)\|_{\psi_{1}} \leq C \cdot\|X\|_{\psi_{1}}
$$

for some absolute constant C. (home work)

### 3.8 Bernstein's inequality

The Bernstein inequality for sub-exponential distributions provides a bound on the probability that the sum of independent sub-exponential random variables deviates from its expected value. One may draw a parallel between Bernstein's inequality and Hoeffding's inequality, both in terms of the statement and the proof method. Historically, Berstein's work was published between 1920 and 1930, Chernoff's paper appeared in 1952, and Hoeffding's inequality was published by him in 1963.

Theorem 3.33 (Bernstein's inequality). Let $X=\left(X_{1}, X_{2}, \cdots, X_{n}\right) \in \mathbb{R}^{n}$ be a random vector, where the $X_{i}$ are independent sub-exponential random variables with $\mathbb{E} X=0$. Then, for all $t \geq 0$,

$$
\mathbb{P}\left(\left|\sum_{i=1}^{n} X_{i}\right| \geq t\right) \leq 2 \exp \left(-c \cdot \min \left(\frac{t^{2}}{\sum_{i=1}^{n}\left\|X_{i}\right\|_{\psi_{1}}}, \frac{t}{\max _{i}\left\|X_{i}\right\|_{\psi_{1}}}\right)\right)
$$

Here $c>0$ is an absolute constant.
Proof. Let $K=\max _{1 \leq i \leq n}\left\|X_{i}\right\|_{\psi_{1}}$. Then, for all $t \in \mathbb{R}$, using Markov's inequality, we see that

$$
\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{n} X_{i} \geq t\right) & =\mathbb{P}\left(\exp \left[\lambda \sum_{i=1}^{n} X_{i}\right] \geq e^{\lambda t}\right) \\
& \leq e^{-\lambda t} \cdot \mathbb{E} \exp \left(\lambda \sum_{i=1}^{n} X_{i}\right) \\
& =e^{-\lambda t} \cdot \prod_{i=1}^{n} \mathbb{E} e^{\lambda X_{i}} .
\end{aligned}
$$

Since the $X_{i}$ are centered, property (e) in Proposition 3.27 implies that

$$
\mathbb{E} e^{\lambda X_{i}} \leq e^{c \lambda^{2}\left\|X_{i}\right\|_{\psi_{1}}^{2}}
$$

for all $\lambda$ satisfying $|\lambda| \leq c / K$. So for $\lambda \leq \frac{1}{K}$,

$$
\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{n} X_{i} \geq t\right) & \leq e^{-\lambda t} \cdot \prod_{i=1}^{n} \mathbb{E} e^{\lambda X_{i}} \\
& \leq e^{-\lambda t} \cdot \prod_{i=1}^{n} e^{c \lambda^{2}\left\|X_{i}\right\|_{\psi_{1}}^{2}} \\
& =e^{-\lambda t+c \lambda^{2} \sigma^{2}}
\end{aligned}
$$

where $\sigma^{2}=\sum_{i=1}^{n}\left\|X_{i}\right\|_{\psi_{1}}^{2}$. To minimize the expression in $\lambda$ under the constraint $\lambda \leq \frac{1}{K}$, we take $\lambda=\min \left\{\frac{t}{2 c \sigma^{2}}, \frac{c}{K}\right\}$. Therefore,

$$
\mathbb{P}\left(\sum_{i=1}^{n} X_{i} \geq t\right) \leq \exp \left(\min \left\{\frac{t^{2}}{4 c \sigma^{2}}, \frac{c t}{2 K}\right\}\right)
$$

By a similar reasoning, we see that

$$
\mathbb{P}\left(\sum_{i=1}^{n} X_{i}<-t\right) \leq \exp \left(\min \left\{\frac{t^{2}}{4 c \sigma^{2}}, \frac{c t}{2 K}\right\}\right) .
$$

The theorem then follows from combining the two bounds above.
Applying the Bernstein's inequality to $a_{i} X_{i}$, we obtain the following inequality:
Corollary 3.34. Let $X=\left(X_{1}, X_{2}, \cdots, X_{n}\right) \in \mathbb{R}^{n}$ be a random vector, where the $X_{i}$ are independent and sub-exponential, $\mathbb{E} X_{i}=0$, and $a \in \mathbb{R}^{n}$. Let $K=\max _{i}\left\|X_{i}\right\|_{\psi_{1}}$. Then for all $t \geq 0$,

$$
\mathbb{P}(|\langle X, a\rangle| \geq t) \leq 2 e^{-c \cdot \min \left\{\frac{t^{2}}{K^{2}|a|^{2}}, \frac{t}{K\|a\| \infty}\right\}}
$$

If $a=\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$, then we obtain the following inequality:

## Corollary 3.35.

$$
\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^{n} X_{i}\right| \geq t\right) \leq 2 \exp \left(-c \cdot \min \left\{\frac{t^{2}}{K^{2}}, \frac{n t}{K}\right\}\right)
$$

If the $X_{i}$ are bounded by a universal constant, then the bound in Bernstein's inequality can be further strengthened. For example, see Theorem 2.8.4 in [53].

### 3.9 Concentration of the norm of a random vector with independent sub-Gaussian coordinates

A model application of Bernstein's inequality is the following inequality:
Theorem 3.36 (Concentration of the norm of a random vector with independent subGaussian coordinates). Let $X=\left(X_{1}, X_{2}, \ldots, X_{n}\right)$ be a random vector, where the $X_{i}$ are independent. Suppose that each $X_{i}$ is sub-Gaussian with the constant at most $K>0$, and $\mathbb{E} X_{i}^{2}=1$. Then,

$$
\|X-\sqrt{n}\|_{\psi_{2}} \leq C K^{2}
$$

where $C$ is an absolute constant.

Remark 3.37. Equivalently, the theorem states that

$$
\mathbb{P}(|X-\sqrt{n}| \geq t) \leq 2 e^{-\frac{c t^{2}}{K^{2}}}
$$

i.e. $X$ belongs, with high probability, to a thin shell around the sphere of radius $\sqrt{n}$. This result holds, and is interesting already when $X$ is Gaussian.

Proof. Since each $X_{i}$ is sub-Gaussian, we know that $X_{i}^{2}$ is sub-exponential and so is $X_{i}^{2}-1$. Furthermore,

$$
\begin{aligned}
\left\|X_{i}^{2}-1\right\|_{\psi_{1}} & \leq C\left\|X_{i}^{2}\right\|_{\psi_{1}} \\
& =C\left\|X_{i}\right\|_{\psi_{2}}^{2} \\
& \leq C K^{2} .
\end{aligned}
$$

Without loss of generality, we may assume $K \geq 1$. Applying Bernstein's inequality (Theorem 3.33) to $\frac{1}{n}|X|^{2}-1$, we see that for any $s>0$

$$
\mathbb{P}\left(\left.\left.\left|\frac{1}{n}\right| X\right|^{2}-1 \right\rvert\, \geq s\right) \leq e^{-\frac{c n}{K^{4}} \min \left\{s, s^{2}\right\}}
$$

where we used that $K \geq 1$. Note that for all $z \geq 0$ and $\delta>0,|z-1| \geq \delta$ implies $\left|z^{2}-1\right| \geq$ $\max \left(\delta^{2}, \delta\right)$. Therefore,

$$
\begin{aligned}
\mathbb{P}\left(\left|\frac{1}{\sqrt{n}}\right| X|-1| \geq \delta\right) & \leq \mathbb{P}\left(\left.\left.\left|\frac{1}{n}\right| X\right|^{2}-1 \right\rvert\, \geq \max \left(\delta^{2}, \delta\right)\right) \\
& \leq 2 e^{-\frac{c n}{K^{4}} \delta^{2}}
\end{aligned}
$$

Finally, taking $t=\delta \sqrt{n}$ yields

$$
\mathbb{P}(||X|-\sqrt{n}| \geq t) \leq 2 e^{-\frac{c t^{2}}{K^{4}}}
$$

### 3.10 Sub-Gaussian Random vectors

A random vector is called sub-Gaussian if each of its one-dimensional projections is subGaussian:

Definition 3.38. Let $X \in \mathbb{R}^{n}$ be a random vector. Then, we say that $X$ is sub-Gaussian if for all $\theta \in \mathbb{S}^{n-1}$, the random variable $\langle X, \theta\rangle$ is sub-Gaussian. The associated sub-Gaussian norm $\|\cdot\|_{\psi_{2}}$ is defined by

$$
\|X\|_{\psi_{2}}:=\sup _{\theta \in \mathbb{S}^{n-1}}\|\langle X, \theta\rangle\|_{\psi_{2}} .
$$

We will see that sub-Gaussian random vectors form a rich class which includes many examples. Firstly, consider the standard Gaussian random vector $X=\sim \mathcal{N}(0$, Id $)$ on $\mathbb{R}^{n}$ : the random vector whose coordinates $X_{i}$ are independent standard normal random variables. Then $X$ is sub-Gaussian since all the $\langle X, \theta\rangle \sim \mathcal{N}(0,1)$ by (13), and therefore, they are also sub-Gaussian.

Proposition 3.39 (An example of a sub-Gaussian random vector). Let $X \in \mathbb{R}^{n}$ be a random vector such that $X=\left(X_{1}, \cdots, X_{n}\right)$ is a random vector, the coordinates $X_{i}$ are independent, $\mathbb{E} X_{i}=0$ and $\left\|X_{i}\right\|_{\psi_{2}} \leq K$ for $K>0$. Then $X$ is a sub-Gaussian random vector and $\|X\|_{\psi_{2}} \leq c \cdot K$.

Proof. By Proposition 3.19, for any $\theta \in \mathbb{S}^{n-1}$,

$$
\|\langle X, \theta\rangle\|_{\psi_{2}}^{2}=\left\|\sum_{i=1}^{n} \theta_{i} X_{i}\right\|_{\psi_{2}}^{2} \leq C \sum_{i=1}^{n}\left|\theta_{i}\right|^{2}\left\|X_{i}\right\|_{\psi_{2}}^{2} \leq C \max _{1 \leq i \leq n}\left\|X_{i}\right\|_{\psi_{2}}^{2}=C K^{2}
$$

This implies the Proposition.
Therefore, the random vector uniformly distributed in $\{-1,1\}^{n}$ is sub-Gaussian. Also the uniform random vector in a cube is sub-Gaussian. More generally, any random vector whose coordinates are independent and bounded is sub-Gaussian.

Remark 3.40. What if we do not assume independence in general? Let $X=\left(X_{1}, \ldots, X_{n}\right)$ and all of its coordinates are $K-$ sub-Gaussian, but possibly dependent. Since $\|\cdot\|_{\psi_{2}}$ is a norm, we have

$$
\|\langle X, \theta\rangle\|_{\psi_{2}} \leq \sum_{i=1}^{n}\left|\theta_{i}\right| \cdot\left\|X_{i}\right\|_{\psi_{2}} \leq K \cdot \sum_{i=1}^{n}\left|\theta_{i}\right| \leq \sqrt{n} K
$$

In the last passage we used the fact that for any $\theta \in \mathbb{S}^{n-1}$,

$$
\sum_{i=1}^{n}\left|\theta_{i}\right| \leq \sqrt{n} \sqrt{\sum_{i=1}^{n} \theta_{i}^{2}}=\sqrt{n}
$$

and in fact the equality in the above is attained for $\theta=(1 / \sqrt{n}, \ldots, 1 / \sqrt{n})$. Therefore,

$$
\mathbb{P}(|\langle X, \theta\rangle| \geq t) \leq 2 e^{\frac{c t^{2}}{n K^{2}}}
$$

or in other words, $X$ is a $\sqrt{n} K$-sub-Gaussian random vector. The loss of $\sqrt{n}$ can be necessary (as an example, one may consider $X=(a, \ldots, a)$ where $a$ is a fixed random variable), but it could be problematic for applications.

In light of Remark 3.40 one may wonder if there is a sub-Gaussian random vector with the sub-Gaussian constant which does not depend on the dimension but with dependencies among the coordinates. Below we show one such example:

Proposition 3.41. The uniform distribution on the sphere $X \sim \operatorname{Unif}\left(\sqrt{n} S^{n-1}\right)$ is subGaussian and $\|X\|_{\psi_{2}} \leq C$ for some absolute constant $C>0$ that does not depend on the dimension.

Remark 3.42. The normalization $\sqrt{n} \mathbb{S}^{n-1}$ is natural because this means that $\mathbb{E} X_{i}^{2}=1$ for all $i$, so the Propositions 3.39 and 3.41 compare in a natural way.

Proof. For the standard Gaussian random vector $g \sim N(0, \mathrm{Id})$, the normalized random vector $\frac{g}{\|g\|} \sim \operatorname{Unif}\left(\mathbb{S}^{n-1}\right)$ (we live this as a home work). Therefore we can represent $X=\frac{\sqrt{n} g}{|g|}$. Therefore, using the notation $Z$ for the standard normal random variable, we get

$$
\begin{aligned}
\mathbb{P}(\langle X, \theta\rangle \geq t)= & \mathbb{P}\left(\left\langle\frac{\sqrt{n} g}{\|g\|}, \theta\right\rangle \geq t\right) \\
= & \mathbb{P}\left(\frac{\sqrt{n}}{\|g\|} \cdot Z \geq t\right) \\
= & \mathbb{P}\left(\left\{\frac{\sqrt{n}}{\|g\|} \cdot Z \geq t\right\} \cap\{\|\|g\|-\sqrt{n}\| \geq c t \sqrt{n}\}\right) \\
& \quad+\mathbb{P}\left(\left\{\frac{\sqrt{n}}{\|g\|} \cdot Z \geq t\right\} \cap\{\| \| g\|-\sqrt{n}\|<c t \sqrt{n}\}\right) \\
\leq & \mathbb{P}(Z \geq(1-c) t)+e^{-c^{\prime} t^{2}} \leq 2 e^{-c^{\prime \prime} t^{2}},
\end{aligned}
$$

where in the last passage we used Propositions 11 and 3.36.

### 3.11 Grothendieck's inequality

We now apply our results about sub-Gaussian random vectors to deduce the following fact from linear algebra, which is useful in semi-definite programming (see Vershynin [53]).

Theorem 3.43 (Grothendieck's inequality). Suppose $A$ is an $m \times n$ matrix with real entries such that for all $x, y \in\{-1,1\}^{n}$ (set of vectors with coordinates with +1 and -1 ) we have

$$
|\langle A x, y\rangle| \leq 1
$$

Then, for all vectors $u_{i}, v_{j} \in \mathbb{S}^{n-1}$ we have

$$
\left|\sum_{i, j} a_{i j}\left\langle u_{i}, v_{j}\right\rangle\right| \leq K
$$

where $K>0$ is an absolute constant.
Proof. First of all, note that the assumption can be equivalently stated as

$$
\left|\sum_{i, j} a_{i j} x_{i} y_{j}\right| \leq \max _{i}\left|x_{i}\right| \max _{j}\left|y_{j}\right|
$$

for any collection $x_{1}, \ldots, x_{m}, y_{1}, \ldots, y_{n}$. The conclusion appropriately changes to

$$
\left|\sum_{i, j} a_{i j}\left\langle u_{i}, v_{j}\right\rangle\right| \leq K \max _{i}\left|u_{i}\right| \max _{j}\left|v_{j}\right|
$$

If we did not wish for an absolute constant, we could take $K=\sum_{i, j}\left|a_{i j}\right|$. However, our goal is to obtain an absolute constant $K$ (bounded above by 288, following [53]).

Suppose $K>0$ is the optimal number for which Grothendick's theorem holds for any Hilbert space, and fix a collection of vectors $u_{i}, v_{j} \in \mathbb{S}^{n-1}$ where $\sum_{i, j=1}^{n} a_{i j}\left\langle u_{i}, v_{j}\right\rangle=K$.

Consider random variables $U_{i}$ and $V_{j}$, given by $U_{i}=\left\langle g, u_{i}\right\rangle$ and $V_{j}=\left\langle g, v_{j}\right\rangle$ where $g \sim N(0, \mathrm{Id})$. One can show that $\mathbb{E}\left(U_{i} V_{j}\right)=\left\langle u_{i}, v_{j}\right\rangle$ (which is left as a home work).

By construction, we have

$$
K=\sum_{i, j} a_{i j}\left\langle u_{i}, v_{j}\right\rangle=\sum_{i, j} a_{i j} \mathbb{E}\left(U_{i} V_{j}\right)=\mathbb{E} \sum_{i, j} a_{i j} U_{i} V_{j} .
$$

If $\left|U_{i}\right| \leq R$ and $\left|V_{j}\right| \leq R$ almost surely for some $R>0$, then

$$
K \leq R^{2} \sum_{i, j} a_{i j} \leq R^{2}
$$

where the last passage follows from the assumption of the Theorem: indeed, if we take all the coordinates of $x$ and $y$ to be 1 , we get $\sum a_{i j} \leq 1$.

Write $U_{i}=U_{i}^{+}+U_{i}^{-}$, where $U_{i}^{+}=U_{i} \cdot \mathbf{1}_{\left\{\left|U_{i}\right| \leq R\right\}}$ and $U_{i}^{-}=U_{i} \cdot \mathbf{1}_{\left\{\left|U_{i}\right|>R\right\}}$. Similarly, decompose $V_{i}=V_{i}^{+}+V_{i}^{-}$.

Note that $\left|U_{i}^{+}\right| \leq R$ almost surely. We have

$$
\begin{aligned}
K & =\mathbb{E} \sum_{i, j} a_{i j} U_{i} V_{j}=\mathbb{E}\left(\sum_{i, j} a_{i j} U_{i}^{+} V_{j}^{+}+\sum_{i, j} a_{i j} U_{i}^{+} V_{j}^{-}+\sum_{i, j} a_{i j} U_{i}^{-} V_{j}^{+}+\sum_{i, j} a_{i j} U_{i}^{-} V_{j}^{-}\right) \\
& \leq R^{2} \cdot \sum_{i, j} a_{i j}+\mathbb{E}\left(\sum_{i, j} a_{i j} U_{i}^{+} V_{j}^{-}+\sum_{i, j} a_{i j} U_{i}^{-} V_{j}^{+}+\sum_{i, j} a_{i j} U_{i}^{-} V_{j}^{-}\right) \\
& \leq R^{2}+\mathbb{E}\left(\sum_{i, j} a_{i j} U_{i}^{+} V_{j}^{-}+\sum_{i, j} a_{i j} U_{i}^{-} V_{j}^{+}+\sum_{i, j} a_{i j} U_{i}^{-} V_{j}^{-}\right) .
\end{aligned}
$$

The first can be bounded using the fact that $U_{i}^{+}, V_{i}^{+} \leq R$ by construction. For the other three terms, we use the fact that $\mathbb{E}\left|U_{i}^{-}\right|^{2} \leq \frac{4}{R^{2}}$ and $\mathbb{E}\left|V_{j}^{-}\right|^{2} \leq \frac{4}{R^{2}}$ (which we leave as a home work). Using Grothendieck's inequality with the constant $K$ (indeed, by our assumption it holds with the constant $K$ on any Hilbert space, so we are using it on the space of random variables with the scalar product $\mathbb{E}(X Y)$ ), and using also the bound $\mathbb{E}\left|V_{j}^{+}\right|^{2} \leq \mathbb{E}\left|V_{j}\right|^{2} \leq 1$, we get

$$
\mathbb{E} \sum_{i, j} a_{i j} U_{i}^{-} V_{j}^{+}=\sum_{i, j} a_{i j} \mathbb{E}\left(U_{i}^{-} V_{j}^{+}\right) \leq K\left(\mathbb{E}\left|U_{i}^{-}\right|^{2}\right)^{1 / 2}\left(\mathbb{E}\left|V_{j}^{+}\right|^{2}\right)^{1 / 2} \leq \frac{2 K}{R}
$$

The other two sums can be bounded in the same way. Putting everything together, we get

$$
K \leq R^{2}+\frac{6 K}{R}
$$

Choosing $R=12$ and solving the resulting inequality leads to $K \leq 288$.

## 4 Random Matrices

Definition 4.1. A random $N \times n$ matrix $A \in \mathbb{R}^{N \times n}$ is a matrix drawn in some random way. We shall use notation

$$
\left[\begin{array}{cccc}
a_{11} & a_{12} & \cdots & a_{1 n}  \tag{15}\\
a_{21} & a_{22} & \cdots & a_{2 n} \\
\vdots & \vdots & \ddots & \vdots \\
a_{N 1} & a_{N 2} & \cdots & a_{N n}
\end{array}\right]
$$

where the entries $a_{i j}$ are random variables.
Example 4.2. Here are some examples of random matrices:

- When $a_{i j} \sim N(0,1)$ and are independent, $A$ is sometimes called a Gaussian random matrix;
- More generally, we may consider $a_{i j}$ to be independent random variables selected according to some distributions;
- We may also consider a symmetric random matrix, by selecting the upper corner entries independently, and reinforcing the rule $a_{i j}=a_{j i}$;
- We may also select one random variable a and fill each entry with it;
- We may select some specific entries randomly independently, while other entries would have fixed values (for example zeroes);
- another example of a random matrix is a random rotation (selected uniformly from the compact set of rotations).

We think about the random $N \times n$ matrix as about the operator on $\mathbb{R}^{n}$ into $\mathbb{R}^{N}$. We will study various properties of random matrices assuming that $N$ and $n$ are very large (but not reaching the infinity limit). This is what is informally called a non-asymptotic random matrix theory. We will be using the methods of High-dimensional Probability which is only one out of the myriad of methods and theories that come useful to study random matrices. We will leave a lot of the relevant methods and questions and theories beyond the scope of this course.

For convenience we will assume throughout that $N \geq n$ since most of the properties of a matrix $A$ which we study can be easily transferred to properties of $A^{T}$.

What properties of random matrices can be studied? One example is singular values. Recall that the singular values of a matrix $A \in \mathbb{R}^{N \times n}$ are given by $\sigma_{i}(A)=\sqrt{\lambda_{i}\left(A A^{T}\right)}$, where $\lambda_{i}$ are the eigenvalues, and we suppose that $\sigma_{1}(A) \geq \cdots \geq \sigma_{n}(A)$. Recall that

$$
\begin{align*}
\sigma_{1}(A) & =\|A\|=\sup _{x \in \mathbb{S}^{n-1}}|A x| ; \\
\sigma_{n}(A) & =\inf _{x \in \mathbb{S}^{n-1}}|A x| ;  \tag{16}\\
\sigma_{i}(A) & =\sup _{\operatorname{dim}(E)=i} \inf _{x \in \mathbb{S}(E)}|A x|,
\end{align*}
$$

where $\mathbb{S}(E)$ denotes the unit sphere in the subspace $E$. See Vershynin [53] for the details.
Definition 4.3. The condition number of a matrix $A$ is defined as $\kappa(A)=\frac{\sigma_{1}(A)}{\sigma_{n}(A)}$.
The condition number measures how far a matrix $A$ is from an isometry: indeed, when $A$ is an isometry, we have $\kappa(A)=1$, and if $A$ stretches the space in some direction then $\kappa(A)$ is large. The parameter $\kappa(A)$ is directly involved in the speed of various algorithms for solving systems of linear equations, and it is important in many applications to know that a certain random matrix model has small enough condition number with high probability.

### 4.1 Norm of a sub-Gaussian random matrix

We start by studying the norm of a class of random matrices.
Theorem 4.4 (Norm of matrices with independent mean zero sub-Gaussian entries). Let $A$ be an $N \times n$ random matrix whose entries $a_{i j}$ are independent, mean zero, sub-Gaussian random variables. Suppose $N \geq n$. Then, for any $t>0$ we have

$$
\begin{equation*}
\|A\| \leq C K(\sqrt{N}+t) \tag{17}
\end{equation*}
$$

with probability at least $1-2 e^{-t^{2}}$. Here $K=\max _{i, j}\left\|a_{i j}\right\|_{\psi_{2}}$ and $C>0$ is some constant.
Proof. According to Theorem 5, let $\epsilon=\frac{1}{4}$, we can find an $\epsilon$-net $\mathcal{N} \subset \mathbb{S}^{n-1}$ and an $\epsilon$-net $\mathcal{M} \subset \mathbb{S}^{N-1}$ where

$$
\begin{equation*}
\# N \leq\left(\frac{2+\epsilon}{\epsilon}\right)^{n}, \# M \leq\left(\frac{2+\epsilon}{\epsilon}\right)^{N} \tag{18}
\end{equation*}
$$

Note that

$$
\sup _{x \in \mathbb{S}^{n-1}, y \in \mathbb{S}^{N-1}}\langle A x, y\rangle \geq \sup _{x \in \mathcal{N}, y \in \mathcal{M}}\langle A x, y\rangle,
$$

since the supremum over a larger set is larger than the supremum over a smaller set. However,

$$
\begin{equation*}
\sup _{x \in \mathbb{S}^{n-1}, y \in \mathbb{S}^{N-1}}\langle A x, y\rangle \cdot(1-2 \epsilon) \leq \sup _{x \in \mathcal{N}, y \in \mathcal{M}}\langle A x, y\rangle \tag{19}
\end{equation*}
$$

Indeed, suppose $x \in \mathbb{S}^{n-1}, \tilde{x} \in \mathcal{N}$ so that $|x-\tilde{x}| \leq \epsilon$, and $y \in \mathbb{S}^{N-1}, \tilde{y} \in \mathcal{M}$ so that $|y-\tilde{y}| \leq \epsilon$. We have

$$
\begin{align*}
|\langle A x, y\rangle-\langle A \tilde{x}, \tilde{y}\rangle| & =|\langle A x, y\rangle-\langle A x, \tilde{y}\rangle+\langle A x, \tilde{y}\rangle-\langle A \tilde{x}, \tilde{y}\rangle| \\
& \leq|\langle A x, y-\tilde{y}\rangle|+|\langle A(x-\tilde{x}), \tilde{y}\rangle| \\
& \leq 2\|A\| \cdot \epsilon  \tag{20}\\
& =2 \varepsilon \cdot \sup _{x \in \mathbb{S}^{n-1}, y \in \mathbb{S}^{N-1}}\langle A x, y\rangle,
\end{align*}
$$

which implies (19).
Going back to our aim, we have

$$
\begin{aligned}
\mathbb{P}(\|A\| \geq C \cdot(\sqrt{N}+t)) & =\mathbb{P}\left(\sup _{x \in \mathbb{S}^{n-1}, y \in \mathbb{S}^{N-1}}\langle A x, y\rangle \geq C \cdot(\sqrt{N}+t)\right) \\
& \leq \mathbb{P}\left(\sup _{x \in \mathcal{N}, y \in \mathcal{M}}\langle A x, y\rangle \geq \frac{C}{2} \cdot(\sqrt{N}+t)\right) \\
& =\mathbb{P}\left(\bigcup_{x \in \mathcal{N}, y \in \mathcal{M}}\{\langle A x, y\rangle \geq \tilde{C} \cdot(\sqrt{N}+t)\}\right) \\
& \leq \sum_{x \in \mathcal{N}, y \in \mathcal{M}} \mathbb{P}(\langle A x, y\rangle \geq \tilde{C} \cdot(\sqrt{N}+t)) \\
& \leq \sup _{x \in \mathcal{N}, y \in \mathcal{M}} \mathbb{P}(\langle A x, y\rangle \geq \tilde{C} \cdot(\sqrt{N}+t)) \cdot 9^{n+N}
\end{aligned}
$$

The first inequality holds because of the inequality (19) and plugging $1-2 \epsilon=\frac{1}{2}$.
It remains to bound the term $\mathbb{P}(\langle A x, y\rangle \geq \tilde{C} \cdot(\sqrt{N}+t))$ for fixed $x \in S^{n-1}, y \in S^{N-1}$. Note that

$$
\begin{equation*}
\mathbb{P}(\langle A x, y\rangle \geq \tilde{C} \cdot(\sqrt{N}+t))=\mathbb{P}\left(\sum_{i j} a_{i j} \cdot x_{i} y_{j} \geq \tilde{C} \cdot(\sqrt{N}+t)\right) \tag{21}
\end{equation*}
$$

We claim that $\sum_{i j} a_{i j} \cdot x_{i} y_{j}$ is sub-Gaussian. Indeed,

$$
\begin{align*}
\left|\sum_{i j} a_{i j} \cdot\left\|x_{i} y_{j}\right\|_{\psi_{2}}^{2}\right| & \leq C^{\prime} \sum_{i j}\left|x_{i} y_{j}\right|^{2} \cdot\left\|a_{i j}\right\|_{\psi_{2}}^{2} \\
& \leq C^{\prime} \sum_{i j} x_{i}^{2} y_{j}^{2}  \tag{22}\\
& =C^{\prime}\left(\sum_{i=1}^{n} x_{i}^{2}\right)\left(\sum_{j=1}^{N} y_{j}^{2}\right)=C^{\prime} .
\end{align*}
$$

The first inequality is because of Hoeffding inequality and the last equality is because $x, y$ are unit vectors. As a result, we have

$$
\begin{equation*}
\mathbb{P}(\langle A x, y\rangle \geq \tilde{C} \cdot(\sqrt{N}+t)) \leq 2 e^{-C^{\prime \prime} \cdot(\sqrt{N}+t)^{2}} \tag{23}
\end{equation*}
$$

Combining the previous results we have

$$
\begin{equation*}
\mathbb{P}(\|A\| \geq C(\sqrt{N}+t)) \leq 9^{n+N} \cdot 2 e^{-C^{\prime \prime} \cdot(\sqrt{N}+t)^{2}} \leq 2 e^{-t^{2}} \tag{24}
\end{equation*}
$$

provided that $C>0$ is selected appropriately.
Corollary 4.5. Under the assumptions of Theorem 3.36 we have $\|A\| \leq C \sqrt{N}$ with probability at least $1-e^{-N}$ (select $\left.t=\sqrt{N}\right)$. This means that $\mathbb{E}\|A\| \leq C \sqrt{N}$.

Corollary 4.6. $A$ is a $n \times n$ symmetric random matrix with the upper corner entries $a_{i j}$ being independent mean zero and $K$-sub-Gaussian. Then for for all $t \geq 0$, we have

$$
\|A\| \leq C K(\sqrt{N}+t)
$$

with probability at least $1-4 e^{-t^{2}}$.
Proof. Home work!

### 4.2 Two-sided bounds for intermediate singular values of tall enough random matrices

Recall the following notion:
Definition 4.7. $X \in \mathbb{R}^{n}$ is called an isotropic random vector in $\mathbb{R}^{n}$ if $\mathbb{E} X=0$, and for all $\theta \in S^{n-1}$,

$$
\mathbb{E}\langle X, \theta\rangle^{2}=1
$$

Note some similarity of this notion and the notion of sub-Gaussian random vectors - the vectors for which $\langle X, \theta\rangle$ is sub-Gaussian for every $\theta \in \mathbb{S}^{n-1}$.

Theorem 4.8 (Two-sided bound on sub-Gaussian matrices). Let $A$ be an $N \times n$ matrix whose rows $A_{i}$ are independent, mean zero, sub-gaussian isotropic random vectors in $\mathbb{R}^{n}$. Then for any $t \geq 0$ we have

$$
\begin{equation*}
\sqrt{N}-C K^{2}(\sqrt{n}+t) \leq \sigma_{n}(A) \leq \sigma_{1}(A) \leq \sqrt{N}+C K^{2}(\sqrt{n}+t) \tag{25}
\end{equation*}
$$

with probability at least $1-2 e^{-t^{2}}$. Here $K=\max _{i}\left\|A_{i}\right\|_{\psi_{2}}$.
Theorem 4.8 is stronger than Theorem 3.36 in the following ways:

- it is more general: for instance, it includes projections composed with matrices from Theorem 3.36, random matrices whose rows are independent and sampled from the unit sphere, and much more.
- it is a two-sided bound for all singular values (rather than just an upper bound for only $\sigma_{1}(A)$.)
- If $N \gg n$, the bound is more precise, recovering the constant 1 in front of $\sqrt{N}$. Note also that the taller the matrix, the lesser is the impact of the sub-Gaussian constant.

However, note that Theorem 4.8 only becomes applicable when $N \geq C n$ for an appropriate $C>0$ that only depends on $K>0$. This is very important to note.

We start the proof by pointing out
Claim 4.9. Let $A$ be an $N \times n$ matrix and $\delta>0$. Suppose that

$$
\begin{equation*}
\left\|A^{\top} A-I_{n}\right\| \leq \max \left(\delta, \delta^{2}\right) \tag{26}
\end{equation*}
$$

Then

$$
\begin{equation*}
(1-\delta)|x| \leq|A x| \leq(1+\delta)|x| \quad \text { for all } x \in \mathbb{R}^{n} \tag{27}
\end{equation*}
$$

The proof of this elementary fact is left as a homework.
Therefore, to establish Theorem 4.8, it is enough to prove that

$$
\left\|\frac{1}{N} A^{T} A-I_{n}\right\| \leq K^{2} \max \left(\delta, \delta^{2}\right)
$$

with high probability, where $\delta=C\left(\sqrt{\frac{n}{N}}+\frac{t}{\sqrt{N}}\right)$. The proof will be done via the epsilon-net argument. As a first step, we shall show the point-wise bound, which relies on the Bernstein inequality which we proved before.

Lemma 4.10. Let $A$ be as in Theorem 4.8. Fix $s>0$ and $x \in \mathbb{S}^{n-1}$. Then

$$
\begin{equation*}
\mathbb{P}\left\{\left.\left.\left|\frac{1}{N}\right| A x\right|^{2}-1 \right\rvert\, \geq s\right\} \leq 2 e^{-C N \min \left\{\frac{s^{2}}{K^{4}}, \frac{s}{K^{2}}\right\}} \tag{28}
\end{equation*}
$$

Proof. Consider the random vector

$$
A x=\left(\begin{array}{c}
\left\langle A^{T} e_{1}, x\right\rangle  \tag{29}\\
\left\langle A^{T} e_{2}, x\right\rangle \\
\vdots \\
\left\langle A^{T} e_{N}, x\right\rangle
\end{array}\right) .
$$

Note that the coordinates of $A x$ are independent by assumption. Also, by assumption, since $x \in \mathbb{S}^{n-1}$, we have $\left\|\left\langle A^{T} e_{i}, x\right\rangle\right\|_{\psi_{2}} \leq K$. Now, isotropicity implies that $\mathbb{E}\left\langle A^{T} e_{i}, x\right\rangle^{2}=1$. Thus, $\frac{1}{N}|A x|^{2}-1$ is a sum of mean-zero sub-exponential independent random variables with sub-Gaussian norms bounded by $\frac{K^{2}}{N}$. An application of Bernstein's inequality (Theorem 3.33) thus finishes the proof.

Proof of Theorem 4.8. Recall that there exists a $1 / 4$-net $\mathcal{N} \subset \mathbb{S}^{n-1}$ with $\# \mathcal{N} \leq 9^{n}$ such that for any $x \in \mathbb{S}^{n-1}$ there exists a $y \in \mathcal{N}$ such that $|x-y| \leq \frac{1}{4}$, and therefore,

$$
\begin{equation*}
\left.\sup _{x \in \mathbb{S}^{n-1}}\left\|\left(\frac{1}{N} A^{\top} A-I_{n}\right) x\right\| \leq\left. 2 \max _{x \in \mathcal{N}}\left|\frac{1}{N}\right| A x\right|^{2}-1 \right\rvert\, \tag{30}
\end{equation*}
$$

The above conclusion is obtained in a similar fashion to the argument in the proof of the Theorem 4.4. Therefore, using the union bound as before, we see:

$$
\begin{align*}
\mathbb{P}\left\{\left\|\frac{1}{N} A^{T} A-I_{n}\right\| \geq K^{2} \max \left(\delta, \delta^{2}\right)\right\} & \leq \mathbb{P}\left\{\max _{x \in \mathcal{N}}\left|\frac{1}{N}\|A x\|^{2}-1\right| \geq 2 K^{2} \max \left(\delta, \delta^{2}\right)\right\} \\
& \leq 9^{n} \cdot \sup _{x \in \mathbb{S}^{n-1}} \mathbb{P}\left\{\left|\frac{1}{N}\|A x\|^{2}-1\right| \geq C K^{2} \max \left(\delta, \delta^{2}\right)\right\} \\
& \leq 2 e^{-C t^{2}} \tag{31}
\end{align*}
$$

where $\delta=C\left(\sqrt{\frac{n}{N}}+\frac{t}{\sqrt{N}}\right)$. The last passage follows from plugging (31) into Lemma 4.10. The result now follows in view of the Claim 4.9.

### 4.3 Matrix Bernstein Inequality

Theorem 4.11 (Matrix Bernstein Inequality). Let $X_{1}, X_{2}, \cdots, X_{N}$ be independent meanzero $n \times n$ positive definite symmetric random matrices. Suppose $\left\|X_{i}\right\| \leq K$ for all $i=$ $1,2, \ldots, n$ almost surely (i.e. with probability 1 ). Then, for all $t \geq 0$,

$$
\begin{equation*}
\mathbb{P}\left\{\left\|\sum_{i=1}^{N} X_{i}\right\| \geq t\right\} \leq 2 n \cdot e^{-\frac{t^{2} / 2}{\sigma^{2}+K t / 3}} \tag{32}
\end{equation*}
$$

where $\sigma^{2}=\left\|\sum_{i=1}^{N} \mathbb{E} X_{i}^{2}\right\|$. Equivalently,

$$
\begin{equation*}
\mathbb{P}\left\{\left\|\sum_{i=1}^{N} X_{i}\right\| \geq t\right\} \leq 2 n \cdot e^{-C \min \left\{\frac{t^{2}}{\sigma^{2}}, \frac{t}{k}\right\}} . \tag{33}
\end{equation*}
$$

Note that $\left\|\sum_{i=1}^{N} X_{i}\right\|$ is the norm of a random matrix and $\left\|\sum_{i=1}^{N} \mathbb{E} X_{i}^{2}\right\|$ is a norm of a deterministic matrix.

Is it possible to deduce Corollary 4.6 from Theorem 4.11?
Suppose that

$$
A=\left(\begin{array}{ccccc}
a_{11} & a_{12} & a_{13} & \cdots & a_{1 n} \\
a_{21} & a_{22} & a_{23} & \cdots & a_{2 n} \\
a_{31} & a_{32} & a_{33} & \cdots & a_{3 n} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
a_{n 1} & a_{n 2} & a_{n 3} & \cdots & a_{n n}
\end{array}\right)
$$

and we decompose $A$ into $N-C n^{2}$ matrices where

$$
\begin{aligned}
X_{1} & =\left(\begin{array}{ccccc}
a_{11} & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 0
\end{array}\right), X_{2}=\left(\begin{array}{ccccc}
0 & a_{12} & 0 & \cdots & 0 \\
a_{21} & 0 & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 0
\end{array}\right), \\
X_{3} & =\left(\begin{array}{ccccc}
0 & 0 & a_{13} & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
a_{31} & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 0
\end{array}\right), \cdots
\end{aligned}
$$

Then, the statement of Corollary 4.6 follows:

$$
\begin{equation*}
\mathbb{P}\{\|A\| \leq(\sqrt{n}+t) K\} \geq 1-2 e^{-C t^{2}} \tag{34}
\end{equation*}
$$

We leave as a homework exercise to check the above consideration.
Theorem 4.12 (Lieb's Inequality). Let $H$ be an $n \times n$ symmetric matrix. Consider $f(X)=$ $\operatorname{tr}\left(e^{H+\log X}\right)$ to be a function of the symmetric positive definite matrix $X$. Then $f$ is a concave on this space:

$$
\begin{equation*}
\operatorname{tr}\left(e^{H+\log \frac{X+Y}{2}}\right) \leq \frac{1}{2}\left(\operatorname{tr}\left(e^{H+\log X}\right)+\operatorname{tr}\left(e^{H+\log Y}\right)\right) \tag{35}
\end{equation*}
$$

Recall Jensen's Inequality (which works also for matrices): if $f$ in concave, then $\mathbb{E} f(X) \leq$ $f(\mathbb{E} X)$. As a result, we get

Corollary 4.13 (Lieb's Inequality + Jensen's Inequality). Let $H$ be a fixed $n \times n$ symmetric matrix and $Z$ be a random $n \times n$ symmetric matrix. Then we have

$$
\begin{equation*}
\mathbb{E}\left(\operatorname{tr}\left(e^{H+Z}\right)\right) \leq \operatorname{tr}\left(e^{H+\log \mathbb{E} e^{Z}}\right) \tag{36}
\end{equation*}
$$

We are now ready to prove Theorem 4.11.
Proof of Theorem 4.11. Denote $S=\sum_{i=1}^{N} X_{i}$, then

$$
\begin{equation*}
\|S\|=\max _{i=1, \cdots, N}\left|\lambda_{i} S\right|=\max \left\{\lambda_{\max }(S), \lambda_{\max }(-S)\right\} \tag{37}
\end{equation*}
$$

Applying Chernoff trick, for all $\lambda \in \mathbb{R}$,

$$
\begin{align*}
\mathbb{P}\left\{\lambda_{\max }(S) \geq t\right\} & =\mathbb{P}\left\{e^{\lambda \lambda_{\max }(S)} \geq e^{\lambda t}\right\} \\
& \leq e^{-\lambda t} \mathbb{E} e^{\lambda \lambda_{\max }(S)} \tag{38}
\end{align*}
$$

Using (37), we have $P\{\|S\| \geq t\} \leq 2 n P\left\{\lambda_{\max }(S) \geq t\right\}$.

Our goal is

$$
\begin{equation*}
\mathbb{E} e^{\lambda \lambda_{\max }(S)}=\mathbb{E} \lambda_{\max }\left(e^{\lambda S}\right) \leq \mathbb{E}\left(\operatorname{tr}\left(e^{\lambda S}\right)\right) \tag{39}
\end{equation*}
$$

The first equation is by definition 1.5 of functions on matrices, and the following inequality is because for all non-negative definite matrices $A$ with eigenvalues $\lambda_{1} \geq \lambda_{2} \geq \cdots \geq \lambda_{n} \geq 0$, we have $\lambda_{1} \leq \lambda_{1}+\cdots+\lambda_{n}=\operatorname{tr}(A)$.

Applying Theorem 4.12, we get

$$
\begin{aligned}
\mathbb{E}\left(\operatorname{tr}\left(e^{\lambda S}\right)\right) & =\mathbb{E}\left(\operatorname{tr}\left(e^{\sum_{i=1}^{N-1} \lambda X_{i}+\lambda X_{N}}\right)\right) \\
& \left.\leq \mathbb{E}\left(\operatorname{tr}\left(e^{\sum_{i=1}^{N-1} \lambda X_{i}+\log \mathbb{E} e^{\lambda X_{N}}}\right)\right) \quad \text { (applying } N \text { times on } X_{N-1}, X_{N-2}, \ldots\right) \\
& \leq \operatorname{tr}\left(e^{\sum_{i=1}^{N} \log \mathbb{E} e^{\lambda X_{i}}}\right) \\
& =\operatorname{tr}\left(e^{\log \prod_{i=1}^{N} \mathbb{E} e^{\lambda X_{i}}}\right) \\
& =\operatorname{tr}\left(\prod_{i=1}^{N} \mathbb{E} e^{\lambda X_{i}}\right)
\end{aligned}
$$

The second inequality here follows from conditioning: considering the random matrix $X=$ $\lambda X_{N}$ and the fixed realization of the matrix $H=\sum_{i=1}^{N-1} \lambda X_{i}$, and then integrating the expectation.

Now, all that remains is to bound $\mathbb{E} e^{\lambda X_{i}}$.
Lemma 4.14 (Homework). Let $X$ be an $n \times n$ symmetric random matrix, $\mathbb{E} X=0$. Suppose that $\mathbb{E}\|X\| \leq K>0$ almost surely, then

$$
\begin{equation*}
\mathbb{E} e^{\lambda X} \leq e^{g(\lambda) \mathbb{E} X^{2}} \tag{40}
\end{equation*}
$$

where $g(\lambda)=\frac{\lambda^{2} / 2}{1-|\lambda| K / 3}$ for $|\lambda| \leq 3 / K$.
Remark 4.15. Note that the inequality is in the matrix sense (that is, $A \geq 0$ if for all $x \in \mathbb{R}^{n}$, one has $\langle A x, x\rangle \geq 0$ ).

We will leave the remaining proof of Theorem 4.11 to homework.

### 4.4 Non-asymptotic bounds for the smallest singular value of random matrices

### 4.4.1 General discussion about the smallest singular value of a random matrix

Theorem 4.8 implies in particular that for $N \times n$ matrices with sub-Gaussian independent entries $a_{i j}$ such that $\mathbb{E} a_{i j}=0$, and $\mathbb{E} a_{i j}^{2}=1$, with probability $\geq 1-e^{c t^{2}}$, we have

$$
\sqrt{N}-c K(\sqrt{n}-t) \leq \sigma_{n}(A) \leq \cdots \leq \sigma_{1}(A) \leq \sqrt{N}+c K(\sqrt{n}+t)
$$

Keep in mind that this is only meaningful when the matrix is tall enough, that is, $N \geq C K n$. We discussed that in order to have the upper-bound on $\sigma_{1}(A)$ with high probability for a random matrix $A$ whose entries are independent, one really needs strong assumptions such as sub-Gaussian entries, and even to guarantee the bound of order $\sqrt{N}$ on average for $\sigma_{1}(A)$, the boundedness of $\mathbb{E} a_{i j}^{4}$ is required (see e.g. Litvak, Spektor [24]). In sharp contrast, it turns out that the sub-Gaussian assumption is not necessary for bounding the smallest singular value

$$
\sigma_{n}(A)=\inf _{x \in \mathbb{S}^{n-1}}|A x|
$$

from below, and in fact, much weaker assumptions on the matrix suffice! This phenomenon was discovered in a series of works of Tikhomirov [49], [50], followed by Rebrova, Tikhomirov [32], Livshyts [25], Guedon, Litvak, Tatarko [12], Livshyts, Tikhomirov, Vershynin [26], and others.

Let us investigate how the bound of the type

$$
\begin{equation*}
\sigma_{n}(A) \geq \boldsymbol{\phi}, \tag{41}
\end{equation*}
$$

can be proven (on average or with high probability), and what might be required of the random matrix $A$ for this. For a square $n \times n$ matrix $A$, the condition $\sigma_{n}(A)=0$ is equivalent to $A$ being non-invertible. Therefore, in the case $N=n$, the condition (41) means that the matrix $A$ is "well invertible", and having such an information about a random matrix could be valuable for applications in various situations.

Let us start by making a naive attempt to use the net argument to bound $\sigma_{n}(A)$ from below. Let $\mathcal{N} \subset S^{n-1}$ be an $\varepsilon$-net of size $\left(\frac{3}{\varepsilon}\right)^{n}$. Then for all $x \in S^{n-1}$, there exits $y \in \mathcal{N}$ such that $|x-y| \leq \varepsilon$. That implies that $|A(x-y)| \leq\|A\| \cdot \varepsilon$. By triangle inequality, $|A x| \geq|A y|-\|A\| \cdot \varepsilon$. Taking infimum on both sides, we get

$$
\inf _{x \in S^{n-1}}|A x| \geq \inf _{y \in \mathcal{N}}|A y|-\|A\| \cdot \varepsilon
$$

Therefore,

$$
\begin{aligned}
\mathbb{P}\left(\sigma_{n}(A) \leq \boldsymbol{\phi}\right) & =\mathbb{P}\left(\inf _{x \in \mathbb{S}^{n-1}}|A x| \leq \boldsymbol{\phi}\right) \leq \mathbb{P}\left(\inf _{y \in \mathcal{N}}|A y| \leq \boldsymbol{\phi}+\|A\| \cdot \varepsilon\right) \\
& =\mathbb{P}\left(\bigcup_{y \in \mathcal{N}}\{|A y| \leq \boldsymbol{\infty}+\|A\| \cdot \varepsilon\}\right) \\
& \leq \# \mathcal{N} \cdot \sup _{y \in \mathcal{N}} \mathbb{P}(|A y| \leq \boldsymbol{\phi}+\|A\| \cdot \varepsilon) .
\end{aligned}
$$

In case the assumptions on the random matrix allow us to have a good control of $\|A\|$ then this is a promising start! However, as we mentioned earlier, one should not need to make such strong assumptions in principle. Thus we are going to apply Theorem 2.23 (about a more involved net argument) instead, in order to aim for a more general result. Theorem 2.23 was valid for all deterministic matrices, so we point out the following Corollary for random matrices:

Corollary 4.16. Suppose $A$ is any random matrix, with $\mathbb{E}\|A\|_{H S}^{2}<\infty$. Then for any $\varepsilon>0$, there is a net $\mathcal{N} \subset \frac{3}{2} \mathbf{B}_{2}^{n} \backslash \frac{1}{2} \mathbf{B}_{2}^{n}$ with $\# \mathcal{N} \leq(C / \varepsilon)^{n}$ such that with probability at least 0.9 we have for all $x \in \mathbb{S}^{n-1}$ there is some $y \in \mathcal{N}$ such that

$$
|A(x-y)| \leq \sqrt{10} \varepsilon \cdot \frac{\sqrt{\mathbb{E}\|A\|_{H S}^{2}}}{\sqrt{n}}
$$

Proof. Take the net $\mathcal{N}$ from Theorem 2.23. Then for all $x \in \mathbb{S}^{n-1}$ there is some $y \in \mathcal{N}$ such that

$$
|A(x-y)| \leq \varepsilon \cdot \frac{\|A\|_{H S}}{\sqrt{n}} .
$$

By Markov's inequality,

$$
\mathbb{P}\left(\|A\|_{H S}^{2} \leq 10 \mathbb{E}\|A\|_{H S}^{2}\right)=1-\mathbb{P}\left(\|A\|_{H S}^{2} \geq 10 \mathbb{E}\|A\|_{H S}^{2}\right) \geq 1-0.1=0.9
$$

This implies that with probability at least 0.9 , for all $x \in \mathbb{S}^{n-1}$ there is some $y \in \mathcal{N}$ such that

$$
|A(x-y)| \leq \varepsilon \cdot \frac{\sqrt{10 \mathbb{E}\|A\|_{H S}^{2}}}{\sqrt{n}}
$$

### 4.4.2 Small ball (or anti-concentration) assumption and the tensorization lemma

In order to apply any kind of net argument, we need to be able to upper bound $\mathbb{P}(|A y| \leq \boldsymbol{\mu})$ for a fixed $y \in \frac{3}{2} \mathbf{B}_{2}^{n} \backslash \frac{1}{2} \mathbf{B}_{2}^{n}$. To this end, we shall now have a discussion about a small ball (or anti-concentration) assumption for a random variable and for a random vector.

Definition 4.17. Suppose $\xi$ is a random variable. We say that it satisfies a small ball (or an anti-concentration) assumption if for any $z \in \mathbb{R}, \mathbb{P}(|\xi-z| \leq a) \leq b$ for some $a>0$, $b \in(0,1)$.

In other words, $\xi$ does not concentrate around some point $z$ ("small ball") with too high probability. Here are some examples:

- if $\xi$ has a bounded density $f$ (say, bounded by some constant $K$ ), then for any interval $I$ with length $2 a$,

$$
\mathbb{P}(\xi \in I)=\int_{I} f \leq K|I|=2 a K
$$

One may let $a=\frac{1}{4 K}, b=\frac{1}{2}$, and then for any $z \in \mathbb{R}, \mathbb{P}(|\xi-z| \leq a) \leq b$.

- Consider the symmetric Bernoulli

$$
\xi=\left\{\begin{array}{l}
1, \text { with probability } \frac{1}{2} \\
-1, \text { with probability } \frac{1}{2}
\end{array}\right.
$$

Then for any $z \in \mathbb{R}, \mathbb{P}(|\xi-z| \leq 0.9) \leq \frac{1}{2}$.

- If $\xi$ is a sub-Gaussian (in particular, if it is bounded), then it satisfies small ball estimate for some $a, b$ depending on $K$ (this is left as a home work).

We shall rely heavily on the following Lemma which appears in Rudelson, Vershynin [35].
Lemma 4.18 (Tensorization Lemma). Let $\xi_{1}, \xi_{2}, \ldots, \xi_{n}$ be independent non-negative random variables. Fix $\varepsilon_{0}>0, K>0$, and suppose that for any $\varepsilon \geq \varepsilon_{0}$,

$$
\mathbb{P}\left(\xi_{k} \leq \varepsilon\right) \leq K \varepsilon, k=1,2, \ldots, n .
$$

Then for some constant $C>0$

$$
\mathbb{P}\left(\sum_{j=1}^{n} \xi_{j}^{2} \leq \varepsilon^{2} n\right) \leq(C K \varepsilon)^{n} .
$$

In other words, the length of a random vector with independent coordinates satisfying small ball satisfies small ball estimate.

Proof. Assume $\varepsilon \geq \varepsilon_{0}$. Using Markov's inequality and the independence of $\xi_{j}$ 's we have:

$$
\begin{aligned}
\mathbb{P}\left(\sum_{j=1}^{n} \xi_{j}^{2} \leq \varepsilon^{2} n\right) & =\mathbb{P}\left(\exp \left\{-\frac{1}{\varepsilon^{2}} \sum_{j=1}^{n} \xi_{j}^{2}\right\} \geq e^{-n}\right) \\
& \leq e^{n} \cdot \mathbb{E} \exp \left\{-\frac{1}{\varepsilon^{2}} \sum_{j=1}^{n} \xi_{j}^{2}\right\} \\
& =e^{n} \cdot \prod_{j=1}^{n} \mathbb{E} \exp \left\{-\frac{1}{\varepsilon^{2}} \xi_{j}^{2}\right\} .
\end{aligned}
$$

Writing out the expectation and changing variables $t=e^{-s^{2}}$, we get

$$
\begin{aligned}
\mathbb{E} \exp \left\{-\frac{1}{\varepsilon^{2}} \xi_{j}^{2}\right\} & =\int_{0}^{\infty} \mathbb{P}\left(\exp \left\{-\frac{1}{\varepsilon^{2}} \xi_{j}^{2}\right\} \geq t\right) d t \\
& =\int_{0}^{\infty} 2 s e^{-s^{2}} \mathbb{P}\left(\xi_{j} \leq \varepsilon s\right) d s
\end{aligned}
$$

Since $\varepsilon \geq \varepsilon_{0}$, if $s \geq 1$, then $\varepsilon s \geq \varepsilon_{0}$ and $\mathbb{P}\left(\xi_{j} \leq \varepsilon s\right) \leq K s \varepsilon$. If $s \leq 1$, then

$$
\mathbb{P}\left(\xi_{j} \leq \varepsilon s\right) \leq \mathbb{P}\left(\xi_{j} \leq \varepsilon\right) \leq K \varepsilon
$$

So we can break the integral into two parts and get, for some constant $C_{1}>0$ :

$$
\int_{0}^{\infty} 2 s e^{-s^{2}} \mathbb{P}\left(\xi_{j} \leq \varepsilon s\right) d s \leq \int_{0}^{1} 2 K \varepsilon s e^{-s^{2}} d s+\int_{1}^{\infty} 2 K \varepsilon s^{2} s e^{-s^{2}} d s \leq C_{1} K \varepsilon
$$

Combining all of the above and letting $C=e \cdot C_{1}$, we get

$$
\mathbb{P}\left(\sum_{j=1}^{n} \xi_{j}^{2} \leq \varepsilon^{2} n\right) \leq e^{n} \cdot\left(C_{1} K \varepsilon\right)^{n}=(C K \varepsilon)^{n}
$$

Definition 4.19 (High-dimensional version of the small ball assumption). A random vector $X$ is called anti-concentrated if there exist $a>0, b \in(0,1)$ such that for any $\theta \in \mathbb{S}^{n-1}$, $\mathbb{P}(|\langle x, \theta\rangle|<a)<b$.

Claim 4.20 (Home work). Let $X$ be a random vector with independent entries. If for each $i=1,2, \ldots, n, \mathbb{P}\left(\left|X_{i}\right|<a\right)<b$, then there is some constant $C$ such that for any $\theta \in \mathbb{S}^{n-1}$,

$$
\mathbb{P}(|\langle x, \theta\rangle|<a)<C b
$$

The following Lemma is a slightly stronger version of Lemma 4.18, whose proof we leave as a home work.

Lemma 4.21. Suppose $A$ is an $N \times n$ random matrix with independent rows $A^{\top} e_{i}, i=$ $1,2, \cdots, N$, and suppose for any $\theta \in S^{n-1}$,

$$
\mathbb{P}\left(\left|\left\langle A^{\top} e_{i}, \theta\right\rangle\right|<a\right)<b
$$

Then for any $x \in \mathbb{S}^{n-1}$,

$$
\mathbb{P}(|A x| \leq c \sqrt{N}) \leq e^{-c_{1} N}
$$

where $c, c_{1}>0$ only depend on $a$ and $b$.
We point out also the following
Corollary 4.22 (Claim 4.20 combined with Lemma 4.21). Suppose $A$ is an $N \times n$ random matrix with independent entries $a_{i j}$. If for all $i, j$ and some $a>0, b \in(0,1)$,

$$
\mathbb{P}\left(\left|a_{i j}\right|<a\right)<b
$$

then for any $x \in \mathbb{S}^{n-1}$,

$$
\mathbb{P}(|A x| \leq c \sqrt{N}) \leq e^{-c_{1} N}
$$

where $c, c_{1}>0$ only depend on $a$ and $b$.

### 4.4.3 The smallest singular value of tall random matrices

We are finally ready to establish our first bound on the smallest singular value of a random matrix. So far, we will require the strong assumption that $N \geq C n$. The statement below appears as Proposition 1 in Livshyts [25].
Proposition 4.23 (Tall matrices with possibly dependent columns). Let $A$ be an $N \times n$ random matrix whose rows $A^{T} e_{i}$ are independent. Given $a>0$ and $b \in(0,1)$, suppose that for every row $A^{\top} e_{i}$ and any $\theta \in S^{n-1}$,

$$
\mathbb{P}\left(\left|\left\langle A^{\top} e_{i}, \theta\right\rangle\right|<a\right)<b
$$

Also assume that

$$
\mathbb{E}\|A\|_{H S}^{2}=\sum_{i, j} \mathbb{E} a_{i j}^{2} \leq K N n
$$

for some $K>0$. Then there exist $C_{1} \geq 1$, and $c_{2}, c_{3}>0$ which only depend on $a, b$ and $K$ such that whenever $N \geq C_{1} n$ then

$$
\mathbb{P}\left(\sigma_{n}(A) \leq c_{2} \sqrt{N}\right) \leq 0.2
$$

Consequently, $\mathbb{E} \sigma_{n}(A) \geq c_{3} \sqrt{N}$.
Proof. Since $\mathbb{E}\|A\|_{H S}^{2} \leq K N n$, applying Corollary 4.16 with $\varepsilon=1 / 4$, we get that there is a net $\mathcal{N}$ such that with probability at least 0.9 for any $x \in \mathbb{S}^{n-1}$, there is some $y \in \mathcal{N}$ such that

$$
|A(x-y)| \leq \sqrt{10} \varepsilon \frac{\sqrt{K N n}}{\sqrt{n}}=\tilde{C} \sqrt{N} .
$$

We use the net argument to upper bound the probability

$$
\begin{aligned}
\mathbb{P}\left(\sigma_{n}(A) \leq c_{2} \sqrt{N}\right) & =\mathbb{P}\left(\inf _{x \in \mathbb{S}^{n-1}}|A x| \leq c_{2} \sqrt{N}\right) \\
& \leq \mathbb{P}\left(\inf _{y \in \mathcal{N}}|A y| \leq \tilde{c_{2}} \sqrt{N}\right)+0.1 \\
& \leq \# \mathcal{N} \cdot \sup _{x \in \mathbb{S}^{n}-1} \mathbb{P}\left(|A y| \leq \tilde{c_{2}} \sqrt{N}\right)+0.1 .
\end{aligned}
$$

By Corollary 4.16 we have $\# \mathcal{N} \leq e^{C n}$ for some constant $C>0$. By Lemma 4.21, we have

$$
\sup _{x \in \mathbb{S}^{n-1}} \mathbb{P}\left(|A y| \leq \tilde{c_{2}} \sqrt{N}\right) \leq e^{-c_{1} N}
$$

Hence if $N \geq C_{1} n$ for $C_{1} \geq 1$ large enough, we are able to conclude that

$$
\mathbb{P}\left(\sigma_{n}(A) \leq c_{2} \sqrt{N}\right) \leq e^{c n} \cdot e^{-c_{1} N}+0.1 \leq 0.2
$$

Proposition 4.23 should be compared with Theorem 4.8: while it only discusses the smallest singular value, the assumptions on the matrix are a lot less demanding. Keep in mind also that Theorem 4.8 implies sharpness of the estimate in the Proposition 4.23 at least in many important situations. In both cases the assumption $N \geq C n$ is crucial. Soon we will see that the situation is quite different for square matrices, and the related results are far more complicated.

Let us formulate a corollary of Proposition 4.23 combined with Claim 4.20:
Corollary 4.24. Let $A$ be an $N \times n$ random matrix with the following properties:

1. All entries are independent;
2. The entries are Uniformly Anti-Concentrated (UAC), that is, there exist $a>0$ and $b \in(0,1)$ such that $\mathbb{P}\left(\left|a_{i j}\right| \leq a\right) \leq b$;
3. $\mathbb{E}\|A\|_{H S}^{2} \leq K \cdot N n$.

If $N \geq C n$ for an appropriate $C \geq 1$ that depends on a and $b$, then $\mathbb{E} \sigma_{n}(A) \geq C_{1} \sqrt{N}$.
What if we want this estimate to hold with high probability rather than on average, that is, would it be possible to prove

$$
\mathbb{P}\left(\sigma_{n}(A) \leq C \sqrt{N}\right) \leq e^{-G N}
$$

in place of

$$
\mathbb{P}\left(\sigma_{n}(A) \leq C \sqrt{N}\right) \leq 0.1 ?
$$

Turns out, this is possible! Analyzing the proof, one may note that the 0.1 error in the probability estimate came from the fact that our net only works with constant probability (as was obtained in Corollary 4.16 using Markov's inequality). But it is possible to construct a net with similar properties which would in fact work well with high probability! We discuss this construction below.

### 4.4.4 A net construction which works with high probability for matrices with independent columns

Theorem 4.25 (Livshyts [25]). Fix $n \in N, \epsilon \in(0,0.1)$. Then there exists a net $\mathcal{N} \subset$ $\frac{3}{2} \mathbf{B}_{2}^{n} \backslash \frac{1}{2} \mathbf{B}_{2}^{n}$ with $\# \mathcal{N} \leq\left(\frac{50}{\epsilon}\right)^{n}$ such that whenever we consider an $N \times n$ random matrix $A$ with independent columns then with probability at least $1-e^{-5 n}$, the following holds: for all $x \in \mathbb{S}^{n-1}$, there exists $y \in \mathcal{N}$ such that $|A(x-y)| \leq \frac{10}{\sqrt{n}} \cdot \sqrt{\mathbb{E}\|A\|_{H S}^{2}}$

One should compare Theorem 4.25 to Corollary 4.16. On one hand, the probability in the new result is much larger: $1-e^{-5 n} \gg 0.9$ when $n \rightarrow \infty$. On the other hand, we pay by requiring the matrix to have independent columns.

Before we discuss the proof of Theorem 4.25, we state the following improvement of Corollary 4.24:

Corollary 4.26. Let $A$ be an $N \times n$ random matrix with the following properties

1. All entries are independent
2. The entries are Uniformly Anti-Concentrated (UAC).

There exists $a>0, b \in(0,1)$ such that $P\left(\left\|a_{i j}\right\| \leq a\right) \leq b$
3. $\mathbb{E}\|A\|_{H S}^{2} \leq K \cdot N n$

This implies that if $N \geq C n$, then $\left.\mathbb{P}\left(b_{n}\right)(A) \leq C \sqrt{N}\right) \leq e^{-C n}$
Corollary 4.26 follows from Theorem 4.25 verbatim as Corollary 4.24 follows from Corollary 4.16, so we leave the details to the reader.

### 4.5 Proof of Theorem 4.25.

Recall that earlier, we constructed a net on the sphere which was a subset of the cubic lattice net. Here, inspired by Rebrova and Tikhimorov [32], we will instead cover the sphere by coordinate boxes, that is, sets of the form $\left[a_{1}, b_{1}\right] \times\left[a_{2}, b_{2}\right] \times \cdots \times\left[a_{n}, b_{n}\right]$.

Step 1: Cover the sphere by coordinate boxes and consider the associated random rounding. Up to translation, a coordinate box $P$ is determined by the lengths of its sides $\alpha_{1}, \alpha_{2}, \ldots, \alpha_{n} \in[0,1]$, or in other words, a coordinate box

$$
P=\left[0, \alpha_{1}\right] \times \ldots \times\left[0, \alpha_{n}\right]
$$

is determined by the vector $\alpha=\left(\alpha_{1}, \ldots \alpha_{n}\right)$.


Fix also a parameter $\kappa \geq 1$.
Consider

$$
\Omega_{\kappa}=\left\{\alpha \in \mathbb{R}^{n}: \alpha_{i} \in[0,1] \forall i \text { and } \prod_{i=1}^{n} \alpha_{i} \geq \kappa^{-n}\right\}
$$

This is the "set of admissible parallelepipeds", a set of coordinate boxes that fit inside the cube with volume no smaller than $\kappa$.

Lemma 4.27. Fix $\kappa \geq 1, \epsilon \in(0,0.1), \alpha \in \Omega_{\kappa}$, and consider $P_{\alpha}=\left[0, \alpha_{1}\right] \times \cdots \times\left[0, \alpha_{n}\right]$. Then, there exist $y_{1}, \ldots, y_{m} \in 1.4 B_{2}^{n}$ so that $\mathbb{S}^{n-1} \subset \bigcup_{i=1}^{m}\left(y_{i}+\frac{\epsilon}{\sqrt{n}} P\right)$ and $m \leq\left(\frac{10}{\epsilon} \kappa\right)^{n}$.

This fact is proved by considering

$$
\begin{equation*}
\mathcal{F}_{\alpha}=\left(\frac{\epsilon \alpha_{1}}{\sqrt{n}} \mathbb{Z} \times \ldots \times \frac{\epsilon \alpha_{n}}{\sqrt{n}} \mathbb{Z}\right) \cap\left(\frac{3}{2} B_{2}^{n} \backslash \frac{1}{2} B_{2}^{n}\right) \tag{42}
\end{equation*}
$$

and letting $y_{i}$ to be the centers of the lattice boxes forming $\mathcal{F}_{\alpha}$. The details are left as a home work.
Remark 4.28. Compare with the corresponding statement about covering the sphere with the cubes: the size of the net is $\kappa^{n}$ larger.

Using the net from the Lemma 4.27 and the random rounding construction, we deduce the following.
Lemma 4.29. Pick an $\alpha \in \Omega_{\kappa}$. Let $A$ be any $N \times n$ matrix. There exists a finite set $\mathcal{F}_{\alpha} \subset \frac{3}{2} \mathbf{B}_{2}^{n} \backslash \frac{1}{2} \mathbf{B}_{2}^{n}$ satisfying $\# \mathcal{F}_{\alpha} \leq\left(\frac{10 \kappa}{\epsilon}\right)^{n}$, such that for all $x \in \mathbb{S}^{n-1}$, there is $y \in \mathcal{F}_{\alpha}$ with

$$
|A(x-y)| \leq \frac{\varepsilon}{\sqrt{n}} \sqrt{\sum_{i=1}^{n} \alpha_{i}^{2}\left|A e_{i}\right|^{2}}
$$

Keep in mind here that $A e_{i}$ are columns of A and $\left|A e_{i}\right|$ are lengths of columns. Recall that we earlier proved this when all $\alpha_{1}=\alpha_{2}=\ldots=\alpha_{n}$, in which case we get $\sum_{i=1}^{n} \alpha_{i}^{2}\left|A e_{i}\right|^{2}=$ $\|A\|_{H S}^{2}$.
Proof. Consider $\mathcal{F}_{\alpha}$ as in (42). By Lemma 4.27, the size of this net is appropriate, and we have the covering

$$
\mathbb{S}^{n-1} \subset \bigcup_{i=1}^{m}\left(y_{i}+\frac{\epsilon}{\sqrt{n}} P\right)
$$

Fix $x \in \mathbb{S}^{n-1}$ and consider a box $P$ from this covering which contains $x$. Construct an " $\alpha$-associated random rounding" $\eta_{x}$, that is, a random vector $\eta_{x}$ which takes values in the vertices of $P$, such that the coordinates of $\eta_{x}$ are independent and $\mathbb{E} \eta_{x}=x$.


Then, considering the expected value with respect to the randomness of the random rounding, we get, as before,

$$
\begin{aligned}
\mathbb{E}\left|A\left(\eta_{x}-x\right)\right|^{2} & =\mathbb{E} \sum_{i=1}^{N}\left\langle A^{T} e_{i}, \eta_{x}-x\right\rangle^{2} \\
& \leq \frac{\epsilon^{2}}{n} \sum_{i=1}^{n} \alpha_{i}^{2} \cdot\left|A e_{i}\right|^{2}
\end{aligned}
$$

by the same reasoning as in the proof of Theorem 2.23 , in the equation (8). Therefore, by Markov's inequality, for all $x \in \mathbb{S}^{n-1}$, there exists $y \in \mathcal{F}_{\alpha}$ such that

$$
|A(x-y)|=\frac{\epsilon}{\sqrt{n}} \sqrt{\sum_{i=1}^{n} \alpha_{i}^{2} \cdot\left|A e_{i}\right|^{2}}
$$

Note that we can indeed guarantee that $y \in \mathcal{F}_{\alpha}$ because we know that $y$ takes values in the vertices of one of the boxes forming the net $\mathcal{F}_{\alpha}$.

This inspires us to formulate the following definition:
Definition 4.30 (A proxy for the Hilbert Schmidt Norm). Let $A$ be an $N \times n$ matrix. We define a functional that will serve as a proxy for the squared Hilbert-Schmidt norm by

$$
\mathcal{B}_{\kappa}(A):=\min _{\alpha \in \Omega_{\kappa}} \sum_{i=1}^{n} \alpha_{i}^{2} \cdot\left|A e_{i}\right|^{2}
$$

Note that $\mathcal{B}_{\kappa}(A) \leq\|A\|_{H S}^{2}=\sum_{i=1}^{n}\left|A e_{i}\right|^{2}$. Therefore, using Lemma 4.29 we can extend Theorem 2.23 as follows:

Corollary 4.31. Fix $\kappa \geq 1$ and $\epsilon \in(0,0.1)$. Let $A$ be an $N \times n$ matrix. Then there exists $a$ net $\mathcal{F} \subset \frac{3}{2} B_{2}^{n} \backslash \frac{1}{2} B_{2}^{n}$ such that for all $x \in S^{n-1}$ there exists $y \in \mathcal{F}$ with

$$
|A(x-y)| \leq \frac{\epsilon}{\sqrt{n}} \sqrt{\mathcal{B}_{\kappa}(A)}
$$

such that $\# \mathcal{F} \leq\left(\frac{10 \kappa}{\epsilon}\right)^{n}$.
Unfortunately, however, Corollary 4.31 is useless! Indeed, the value of the vector $\alpha$ which gives the minimum in $\mathcal{B}_{\kappa}(A)$ depends on the matrix $A$, and therefore the net in Corollary 4.31 depends on the matrix. In order to apply this to study random matrices, we would like to consider the union of all the nets associated with the admissible boxes in $\Omega_{\kappa}$, but there is infinitely many of them! However, it turns out that we can discretize the set $\Omega_{\kappa}$, and only consider finitely many nets which would "represent" all the possible nets. We will then be able to construct our desired net as a union of those representative nets.

Step 2: "Nets on nets" Next, we would like to switch the quantifiers in the previous statement: in place of the net that depends on the matrix, we need to have a fixed net, which serves all matrices. For that purpose we shall consider a net on the set of admissible nets.

Lemma 4.32 (nets on nets). There exist absolute constants $C, C^{\prime}, C^{\prime \prime}>0$ such that for any $\kappa>1$ and $\mu \in(0, \sqrt{n})$ there exists a collection $\mathcal{F} \subset \Omega_{\kappa^{1+\mu}}$ of cardinality

$$
\begin{equation*}
\max \left(\left(\frac{C}{\mu}\right)^{n-1},\left(C^{\prime} \mu\right)^{\frac{C^{\prime \prime} n}{\mu^{2}}}\right) \tag{43}
\end{equation*}
$$

such that for any $\alpha \in \Omega_{\kappa}$ there exists a $\beta \in \mathcal{F}$ such that for all $i=1, \ldots, n$ we have $\alpha_{i}^{2} \geq \beta_{i}^{2}$.
In particular, for any $N \times n$ matrix $A$, we have

$$
\mathcal{B}_{\kappa}(A) \geq \min _{\beta \in \mathcal{F}} \sum_{i=1}^{n} \beta_{i}^{2}\left|A e_{i}\right|^{2}
$$

Proof. Consider a transformation $T: \mathbb{R}^{n} \rightarrow \mathbb{R}^{n}$ given by

$$
T \alpha=\left(\ldots, \sqrt{\frac{\log \left|\frac{1}{\alpha_{\alpha}}\right|}{n \log \kappa}}, \ldots\right) .
$$

Denote $B=B_{2}^{n} \cap\left\{x_{i} \geq 0 \forall i=1, \ldots, n\right\}$. Then, by definition of $\Omega_{\kappa}$ we have

$$
T \Omega_{\kappa}=B
$$

and

$$
T^{-1}((1+\mu) B)=\Omega_{\kappa^{1+\mu}} .
$$

Note that this mapping is a bijection on $\Omega_{\kappa}$ as well as on $\Omega_{\kappa^{1+\mu}}$.
Consider a lattice covering $\mathcal{N}$ of the boundary of $B$ with translates of $\frac{\mu}{\sqrt{n}} B_{\infty}^{n}$. In each cube $x+\frac{\mu}{\sqrt{n}} B_{\infty}^{n}$ from this covering, pick such a vertex $v(x)$ that for all $y \in x+\frac{\mu}{\sqrt{n}} B_{\infty}^{n}$, and for all $i=1, \ldots, n$, one has $y_{i} \leq v(x)_{i}$. Define $\mathcal{S}=\{v(x): x \in \mathcal{N}\}$. Note that $\mathcal{S} \subset(1+\mu) B$, and that

$$
\# \mathcal{S}=\# \mathcal{N} \leq \min \left(\left(\frac{C}{\mu}\right)^{n-1},\left(C^{\prime} \mu\right)^{\frac{C^{\prime \prime}{ }_{n}}{\mu^{2}}}\right)
$$

Note that the power $n-1$ comes from the fact that we are covering the sphere rather than the ball.

Let $\mathcal{F}=T^{-1} \mathcal{S} \subset \Omega_{\kappa^{1+\mu}}$. For every $\alpha \in \Omega_{\kappa}$ let $a=T \alpha \in B$. Then take the $b \in \mathcal{S} \subset$ $(1+\mu) B$ such that $a_{i}^{2} \leq b_{i}^{2}$; consider $\beta \in \mathcal{F}$ defined as $\beta=T^{-1} b$. Since $T$ is coordinate-vise decreasing, we have, for all $i \in\{1, \ldots, n\}$, the inequality $\alpha_{i}^{2} \geq \beta_{i}^{2}$, as desired.


Finally, we deduce the result about a net which serves all deterministic matrices.
Theorem 4.33 (A net for deterministic matrices). Fix $n \in \mathbb{N}, \epsilon \in\left(0, \frac{1}{10}\right), \kappa \geq 1$ Consider any $\mathcal{S} \subset \mathbb{S}^{n-1}$. There exists a net $\mathcal{N} \subset \frac{3}{2} \mathbf{B}_{2}^{n} \backslash \frac{1}{2} \mathbf{B}_{2}^{n}$, such that for any $N \times n$ matrix $A$, the following holds: for every $x \in \mathcal{S}$ there exists $y \in \mathcal{N}$ such that

$$
|A(x-y)| \leq \frac{100}{\sqrt{n}} \sqrt{\mathcal{B}_{\kappa}(A)},
$$

and

$$
\# \mathcal{N} \leq\left(\frac{50 \kappa \log \kappa}{\epsilon}\right)^{n}
$$

Proof. Let $\mu=2$ and consider a net $\beta_{1}, \beta_{2}, \ldots, \beta_{m}$ on $\Omega_{\kappa}$ with $m \leq 5^{n}$, as was described in Lemma 4.32. For each $\beta_{i}$, we consider a box $P_{\beta_{i}}$, and the lattice net $\mathcal{F}_{i}$ generated by $\frac{\epsilon}{\sqrt{n}} P_{B_{i}}$.

By Lemma 4.27 applied with $\alpha=\beta_{i}$, for all $x \in \mathbb{S}^{n-1}$, there exists $y \in \mathcal{F}_{i}$ such that

$$
|A(x-y)| \leq \frac{\epsilon}{\sqrt{n}} \sqrt{\sum_{j=1}^{n}\left(\beta_{i}^{j}\right)^{2}\left|A_{j} e_{j}\right|^{2}}
$$

By Lemma 4.32,

$$
\min _{i} \sum_{j=1}^{n}\left(\beta_{i}^{j}\right)^{2}\left|A_{j} e_{j}\right|^{2} \leq \mathcal{B}_{\kappa}(A) .
$$

Therefore, if we consider our net to be $\mathcal{N}=\cup_{i=1}^{m} \mathcal{F}_{i}$, the conclusion follows.
Note that Theorem 4.33 improves upon Theorem 2.23 since $\mathcal{B}_{\kappa}(A) \leq\|A\|_{H S}^{2}$. It turns out that when the matrix $A$ is random and it has independent columns, this improvement is really crucial, since $\mathcal{B}_{\kappa}(A)$ has strong large deviation properties, while $\|A\|_{H S}^{2}$ is only larger than a multiple of its average with constant probability (as follows from Markov's inequality). We explore these strong properties in our next step.

Step 3: Large deviation of $\mathcal{B}_{\kappa}$. Note that if $y_{1}, \ldots, y_{n} \geq 0$ are fixed and we constrain $\prod_{i=1}^{n} a_{i}=C_{1}$ for some constant $C_{1}>0$, then the quantity $\sum_{i} a_{i} y_{i}$ is minimized precisely when $a_{i}=\frac{C_{2}}{y_{i}}$ for all $i$ and some constant $C_{2}>0$. This inspires our next proof.

Lemma 4.34. Let $A$ be an $N \times n$ matrix with independent columns. Fix $\kappa \geq 1$. It holds that

$$
\mathbb{P}\left\{\mathcal{B}_{\kappa}(A) \geq 10 \mathbb{E}\|A\|_{H S}^{2}\right\} \leq(C \kappa)^{-2 n}
$$

for some absolute constant $C>0$.
Proof. For all $i \in\{1, \ldots, n\}$, define the random variables $Y_{i}:=\left|A e_{i}\right|$ and

$$
a_{i}:=\sqrt{\min \left\{1, \frac{\left.\mathbb{E} Y_{i}^{2}\right)}{Y_{i}^{2}}\right\}}
$$

Let $a$ be a vector with coordinates $a_{i}$. We note that

$$
\begin{aligned}
\mathbb{P}\left\{\mathcal{B}_{\kappa}(A) \geq 10 \mathbb{E}\|A \mid\|_{H S}^{2}\right\} & =\mathbb{P}\left\{\min _{\alpha \in \Omega_{\kappa}} \sum_{i=1}^{n} \alpha_{i}^{2} Y_{i}^{2} \geq 10 \mathbb{E}\|A\|_{H S}^{2}\right\} \\
& \leq \mathbb{P}\left\{\sum_{i=1}^{n} a_{i}^{2} Y_{i}^{2} \geq 10 \mathbb{E}\|A\|_{H S}^{2}\right\}+\mathbb{P}\left\{a \notin \Omega_{\kappa}\right\} \\
& =\mathbb{P}\left\{\sum_{i=1}^{n} \min \left\{1, \frac{\left.\mathbb{E} Y_{i}^{2}\right)}{Y_{i}^{2}}\right\} Y_{i}^{2} \geq 10 \mathbb{E}\|A\|_{H S}^{2}\right\}+\mathbb{P}\left\{a \notin \Omega_{\kappa}\right\} \\
& \leq \mathbb{P}\left\{\sum_{i=1}^{n} \mathbb{E} Y_{i}^{2} \geq 10 \mathbb{E}\|A\|_{H S}^{2}\right\}+\mathbb{P}\left\{\prod_{i=1}^{n} \frac{\mathbb{E} Y_{i}^{2}}{Y_{i}^{2}} \leq \kappa^{-2 n}\right\} \\
& \leq(C \kappa)^{-2 n} \mathbb{E} \prod_{i=1}^{n} \frac{Y_{i}^{2}}{\mathbb{E} Y_{i}^{2}}=(C \kappa)^{-2 n},
\end{aligned}
$$

where the last inequality follows by noting $\sum_{i} \mathbb{E} Y_{i}^{2}=\mathbb{E}\|A\|_{H S}^{2}$ (and therefore, the first summand is zero), and by applying Markov's inequality to the second term.

$$
|A(x-y)| \leq \frac{c \sqrt{\mathbb{E}\|A\|_{H S}^{2}}}{\sqrt{n}}
$$

where $c>0$ is an absolute constant.
Proof of Theorem 4.25. The assertion follows by applying Lemma 4.34 and Theorem 4.33 with $\kappa=5$.

Remark 4.35 (Important remark). Suppose $\epsilon=0.1$ and $\kappa=3$. Then the net described in Theorem 4.25 is roughly of order $200^{n}$ whereas the usual $\epsilon$-net would be of order roughly $30^{n}$. Note furthermore that the net from Theorem 4.25 cannot be smaller than of size $2^{n}$,
because each box in the net has as many as $2^{n}$ vertices, and the net is formed by the vertices of many such boxes. Often this aspect does not matter.

However, suppose $S \subset \mathbb{R}^{n}$ is a "small set" which can be covered by $m \in$-balls, where $m \leq 1.01^{n}$. So $S$ is in this sense a lot smaller than the whole sphere (which, like we said, would require $30^{n}$ balls). Then $S$ can also be covered by parallelepipeds and one can constrict a net on $S$ with properties like the ones in Theorem 4.25. And furthermore, for any $\gamma>0$ one can have a net $\mathcal{F}$ near $S$ of size $\# \mathcal{F} \leq(1+\gamma)^{n} m$ (recall that $m$ is the number of $\epsilon$-balls covering $S$ ), and such that for any $x \in S$ one has $y \in \mathcal{F}$ so that with probability at least $1-e^{-c n}$ one has

$$
|A(x-y)| \leq \frac{C(\gamma) \epsilon \sqrt{\mathbb{E}\|A\|_{H S}^{2}}}{\sqrt{n}}
$$

Here the assumptions on $A$ are the same as in Theorem 4.25. For the full statement and the details, see Livshyts [25].

At the first glance, this may seem surprising: how could one guarantee a net made out of vertices of some boxes of size less than, say $1.2^{n}$, if there is $2^{n}$ vertices already for one box? The trick is that a lot of the vertices could be "dismissed" from this construction. This requires some further ideas and technicalities which we leave beyond the scope of this course. See the details in [25], particularly in Lemma 3.10.

### 4.6 The smallest singular value of square random matrices

The above concentration results for the smallest singular value of $A$ rely upon taking the number of rows of $A$ to be sufficiently large relative to the number of columns. Having $N \geq C n$ allowed the term

$$
\sup _{x \in \mathbb{S}^{n-1}} \mathbb{P}(|A x| \leq \boldsymbol{\varphi}) \leq e^{-c N}
$$

compensate for the $e^{C_{1} n}$ factor coming from the union bound applied over the $\epsilon$-net. The standard $\epsilon$-net argument no longer works if $A$ is a square matrix, even if one imposes additional strong assumptions on the matrix - it doesn't even work for Gaussian random matrices. However, when an idea does not work, it might be a good strategy to find a way to still apply it to make at least some progress.

In what follows we describe a foundational idea of Rudelson and Vershynin about a decomposition of the sphere. Instead of using an $\epsilon$-net to cover the whole sphere $\mathbb{S}^{n-1}$, we can construct an $\epsilon$-net of prescribed size $e^{c n}$ for a "small but well spread" subset $S \subseteq \mathbb{S}^{n-1}$. "Small" in this context means that the set has a small covering number, and "well spread" means that its complement is in some sense predictable and controllable using different ideas. The smallest singular value can then be analyzed by noticing that

$$
\mathbb{P}\left(\inf _{x \in \mathbb{S}^{n-1}}|A x| \leq \boldsymbol{\phi}\right) \leq \mathbb{P}\left(\inf _{x \in S}|A x| \leq \boldsymbol{\phi}\right)+\mathbb{P}\left(\inf _{x \in S^{c}}|A x| \leq \boldsymbol{\phi}\right)
$$

and considering the infima $\inf _{x \in S}|A x|$ and $\inf _{x \in \mathbb{S}^{n-1}}|A x|$ separately.

Definition 4.36. For all $\delta>0$ define the set

$$
\operatorname{Sparse}(\delta):=\left\{x \in \mathbb{S}^{n-1}: \#\left\{i: x_{i}=0\right\} \geq \delta n\right\}
$$

In other words, the set $\operatorname{Sparse}(\delta)$ is the intersection of $\mathbb{S}^{n-1}$ with the union of all coordinate sub-spaces of dimension $\delta n$.


The following Lemma is a crucial fact about sparse vectors: the set of sparse vectors can be covered by a relatively small number of $\epsilon$-balls.

Lemma 4.37. For all $\epsilon>0$ and $\delta \in(0,1 / 2)$ one has

$$
\operatorname{Sparse}(\delta) \subset \bigcup_{i=1}^{m}\left(y_{i}+\epsilon \mathbf{B}_{2}^{n}\right)
$$

where $m \leq \epsilon^{-c_{1} \delta \log \frac{1}{\delta} n+c_{2}}$ and $c_{1}, c_{2}>0$ are absolute constants.
Proof. Suppose for simplicity that $\delta n$ is an integer (the proof works along similar lines either way). There are $\binom{n}{\delta n}$ sub-spaces in $\mathbb{R}^{n}$ of dimension $\delta n$. Note, using Stirling's formula:

$$
\binom{n}{\delta n}=\frac{n!}{(\delta n)!(n-\delta n)!}=C \frac{n^{n} e^{-n}}{(\delta n)^{\delta n} e^{-\delta n}\left((n-\delta n)^{n-\delta n} e^{-n+\delta n}\right.} \leq e^{n \delta \log \frac{1}{\delta}+c}
$$

where in the last passage one may use elementary calculus, see e.g. [58].
We deduce that

$$
m \leq\left(\frac{3}{\epsilon}\right)^{\delta n}\binom{n}{\delta n} \leq \epsilon^{-\delta n} e^{H(\delta) n}=\epsilon^{-\delta n} e^{-(\delta \log \delta+(1-\delta) \log (1-\delta)) n} \leq \epsilon^{-\delta n} e^{2 \delta \log \left(\frac{1}{\delta}\right) n+c}
$$

which yields the conclusion.

### 4.6.1 Rudelson-Vershynin decomposition of the sphere

As we discussed earlier, we want to decompose the unit sphere into two disjoint subsets, one of which would have a small covering. Sparse vectors are a good candidate for this first set. However, that is not enough: the set of sparse vectors has measure zero, and removing it does not achieve much... So instead we will do a net argument on the set of vectors which are close to sparse vectors. These are called compressible vectors. Here is the formal

Definition 4.38 (the Rudelson-Vershynin decomposition of the sphere). For all $\delta>0$ and $\rho>0$, define the compressible vectors to be the set

$$
\begin{aligned}
\operatorname{Comp}(\delta, \rho) & =\left\{x \in \mathbb{S}^{n-1}: d(x, \operatorname{Sparse}(\delta)) \leq \rho\right\} \\
& =\left\{x \in \mathbb{S}^{n-1}: \exists y \in \operatorname{Sparse}(\delta) \text { s.t. }|x-y| \leq \rho\right\},
\end{aligned}
$$

and the incompressible vectors to be their complement

$$
\operatorname{Incomp}(\delta, \rho)=\mathbb{S}^{n-1} \backslash \operatorname{Comp}(\delta, \rho)
$$



Lemma 4.37 implies:
Lemma 4.39. For all $c \in[0,1]$ and $\epsilon \in[0,1]$ there exist $\delta>0$ and $\rho>0$ such that

$$
\operatorname{Comp}(\delta, \rho) \subset \bigcup_{i=1}^{m}\left(y+\epsilon \mathbf{B}_{2}^{n}\right)
$$

where $m \leq \frac{\rho}{\epsilon} e^{c n} \leq e^{c_{1} n}$, with $c$ depending on $\epsilon$ and $\delta$, and $c_{1}$ depending on $\epsilon, \rho, \delta$.
The following claim follows from Lemma 4.39 and Remark 4.35:
Claim 4.40. For any $c \in[0,1]$ there exist $\epsilon \in\left(0, \frac{1}{4}\right), \delta, \rho \in(0,1)$, and an $\epsilon$-net $\mathcal{N} \subset$ $\frac{3}{2} \mathbf{B}_{2}^{n} \backslash \frac{1}{2} \mathbf{B}_{2}^{n}$ with $\# \mathcal{N} \leq e^{c n}$, such that the following holds. Let $A$ be an $N \times n$ random
matrix with independent columns. With probability at least $1-e^{c n}$, one has that for all $x \in \operatorname{Comp}(\delta, \rho)$, there exists $y \in \mathcal{N}$ such that

$$
\|A(x-y)\| \leq \frac{C \sqrt{\mathbb{E}\|A\|_{H S}^{2}}}{\sqrt{n}}
$$

Here $C>0$ depends only on $C, \delta, \rho$.
The Claim above will be useful for us when we estimate the infimum of $|A x|$ over the set of compressible vectors. But for the time being, let us discuss how the smallest singular value of square random matrices actually behaves, and state some relevant results.

### 4.6.2 Survey of results regarding the smallest singular value of square random matrices

Theorem 4.41 (Rudelson-Vershynin). Let $A$ be an $n \times n$ random matrix with entries that are iid, $K$-sub-Gaussian, mean-zero, and variance-one. For all $\epsilon>0$ one has

$$
\mathbb{P}\left\{\sigma_{n}(A) \leq \frac{\epsilon}{\sqrt{n}}\right\} \leq C \epsilon+e^{-c_{1} n}
$$

where $C$ and $c_{1}$ are positive constants depending only on $K$.
In the analogous setting for tall matrices $N \geq C n$, we established that $\mathbb{E} \sigma_{n}(A) \geq \widetilde{c} \sqrt{N}$, which tends to infinity. Theorem 4.41 states something quite different for square matrices: the smallest singular value turns out to be of order $\frac{1}{\sqrt{n}}$ on average, and in fact this estimate is sharp (up to an absolute constant), see Rudelson, Vershynin [36], Tatarko [44]. Let us consider two key examples which demonstrate the sharpness also of the probability estimate in Theorem 4.41.

Example 4.42. Suppose the entries of $A$ are i.i.d. distributed as standard Gaussian random variables. Szarek [43] and Edelman [8] proved that for all $\epsilon>0$ one has

$$
\mathbb{P}\left\{\sigma_{n}(A) \leq \frac{\epsilon}{\sqrt{n}}\right\} \leq \epsilon
$$

Note that this implies $\mathbb{P}\left\{\sigma_{n}(A)=0\right\}=0$, and this phenomenon would more generally be true for any random matrix whose entries have continuous distribution, even without the mean zero assumption, see Tikhomirov [51].

Example 4.43. Suppose the entries $a_{i j}$ are i.i.d. symmetric Bernoulli random variables with parameter $1 / 2$, i.e. $a_{i j} \sim \operatorname{Unif}\{-1,1\}$. What can be said about the invertibility of $A$ ? One has for instance

$$
\begin{aligned}
\mathbb{P}\left\{\sigma_{n}(A)=0\right\} & \geq \mathbb{P}\{\text { two rows are the same or two columns are the same }\} \\
& \geq\left(n^{2}+o(1)\right) 2^{-n} .
\end{aligned}
$$

A conjecture of Erdős stated that this bound is sharp up to a polynomial error. This conjecture was essentially resolved by Tikhomirov, who proved that

$$
\mathbb{P}\left\{\sigma_{n}(A)=0\right\} \leq(2+o(1))^{-n}
$$

This shows that for random matrices with independent discrete entries, one generally expects to have

$$
\mathbb{P}\left\{\sigma_{n}(A)=0\right\}=e^{-n}
$$

Combining the terms in Examples 1 and 2 shows that the probability estimate in Theorem 4.41 is natural.

Following the result from Theorem 4.41, Rebrova and Tikhomirov were able to relax the sub-Gaussian assumption, instead considering uniformly anti-concentrated random variables.

Theorem 4.44 (Rebrova-Tikhomirov [32]). Let A be a random matrix, whose entries $a_{i j}$ are zero-mean, unit variance, and i.i.d. If $a_{i j}$ are uniformly anti-concentrated (UAC), i.e., $\mathbb{P}\left(a_{i j}<a\right)<b$ for fixed $a>0, b \in(0,1)$, then for every $\varepsilon>0$,

$$
\mathbb{P}\left(\sigma_{n}(A) \leq \frac{\varepsilon}{\sqrt{n}}\right) \leq c \varepsilon+e^{-c_{1} n}
$$

where $c, c_{1}$ are constants that depend on a and $b$.
The proof technique for Theorem 4.44 uses a more clever $\varepsilon$-net argument than Theorem 4.41, which depends on the norm of the matrix. In particular, every assumption including $\mathbb{E}\left(a_{i j}\right)=0, \mathbb{E}\left(a_{i j}^{2}\right)=1$, and UAC are used in the net construction. A further generalization of the result investigates removing the assumption that the entries are all identically distributed, zero-mean, and unit variance.
Theorem 4.45 (Livshyts [25]). Let $A$ be a random matrix, whose entries $a_{i j}$ are independent. If $\mathbb{E}\|A\|_{H S}^{2} \leq K n^{2}$, and $a_{i j}$ are uniformly anti-concentrated (UAC), i.e., $\mathbb{P}\left(a_{i j}<a\right)<b$ for fixed $a>0, b \in(0,1)$, then for every $\varepsilon>0$,

$$
\begin{equation*}
\mathbb{P}\left(\sigma_{n}(A) \leq \frac{\varepsilon}{n}\right) \leq c \varepsilon+\frac{c_{1}}{\sqrt{n}}, \tag{44}
\end{equation*}
$$

where $c, c_{1}$ are constants that depend on a and $b$. Moreover, if the rows of $A$ are i.i.d., then

$$
\begin{equation*}
\mathbb{P}\left(\sigma_{n}(A) \leq \frac{\varepsilon}{\sqrt{n}}\right) \leq c \varepsilon+e^{-c_{1} n} \tag{45}
\end{equation*}
$$

where $c, c_{1}$ are constants that depend on a and $b$.
Remark 4.46. The polynomial bound from (44), while easier to prove, is a strictly weaker bound than the exponential bound of (45), which recovers the bound from Theorem 4.44 for the i.i.d. case. Recovering the exponential bound (45) requires an additional tool called the "Least Common Denominator" (LCD) of vectors. If one only cares about the limit as $n \rightarrow \infty$, this is unimportant, however, in the case of square matrices, we may care about the conservatism of this bound. The proof of the polynomial bound (44), is worked out in the next subsection.

In a future result, Livshyts, Tikhomirov, and Vershynin were able to prove the same exponential bound without the i.i.d. row assumption.
Theorem 4.47 (Livshyts, Tikhomirov, Vershynin [26]). Let A be a random matrix, whose entries $a_{i j}$ are independent. If $\mathbb{E}\|A\|_{H S}^{2} \leq K n^{2}$, and $a_{i j}$ are uniformly anti-concentrated $(U A C)$, i.e., $\mathbb{P}\left(a_{i j}<a\right)<b$ for fixed $a>0, b \in(0,1)$, then for every $\varepsilon>0$,

$$
\mathbb{P}\left(\sigma_{n}(A) \leq \frac{\varepsilon}{n}\right) \leq c \varepsilon+e^{-c_{1} n}
$$

where $c, c_{1}$ are constants that depend on a and $b$.
The proof of Theorem 4.47 requires the use of the "randomized LCD", similar to the LCD used in the proof of Theorem 4.45.

Arbitrary aspect ratios We briefly state some results for matrices of arbitrary aspect ratios. Consider the random matrix $A$, of dimension $N \times n$, with i.i.d., zero mean, and unit variance entries $a_{i j}$. Consider the case $N \geq n$, but possibly not $N \gg n$.
Theorem 4.48 (Rudelson, Vershynin [37]). If $a_{i j}$ are sub-Gaussian, then for every $\varepsilon>0$,

$$
\mathbb{P}\left(\sigma_{n}(A) \leq \varepsilon(\sqrt{N+1}-\sqrt{n})\right) \leq(c \varepsilon)^{N-n+1}+e^{-c_{1} n}
$$

for constants $c, c_{1}$.
Theorem 4.49 (Livshyts). If $a_{i j}$ are $U A C$, then for every $\varepsilon>0$,

$$
\mathbb{P}\left(\sigma_{n}(A) \leq \varepsilon(\sqrt{N+1}-\sqrt{n})\right) \leq\left(c \varepsilon \log \frac{1}{\varepsilon}\right)^{N-n+1}+e^{-c_{1} n}
$$

for constants $c, c_{1}$ dependent on $a$ and $b$.

### 4.7 Proof of Theorem 4.45 Part 1

Recall the statement of the theorem. Let $A$ be a random matrix, whose entries $a_{i j}$ are independent. If $\mathbb{E}\|A\|_{H S}^{2} \leq K n^{2}$, and $a_{i j}$ are uniformly anti-concentrated (UAC), i.e., $\mathbb{P}\left(a_{i j}<a\right)<b$ for fixed $a>0, b \in(0,1)$, then for every $\varepsilon>0$,

$$
\mathbb{P}\left(\sigma_{n}(A) \leq \frac{\varepsilon}{n}\right) \leq c \varepsilon+\frac{c_{1}}{\sqrt{n}}
$$

where $c, c_{1}$ are constants that depend on $a$ and $b$.
To prove this result, we will need to use additional tools to separately handle the compressible and incompressible vectors from the Rudelson-Vershynin decomposition of the sphere from Definition 4.38. Recall the following definitions of sparse vectors, compressible vectors (close to sparse), and incompressible vectors (far from sparse). For $\delta, \rho>0$,

$$
\begin{aligned}
\operatorname{Sparse}(\delta) & =\left\{x \in \mathbb{S}^{n-1}: \#\left\{i: x_{i}=0\right\} \geq \delta n\right\} \\
\operatorname{Comp}(\delta, \rho) & =\left\{x \in \mathbb{S}^{n-1}: \exists y \in \operatorname{Sparse}(\delta) \text { s.t. }|x-y| \leq \rho\right\} \\
\operatorname{Incomp}(\delta, \rho) & =\mathbb{S}^{n-1} \backslash \operatorname{Comp}(\delta, \rho)
\end{aligned}
$$

### 4.7.1 Compressible Vectors

After taking the Rudelson-Vershynin decomposition of the sphere, we first consider the compressible vectors in $\operatorname{Comp}(\delta, \rho)$. The following Lemma bounds the behavior of the compressible vectors.

Lemma 4.50 (Compressible vectors). Let $A$ be a random matrix, whose entries $a_{i j}$ are independent. If $\mathbb{E}\|A\|_{H S}^{2} \leq K n^{2}$, and $a_{i j}$ are uniformly anti-concentrated (UAC), i.e., $\mathbb{P}\left(a_{i j}<a\right)<b$ for fixed $a>0, b \in(0,1)$, then for every $\varepsilon>0$,

$$
\mathbb{P}\left(\inf _{x \in \operatorname{Comp}(\delta, \rho)}|A x| \leq \frac{\varepsilon}{\sqrt{n}}\right) \leq \mathbb{P}\left(\inf _{x \in \operatorname{Comp}(\delta, \rho)} \leq c \sqrt{n}\right) \leq e^{-\widetilde{c} n}
$$

Proof. Let $\mathcal{N}$ be a net from Claim 4.40, such that $\mathcal{N} \subset \frac{3}{2} B_{2}^{n} \backslash \frac{1}{2} B_{2}^{n}, \# \mathcal{N} \leq e^{c_{1} n}$, and with probability $e^{-c_{1} n}$, there exists a $y \in \mathcal{N}$ such that

$$
|A(x-y)| \leq \frac{C \mathbb{E}\|A\|_{H S}^{2}}{\sqrt{n}} \leq \widetilde{C} \sqrt{n}
$$

since $\mathbb{E}\|A\|_{H S}^{2} \leq K n^{2}$, with $\widetilde{C}=C \sqrt{K}$. This implies that

$$
\mathbb{P}\left(\inf _{x \in \operatorname{Comp}(\delta, \rho)}|A x| \leq \frac{\varepsilon}{\sqrt{n}}\right) \leq \mathbb{P}\left(\inf _{x \in \mathcal{N}}|A x| \leq \widetilde{C} \sqrt{n}\right)
$$

However, since $\# \mathcal{N} \leq e^{c_{1} n}$ and $\mathcal{N} \subset \frac{3}{2} B_{2}^{n} \backslash \frac{1}{2} B_{2}^{n}$,

$$
\begin{aligned}
\mathbb{P}\left(\inf _{x \in \mathcal{N}}|A x| \leq \widetilde{C} \sqrt{n}\right) & \leq \# \mathcal{N} \sup _{x \in \frac{3}{2} B_{2}^{n} \backslash \frac{1}{2} B_{2}^{n}} \mathbb{P}(|A x| \leq \widetilde{C} \sqrt{n}) \\
& \leq e^{c_{1} n} \sup _{x \in \frac{3}{2} B_{2}^{n} \backslash \frac{1}{2} B_{2}^{n}} \mathbb{P}(|A x| \leq \widetilde{C} \sqrt{n}) .
\end{aligned}
$$

Finally, by the UAC assumption and the Tensorization lemma 4.18 (which we can use by independence), $\mathbb{P}(|A x| \leq \widetilde{C} \sqrt{n}) \leq e^{-c^{\prime} n}$, which implies that

$$
\mathbb{P}\left(\inf _{x \in \mathcal{N}}|A x| \leq \widetilde{C} \sqrt{n}\right) \leq e^{c_{1} n} e^{-c^{\prime} n} \leq e^{-\widetilde{c} n}
$$

if $c_{1}>0$ is chosen small enough.
Remark 4.51. We used all the assumptions from the Theorem statement, including:

- Independent columns, which were used to constrct the net;
- $\mathbb{E}\|A\|_{H S}^{2} \leq K n^{2}$, which crucially ensured that $|A(x-y)| \leq \widetilde{C} \sqrt{n}$;
- Independent rows and uniform anti-concentration, which were used to ensure that $\sup _{x \in} \mathbb{P}(|A x| \leq \widetilde{C} \sqrt{n}) \leq e^{-c^{\prime} n}$.


### 4.7.2 Incompressible Vectors

While the result for the compressible vectors followed immediately from previous results, we will need to use different tools to handle the case of incompressible vectors in $\operatorname{Incomp}(\delta, \rho)$. The following Example demonstrates some crucial behavior of incompressible vectors which will be useful for bounding their behavior.
Example 4.52 (An incompressible vector). Consider the vector $x=\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right) \in$ $\mathbb{S}^{n-1}$. It is clear to see that for any $\delta>0$, any $y \in \operatorname{Sparse}(\delta)$ is such that

$$
|x-y| \geq\|x-y\|_{\infty} \geq \frac{1}{\sqrt{n}}
$$

since $y$ has at least one zero element. Thus, $x$ is in $\operatorname{Incomp}\left(\delta, \frac{1}{\sqrt{n}}\right)$. The vector is "spread", in the sense that there are many nonzero elements where $\left|x_{k}\right|=\frac{1}{\sqrt{n}}$.

Example 4.52 provides some intuition for how an incompressible vector behaves. In what follows, we will try to find some characterization of this behavior, where incompressible vectors will similarly have a substantial subset of coordinates where $\left|x_{k}\right| \sim \frac{1}{\sqrt{n}}$.
Lemma 4.53 (Incompressible vectors are spread). For every $x \in \operatorname{Incomp}(\delta, \rho)$, for $\delta, \rho>0$, there exists a subset of indices $\sigma \subset\{1, \ldots, n\}$ with $\# \sigma \geq \frac{1}{2} \rho^{2} \delta n$ such that for every $k \in \sigma$,

$$
\frac{\rho}{\sqrt{2 n}} \leq\left|x_{k}\right| \leq \frac{1}{\sqrt{\delta n}}
$$

Proof. Let $x \in \operatorname{Incomp}(\delta, \rho)$, and consider $\sigma_{1}, \sigma_{2} \subset\{1, \ldots, n\}$ such that

$$
\sigma_{1}:=\left\{k:\left|x_{k}\right| \leq \frac{1}{\sqrt{\delta n}}\right\}, \quad \sigma_{2}:=\left\{k:\left|x_{k}\right| \geq \frac{\rho}{\sqrt{2 n}}\right\} .
$$

We would like to show that the cardinality of $\sigma=\sigma_{1} \cap \sigma_{2}$ is controlled, i.e., that $\# \sigma \geq c n$ for some $c$. First, since $x \in \mathbb{S}^{n-1},|x|=1$, and thus,

$$
1=|x|^{2}=\sum_{k=1}^{n}\left|x_{k}\right|^{2} \geq \sum_{k \in \sigma_{1}^{\mathrm{C}}}\left|x_{k}\right|^{2} \geq \sum_{k \in \sigma_{1}^{\mathrm{C}}}\left|\frac{1}{\sqrt{\delta n}}\right|^{2} \geq \# \sigma_{1}^{\mathrm{\complement}} \frac{1}{\delta n} .
$$

Thus, $\# \sigma_{1}^{\complement} \leq \sqrt{\delta n}$, so $\# \sigma_{1} \geq n-\delta n$. Next, consider the following projection operator $P_{\sigma}$,

$$
P_{\sigma}(x)=\left(y_{1}, y_{2}, \ldots, y_{n}\right), \quad y_{k}=\left\{\begin{array}{ll}
x_{k} & k \in \sigma \\
0 & k \notin \sigma
\end{array} .\right.
$$

Let $y=P_{\sigma_{1}^{\mathrm{c}}}(x)$. Since $\# \sigma_{1} \geq n-\delta n$, this implies that $y \in \operatorname{Sparse}(\delta)$. But since $x \in$ $\operatorname{Incomp}(\delta, \rho),|x-y| \geq \rho$. Note that

$$
\left|P_{\sigma_{2}^{\mathrm{c}}}(x)\right|^{2}=\sum_{k:\left|x_{k}\right| \leq \frac{\rho}{\sqrt{2 n}}}\left|x_{k}\right|^{2} \leq n\left|\frac{\rho}{\sqrt{2 n}}\right|^{2}=\frac{\rho^{2}}{2}, \text { and }\left|P_{\sigma_{1}}(x)\right|^{2}=|x-y|^{2} \geq \rho^{2}
$$

which implies that

$$
\left|P_{\sigma}(x)\right|^{2} \geq\left|P_{\sigma_{1}}(x)\right|^{2}-\left|P_{\sigma_{2}^{\mathrm{c}}}(x)\right|^{2} \geq \frac{\rho^{2}}{2}
$$

On the other hand,

$$
\left|P_{\sigma}(x)\right|^{2} \leq \# \sigma \max _{k \in \sigma}\left|x_{k}\right|^{2} \leq \frac{1}{\delta n} \# \sigma,
$$

and therefore, $\# \sigma \geq \delta n \frac{\rho^{2}}{2}$.
Lemma 4.54 (Invertibility via distance). Let $A$ be a random matrix, let $X_{j}=A e_{j}$ be the columns of $A$, and $H_{j}=\operatorname{span}\left\{X_{i}: i \neq j\right\}$. Then for every $\delta, \rho \in\left(0, \frac{1}{2}\right), \varepsilon>0$,

$$
\mathbb{P}\left(\inf _{x \in \operatorname{Incomp}(\delta, \rho)}|A x| \leq \frac{\rho \varepsilon}{\sqrt{n}}\right) \leq \frac{1}{\delta n} \sum_{j=1}^{n} \mathbb{P}\left(\operatorname{dist}\left(X_{j}, H_{j}\right)<\varepsilon\right) .
$$

Proof. Let $x \in \operatorname{Incomp}(\delta, \rho)$, and let $X_{k}=A e_{k}$ be the $k$-th column of $A$. Note that for any vector $a$, and any subspace (passing through the origin) $H$, $\operatorname{dist}(a, H)=\inf \{|a-h|: h \in$ $H\} \leq|a-0|=|a|$, since $0 \in H$. Thus, since $A x$ is a vector, and $H_{k}$ is a subspace,

$$
|A x| \geq \max _{k=1, \ldots, n} \operatorname{dist}\left(A x, H_{k}\right)=\max _{k=1, \ldots, n} \operatorname{dist}\left(\sum_{j=1}^{n} x_{j} X_{j}, H_{k}\right)
$$

Note that by definition of $H_{k}, X_{j} \in H_{k}$ unless $j \neq k$. Thus, $\operatorname{dist}\left(X_{j}, H_{k}\right)=0$ for every $j \neq k$, so those components vanish, and

$$
\begin{equation*}
|A x| \geq \max _{k=1, \ldots, n} \operatorname{dist}\left(x_{k} X_{k}, H_{k}\right)=\max _{k=1, \ldots, n}\left|x_{k}\right| \operatorname{dist}\left(X_{k}, H_{k}\right) . \tag{46}
\end{equation*}
$$

Let $p_{k}:=\mathbb{P}\left(\operatorname{dist}\left(X_{k}, H_{k}\right) \leq \varepsilon\right)$. Consider the event $\mathcal{U}$ where $\sigma_{1}=\left\{k: \operatorname{dist}\left(X_{k}, H_{k}\right)>\varepsilon\right\}$ contains more than $(1-\delta) n$ elements. Then, using Markov's inequality,

$$
\begin{align*}
\mathbb{P}\left(\mathcal{U}^{\complement}\right)=\mathbb{P}\left(\# \sigma_{1}^{\complement} \geq \delta n\right) \leq \frac{1}{\delta n} \mathbb{E}\left(\# \sigma_{1}^{\complement}\right) & =\frac{1}{\delta n} \mathbb{E} \#\left\{k: \operatorname{dist}\left(X_{k}, H_{k}\right) \leq \varepsilon\right\} \\
& =\frac{1}{\delta n} \mathbb{E} \sum_{k=1}^{n} 1_{\left\{\operatorname{dist}\left(X_{k}, H_{k}\right) \leq \varepsilon\right\}}=\frac{1}{\delta n} \sum_{k=1}^{n} p_{k} . \tag{47}
\end{align*}
$$

Let $\sigma_{2}(x)=\left\{k:\left|x_{k}\right| \geq \frac{\rho}{\sqrt{n}}\right\}$. Note that $\left\|P_{\sigma_{2}(x)^{\mathrm{c}}}(x)\right\|_{2}^{2}=\sum_{k:\left|x_{k}\right| \leq \frac{\rho}{\sqrt{n}}}\left|x_{k}\right|^{2} \leq n\left|\frac{\rho}{\sqrt{n}}\right|^{2}=\rho^{2}$. This implies that $\sigma_{2}(x)$ has at least $\delta n$ elements, since otherwise we would have $|x-y|=$ $\left|P_{\sigma_{2}(x)^{\mathrm{c}}}(x)\right| \leq \rho$ for sparse vector $y=P_{\sigma_{2}(x)}(x)$, contradicting the incompressibility of $x$.

Next, suppose $\mathcal{U}$ occurs, and consider any $x \in \operatorname{Incomp}(\delta, \rho)$. We have that

$$
\# \sigma_{1}+\# \sigma_{2}(x)>(1-\delta) n+\delta n=n
$$

which implies that $\sigma_{1} \cap \sigma_{2}(x) \neq \emptyset$. Let $k \in \sigma_{1} \cap \sigma_{2}(x)$, then, using (46),

$$
|A x| \geq\left|x_{k}\right| \operatorname{dist}\left(X_{k}, H_{k}\right)>\frac{\rho}{\sqrt{n}} \varepsilon
$$

using $\left|x_{k}\right| \geq \frac{\rho}{\sqrt{n}}$ from $\sigma_{2}(x)$ and $\operatorname{dist}\left(X_{k}, H_{k}\right)>\varepsilon$ from $\sigma_{1}$. We have shown that $\mathcal{U} \Longrightarrow$ $\left\{\inf _{x \in \operatorname{Incomp}(\delta, \rho)}|A x| \geq \frac{\rho \varepsilon}{\sqrt{n}}\right\}$, or equivalently, that $\left\{\inf _{x \in \operatorname{Incomp}(\delta, \rho)}|A x| \leq \frac{\rho \varepsilon}{\sqrt{n}}\right\} \Longrightarrow \mathcal{U}^{\complement}$, so

$$
\mathbb{P}\left(\inf _{x \in \operatorname{Incomp}(\delta, \rho)}|A x| \leq \frac{\rho \varepsilon}{\sqrt{n}}\right) \leq \mathbb{P}\left(\mathcal{U}^{\complement}\right)=\frac{1}{\delta n} \sum_{k=1}^{n} p_{k}
$$

which follows from (47).
Lemma 4.54 reduces the problem from dealing with the incompressible vectors to finding an estimate of the following form

$$
\mathbb{P}(\operatorname{dist}(X, H) \leq \varepsilon) \leq c \varepsilon+\frac{c_{1}}{\sqrt{n}}
$$

where $X$ is a random vector with independent UAC entries, and $H$ is the span of $n-1$ independent random vectors with independent UAC entries.

### 4.7.3 Distance Theorem

Recall that our goal is to prove the following 'distance' theorem
Theorem 4.55 (Distance theorem). Let $X$ be a random vector with uniformly anti-concentrated (UAC) entries . Let $H$ be the span of $n-1$ independent random vectors with independent UAC entries. Then

$$
\begin{equation*}
\mathbb{P}(\operatorname{dist}(X, H) \leq \varepsilon) \leq c \varepsilon+\frac{c_{1}}{\sqrt{n}} \tag{48}
\end{equation*}
$$

Remark 4.56. We remark that 4.55 bounds the probability of the distance to the subspace by $c \varepsilon+\frac{c_{1}}{\sqrt{n}}$, but one can in fact improve the bound to $c \varepsilon+e^{-c_{1} n}$ using more sophisticated techniques [26].

As motivation for our next theorem, we note that the distance between $X$ and $H$ is exactly the length of the component of $X$ that is orthogonal to $H$. In particular dist $(X, H)=\langle x, n\rangle$, where $n$ is the normal unit vector of $H$.

Theorem 4.57 (Rogozin's Theorem). Let $v=\left(v_{1}, \cdots, v_{n}\right)$ be a random vector with independent UAC entries, say

$$
\begin{equation*}
\sup _{z \in \mathbb{R}} \mathbb{P}\left(\left|v_{i}-z\right|<a\right)<b \tag{49}
\end{equation*}
$$

Then for any $u \in \mathbb{R}^{n}$ and any $\varepsilon>c a\|u\|_{\infty}$

$$
\begin{equation*}
\sup _{z \in \mathbb{R}} \mathbb{P}(|\langle u, v\rangle-z|<\varepsilon) \leq \frac{C \varepsilon}{|u|}, \tag{50}
\end{equation*}
$$

where $C, c$ depend only on $a$ and $b$.
Corollary 4.58. Let $c_{1}, c_{2}>0$ and let $u$ be a random vector satisfying

$$
\#\left\{i:\left|u_{i}\right| \geq \frac{c_{1}}{\sqrt{n}}\right\} \geq c_{2} n
$$

If $v$ is a random vector with independent UAC entries then for all $\varepsilon>\frac{c}{\sqrt{n}}$ one has

$$
\sup _{z \in \mathbb{R}} \mathbb{P}[|\langle u, v\rangle-z|<\varepsilon] \leq c_{2} \varepsilon
$$

Proof. Let $\sigma:=\left\{i:\left|u_{i}\right| \geq \frac{c_{1}}{\sqrt{n}}\right\}$. By assumption $|\sigma| \geq c_{2} n$. Note that we may write $\langle u, v\rangle=R+\sum_{i \in \sigma} u_{i} v_{i}$, where $R=\sum_{i \notin \sigma} u_{i} v_{i}$. Therefore by Theorem 4.57

$$
\begin{aligned}
\sup _{z \in \mathbb{R}} \mathbb{P}(|\langle u, v\rangle-z|<\varepsilon) & =\sup _{z \in \mathbb{R}} \mathbb{P}\left(\left|\sum_{i \in \sigma} u_{i} v_{i}-(z-R)\right|<\varepsilon\right) \\
& =\sup _{z \in \mathbb{R}} \mathbb{E}_{R} \mathbb{P}\left(\left|\sum_{i \in \sigma} u_{i} v_{i}-(z-R)\right|<\varepsilon\right) \\
& \leq \mathbb{E}_{R} \sup _{z \in \mathbb{R}} \mathbb{P}\left(\left|\sum_{i \in \sigma} u_{i} v_{i}-(z-R)\right|<\varepsilon\right) \\
& =\sup _{y \in \mathbb{R}} \mathbb{P}\left(\left|\sum_{i \in \sigma} u_{i} v_{i}-y\right|<\varepsilon\right) \\
& \leq \frac{c \varepsilon}{\sqrt{\sum_{i \in \sigma} u_{i}^{2}}}=\tilde{c} \varepsilon,
\end{aligned}
$$

whenever $\varepsilon \geq \sup _{i \in \sigma}\left|u_{i}\right| \geq \frac{c_{1}}{\sqrt{n}}$.
Using Corollary 4.58 we will show that

$$
\sup _{z \in \mathbb{R}} \mathbb{P}(|\langle X, n\rangle-z|<\varepsilon) \leq C \varepsilon+\frac{c_{1}}{\sqrt{n}}
$$

Where $X=A e_{j}$ and $n$ is the the unit normal to $\operatorname{span}\left(A e_{j}: i \neq j\right)$. This, however, will require that $n$ is incompressible.

Lemma 4.59 (Random normal is incompressible). Let $H=\operatorname{span}\left(A e_{i}: i \neq j\right)$, where $A$ is a matrix satisfying the assumptions of Theorem 4.45. Let $n \perp H$ be a unit vector. Then $n$ is incompressible with probability $1-e^{-c n}$.

Proof. Note that the condition $n \perp H$ is equivalent to the condition $B^{\top} n=0$, where $B=\left[A e_{1}, \cdots, A e_{j-1}, A e_{j+1}, \cdots, A e_{n}\right]$. Then

$$
\mathbb{P}(n \in \operatorname{Comp}(\delta, \rho)) \leq \mathbb{P}\left(\inf _{x \in \operatorname{Comp}(\delta, \rho)}\left|B^{\top} x\right|=0\right) \leq e^{-c n}
$$

where the exponential failure probability follows from the net-argument for compressible vectors, as done in Lemma 4.50.

This lemma, in conjunction with Corollary 4.58 , will give us the distance theorem with $X=A e_{i}$ and $H=\operatorname{span}\left(A e_{j}: j \neq i\right)$.

### 4.7.4 Proof of the first part of Theorem 4.45

Recall that $A$ has independent UAC entries and satisfies $\mathbb{E}\|A\|_{H S}^{2} \leq K n^{2}$. Therefore

$$
\begin{aligned}
\mathbb{P}\left(\sigma_{n}(A) \leq \frac{\varepsilon}{\sqrt{n}}\right) & \leq \mathbb{P}\left(\inf _{x \in \operatorname{Comp}(\delta, \rho)}|A x|<\frac{\varepsilon}{\sqrt{n}}\right)+\mathbb{P}\left(\inf _{x \in \operatorname{Incomp}(\delta, \rho)}|A x|<\frac{\varepsilon}{\sqrt{n}}\right) \\
& \leq e^{-c_{1} n}+\mathbb{P}\left(\inf _{x \in \operatorname{Incomp}(\delta, \rho)}|A x|<\frac{\varepsilon}{\sqrt{n}}\right) \\
& \leq e^{-c_{1} n}+\frac{1}{\delta n} \sum_{i=1}^{n} \mathbb{P}\left(\operatorname{dist}\left(A e_{i}, H\right)<\varepsilon\right) .
\end{aligned}
$$

Note that in the first line we used the Rudelson-Vershynin decomposition of the sphere, in the second line we used Lemma 4.50, and in the third line we used Lemma 4.54. Next we have that

$$
\begin{aligned}
\mathbb{P}\left(\operatorname{dist}\left(A e_{i}, H\right)<\varepsilon\right) & \leq \mathbb{P}(|\langle A e, n\rangle| \leq \varepsilon), \\
& \leq e^{-c_{1} n}+\mathbb{P}(|\langle A e, n\rangle| \leq \varepsilon, n \in \operatorname{Incomp}(\delta, \rho)) \\
& \leq e^{-c_{1} n}+\underset{\substack{u \text { s.t. } \\
\#\left\{i:\left|u_{i}\right| \geq c_{1} \\
\sqrt{n}\right\} \geq c_{2} n}}{ } \quad \mathbb{P}\left(\left|\left\langle A e_{i}, u\right\rangle\right| \leq \varepsilon\right) \\
& \leq e^{-c_{1} n}+c \varepsilon+\frac{c_{2}}{\sqrt{n}} .
\end{aligned}
$$

Note that in the second line we used Lemma 4.59, in the third line we used the fact that incompressible vectors are spread, and in the fourth line used Rogozin's theorem. Plugging this estimate in our bound for the smallest singular value yields

$$
\begin{align*}
\mathbb{P}\left(\sigma_{n}(A) \leq \frac{\varepsilon}{\sqrt{n}}\right) & \leq e^{-c_{1} n}+\left(\frac{e^{-c_{1} n}}{\delta}+\frac{c \varepsilon}{\delta}+\frac{c_{2}}{\delta \sqrt{n}}\right)  \tag{51}\\
& \leq c^{\prime} \varepsilon+\frac{c^{\prime \prime}}{\sqrt{n}} \tag{52}
\end{align*}
$$

## 5 Gaussian Random Processes

Recall that a random vector is a collection of $n$ random variables, forming a vector (i.e. $\left.x=\left(x_{1}, \cdots, x_{n}\right)\right)$. A random walk is a sequence of random variables (i.e. $\left\{x_{1}, x_{2}, \ldots\right\}$ ). In 1-D a random process is a collection of random variables indexed by "time". Examples include

$$
\left\{X_{t}: t \in \mathbb{R}\right\}, \quad\left\{X_{t}: t \geq 0\right\}, \quad\left\{X_{t}: t \in[a, b]\right\} .
$$

### 5.1 Basic Concepts and Examples

Example 5.1 (Brownian Motion). A brownian motion $\left\{X_{t}: t \geq 0\right\}$ is a random process having the following properties:

1. For all $s \geq t \geq 0$ the random variable $X_{s}-X_{t}$ is distributed as the normal variable $N(0, t-s)$. This is known as having "gaussian increments".
2. The function $f(t)=X_{t}$ is continuous in $t$ almost surely.
3. For $\gamma \leq \tau \leq t \leq s$ the increments $X_{s}-X_{t}$ and $X_{\tau}-X_{\gamma}$ are independent.

We will be interested in high dimensional random processes (i.e. random processes where "time" is a subset of $\mathbb{R}^{n}$ ).

Example 5.2 (Ocean temperature). To represent the temperature of the ocean as a random process we can take $T \subset \mathbb{R}^{n}$ and for $t \in T$ let $X_{t}$ denote the temperature at $t$.

Example 5.3 (Random Projection). Let $g \sim N\left(0, I_{d}\right)$ be a standard gaussian vector. In other words $g=\left(g_{1}, \ldots, g_{n}\right)$ where the cooordinates $g_{i}$ are independent standard normal gaussians. Then for $t \in T \subset \mathbb{R}^{n}$ we define $X_{t}:=\langle g, t\rangle$.

We now recall the definition of a gaussian random vector. Given a non-negative definite $n \times n$ matrix $A$, and a vector $\mathrm{b} \in \mathbb{R}^{n}$, we define the random variable $X \sim N(b, A)$ whose law has density $e^{\langle A(x-b), x-b\rangle} \cdot c_{n}$, where $c_{n}$ is chosen so that the density integrates to 1 .

Remark 5.4. (Homework) Let $X \sim N(0, \Sigma)$ be a random gaussian vector where $\Sigma=$ $A A^{\top}$. Then there exists $g \sim N\left(0, \operatorname{Id}_{n}\right)$ such that $X_{i}=\left\langle g, u_{i}\right\rangle$ for some $u_{1}, \ldots, u_{n} \in \mathbb{R}^{n}$. Note that the coordinates of $X$ may be dependent.

Definition 5.5. The covariance matrix $\Sigma$ of a random vector $X \in \mathbb{R}^{n}$ is the $n \times n$ matrix with entries

$$
\Sigma_{i j}=\mathbb{E}\left[\left(X-\mathbb{E} X_{i}\right)\left(X-\mathbb{E} X_{j}\right)\right]
$$

Remark 5.6. If $X$ has independent entries then $\Sigma$ is a diagonal matrix.

As a cool fact (HW), we remark that the distribution of a Gaussian vector $X \sim N(0, \Sigma)$ is uniquely determined by its covariance matrix. In general, knowing that a random vector belongs to a certain class of vectors (say Poisson or Exponential) and knowing its covariance matrix are not enough to recover its distribution.

Definition 5.7 (Covariance Function). Let $\left\{X_{t}: t \in T\right\}$ be a "mean zero" random process (i.e. $\mathbb{E}\left[X_{t}\right]=0$ for all $t \in T$ ). We define $\Sigma: T \times T \rightarrow \mathbb{R}$ according to

$$
\Sigma(t, s)=\mathbb{E}\left[X_{t} X_{s}\right] .
$$

$\Sigma$ is known as the covariance function and is the random process analogue to the covariance matrix.

Definition 5.8 (Gaussian Random Process). Let $T \subset \mathbb{R}^{n}$. A random process $\left\{X_{t}: t \in T\right\}$ is called a Gaussian Random Process (GRP) if for every finite subset $T_{0} \subset T$, the vector $\left(X_{t}\right)_{t \in T_{0}}$ is a gaussian vector. An equivalent characterization is that for every finite subset $T_{0} \subset T$ and vector $\left(a_{t}\right)_{t \in T_{0}}$ the linear combination $\sum_{t \in T_{0}} a_{t} X_{t}$ is a gaussian random variable. This equivalence is because the projection of a gaussian vector in any direction is a normal random variable, and a random vector whose projection in every direction is a normal random variable must be a gaussian vector. (HW) A gaussian random process is determined by its covariance function.

Definition 5.9 ("Canonical GRP"). Let $T \subset R^{n}$ and let $g \sim N\left(0, \operatorname{Id}_{n}\right)$. For every $t \in T$ define $X_{t}=\langle g, t\rangle$. Then $\left\{X_{t}: t \in T\right\}$ is known as a Canonical Gaussian Process. Note that this is indeed a Gaussian Random Process since, by the definition of $X_{t}$, each $X_{t}$ is the projection of a gaussian vector and therefore a normal random variable and therefore any linear combination of a finite number of $X_{t}$ is a normal random variable.

Lemma 5.10 (All GRP are canonical). Let $Y_{t}$ be a mean zero Gaussian Random Process. Then there exists $T \subset \mathbb{R}^{n}$ such that $Y_{t}=\langle g, t\rangle$ for all $t \in T$, where $g \sim N(0,1)$.

### 5.2 Slepian's Inequality

In applications, it is useful to have a uniform control on a random process $\left\{X_{t}: t \in T\right\}$, i.e. to have a bound on $\mathbb{E} \sup _{t \in T} X_{t}$.

For some processes, this quantity can be computed exactly. For example, if $\left\{X_{t}: t \in T\right\}$ is a standard Brownian motion, then by reflection principle, we have $\mathbb{E} \sup _{t \leq t_{0}} X_{t}=\sqrt{2 t_{0} / \pi}$ for every $t_{0} \geq 0$. For general random processes, even if they are Gaussian, the problem is emphatically nontrivial.

The first general bound we will prove is Slepian's comparison inequality for Gaussian processes. Intuitively, it states that the faster the process grows (in terms of the magnitude of the increments), the farther it gets.

Theorem 5.11 (Slepian's Inequality). Let $\left\{X_{t}: t \in T\right\},\left\{Y_{t}: t \in T\right\}$ be mean zero gaussian processes indexed by $T$. Suppose that for all $s, t \in T$ it holds that

$$
\mathbb{E} X_{t}^{2}=\mathbb{E} Y_{t}^{2}, \quad \mathbb{E}\left(X_{t}-X_{s}\right)^{2} \leq \mathbb{E}\left(Y_{t}-Y_{s}\right)^{2}
$$

Then for all $\tau \in \mathbb{R}$ it follows that

$$
\begin{equation*}
\mathbb{P}\left[\sup _{t \in T} X_{t} \geq \tau\right] \leq \mathbb{P}\left[\sup _{t \in T} Y_{t} \geq \tau\right] \tag{53}
\end{equation*}
$$

and thus

$$
\begin{equation*}
\mathbb{E} \sup _{t \in T} X_{t} \leq \mathbb{E} \sup _{t \in T} Y_{t} \tag{54}
\end{equation*}
$$

Remark 5.12 (Homework). Instead of the Gaussian processes $\left\{X_{t}: t \in T\right\}$ and $\left\{Y_{t}: t \in\right.$ $T\}$, it suffices to prove the above inequality for the Gaussian random vectors $X$ and $Y$ in $\mathbb{R}^{n}$. Furthermore, it suffices to prove for the case when $X$ and $Y$ are independent. Hence, Theorem 5.11 is equivalent to Theorem 5.19 which will be proved later.

Then the inequalities (53) and (54) are equivalent to

$$
\mathbb{P}\left[\sup _{i \leq n} X_{i} \geq \tau\right] \leq \mathbb{P}\left[\sup _{i \leq n} Y_{i} \geq \tau\right], \quad \mathbb{E} \sup _{i \leq n} X_{i} \leq \mathbb{E} \sup _{i \leq n} Y_{i},
$$

where the first inequality always guarantees the second since

$$
\mathbb{E} \max _{i \leq n} X_{i}^{2}=\int_{0}^{\infty} \mathbb{P}\left(\max _{i \leq n} X_{i}^{2} \geq \tau\right) d \tau \leq \int_{0}^{\infty} \mathbb{P}\left(\max _{i \leq n} Y_{i}^{2} \geq \tau\right) d \tau=\mathbb{E} \max _{i \leq n} Y_{i}^{2}
$$

by the tail formula for non-negative random variables.

### 5.2.1 Gaussian Interpolation

The proof of Slepian's inequality will be based on the technique of Gaussian Interpolation which is described as follows.

Definition 5.13 (Gaussian Interpolation). For any pair of independent Gaussian random vectors $X, Y \in \mathbb{R}^{n}$, not necessarily standard, define a Gaussian random vector $Z(u)$ in $\mathbb{R}^{n}$ that continuous interpolates between $Z(0)=Y$ and $Z(1)=X$ :

$$
Z(u):=\sqrt{u} X+\sqrt{1-u} Y, \quad u \in[0,1] .
$$

Remark 5.14 (Homework). The covriance matrix of $Z(u)$ interpolates linearly between the covariance matrices of $Y$ and $X$. Namely, if $\Sigma(X)$ is the covariance matrix for $X$ and $\Sigma(Y)$ is the covariance matrix for $Y$, then

$$
\Sigma(Z(u))=u \Sigma(X)+(1-u) \Sigma(Y)
$$

Consider the indicator function for vector $x=\left(x_{1}, \cdots, x_{n}\right)$

$$
f(x):=\mathbf{1}_{\left\{\max _{i} x_{i}<\tau\right\}}
$$

which satisfies $\mathbb{E} f(Z(1))=\mathbb{P}\left(\max _{i \leq n} X_{i}<\tau\right)$ and $\mathbb{E} f(Z(0))=\mathbb{P}\left(\max _{i \leq n} Y_{i}<\tau\right)$. Now, if we can show

$$
\mathbb{E} f(Z(1)) \geq \mathbb{E} f(Z(0))
$$

then inequality (53) can be concluded. Our goal now shifts to study how the quantity $\mathbb{E} f(Z(u))$ changes as $u$ increases from 0 to 1 . We approach this goal by starting with the following identity which is a version of integration by parts in Gaussian expectations.

Lemma 5.15 (Gaussian integration by parts). Let $X \sim N(0,1)$. Then for any differentiable function $f: \mathbb{R} \rightarrow \mathbb{R}$ we have

$$
\begin{equation*}
\mathbb{E} f^{\prime}(X)=\mathbb{E} X f(X) \tag{55}
\end{equation*}
$$

Proof. It suffices to argue for the case when $f$ has bounded support, and this identity then can be extended to general functions by a standard approximation argument. By density of standard normal and Integration by parts, we have

$$
\mathbb{E} f^{\prime}(X)=\int_{\mathbb{R}} f^{\prime}(t) \frac{1}{\sqrt{2 \pi}} e^{-t^{2} / 2} d t=0+\int_{\mathbb{R}} t f(t) \frac{1}{\sqrt{2 \pi}} e^{-t^{2} / 2} d t=\mathbb{E} X f(X)
$$

as claimed, where the zero term comes from the fact that $f(t) e^{-t^{2} / 2}$ has limits equal to zero when $t$ goes to both positive and negative infinity if $f$ has bounded support.

Corollary 5.16 (Homework: Multivariate Gaussian Integration by parts). Let $X \sim N(0, \Sigma)$. Then for any differentiable function $f: \mathbb{R}^{n} \rightarrow \mathbb{R}$, we have

$$
\mathbb{E} X f(X)=\Sigma \mathbb{E} \nabla f(X)
$$

where $\nabla f(X)$ is the $n$-dimensional gradient vector with entries $\mathbb{E} \partial_{i} f(X)$.
Lemma 5.17 (Gaussian Interpolation). Consider two independent Gaussian random vectors $X \sim N\left(0, \Sigma^{X}\right)$ and $Y \sim N\left(0, \Sigma^{Y}\right)$. Define the interpolation Gaussian vector

$$
\begin{equation*}
Z(u):=\sqrt{u} X+\sqrt{1-u} Y, \quad u \in[0,1] . \tag{56}
\end{equation*}
$$

Then for any twice-differentiable function $f: \mathbb{R}^{n} \rightarrow \mathbb{R}$, we have

$$
\begin{equation*}
\frac{d}{d u} \mathbb{E} f(Z(u))=\frac{1}{2} \sum_{i, j=1}^{n}\left(\Sigma_{i j}^{X}-\Sigma_{i j}^{Y}\right) \mathbb{E} \frac{\partial^{2} f}{\partial x_{i} \partial x_{j}}(Z(u)), \tag{57}
\end{equation*}
$$

where $\Sigma_{i j}^{X}$ stands for the $(i, j)$ entry of $\Sigma^{X}$.

Proof. By the (multivariate) chain rule, we have

$$
\frac{d}{d u} \mathbb{E} f(Z(u))=\sum_{i=1}^{n} \mathbb{E} \frac{\partial f}{\partial x_{i}}(Z(u)) \frac{d Z_{i}}{d u}=\frac{1}{2} \sum_{i=1}^{n} \mathbb{E} \frac{\partial f}{\partial x_{i}}(Z(u))\left(\frac{X_{i}}{\sqrt{u}}-\frac{Y_{i}}{\sqrt{1-u}}\right)
$$

where the second equality is from the definition (56) of $Z(u)$. Now break this sum into two, and first compute the contribution of terms containing $X_{i}$. To this end, we condition on $Y$ and express

$$
\sum_{i=1}^{n} \frac{1}{\sqrt{u}} \mathbb{E} X_{i} \frac{\partial f}{\partial x_{i}}(Z(u))=\sum_{i=1}^{n} \frac{1}{\sqrt{u}} \mathbb{E} X_{i} g_{i}(X)
$$

where

$$
g_{i}(X)=\frac{\partial f}{\partial x_{i}}(\sqrt{u} X+\sqrt{1-u} Y) .
$$

Apply the multivariate Gaussian integration by parts (Corollary 5.16), we have

$$
\mathbb{E} X_{i} g_{i}(X)=\sum_{j=1}^{n} \Sigma_{i j}^{X} \mathbb{E} \frac{\partial g_{i}}{\partial x_{j}}(X)=\sum_{j=1}^{n} \Sigma_{i j}^{X} \mathbb{E} \frac{\partial^{2} f}{\partial x_{i} \partial x_{j}}(X)(\sqrt{u} X+\sqrt{1-u} Y) \sqrt{u},
$$

where the second equality is by definition of $g_{i}$. Substituting this back into the previous equation gives

$$
\sum_{i=1}^{n} \frac{1}{\sqrt{u}} \mathbb{E} X_{i} \frac{\partial f}{\partial x_{i}}(Z(u))=\sum_{i, j=1}^{n} \Sigma_{i j}^{X} \mathbb{E} \frac{\partial^{2} f}{\partial x_{i} \partial x_{j}}(Z(u))
$$

Taking expectation of both sizes with respect to $Y$, we remove the conditioning on $Y$. Similar discussion works for the contribution of terms containing $Y_{i}$, and that should yield

$$
\sum_{i=1}^{n} \frac{1}{\sqrt{1-u}} \mathbb{E} X_{i} \frac{\partial f}{\partial x_{i}}(Z(u))=\sum_{i, j=1}^{n} \Sigma_{i j}^{Y} \mathbb{E} \frac{\partial^{2} f}{\partial x_{i} \partial x_{j}}(Z(u))
$$

Combining these two equalities, together with the first line of the proof, gives the desired relation (57).

### 5.2.2 Proof of Slepian's Inequality

We are ready to establish the key lemma in proving Slepian's inequality which is also known as a preliminary functional form of Slepian's inequality.

Lemma 5.18 (Slepian 's inequality: functional form). Consider two mean-zero independent Gaussian random vectors $X$ and $Y$ in $\mathbb{R}^{n}$. Assume that for all $i, j=1, \cdots, n$, we have

$$
\begin{equation*}
\mathbb{E} X_{i}^{2}=\mathbb{E} Y_{i}^{2} \quad \text { and } \quad \mathbb{E}\left(X_{i}-X_{j}\right)^{2} \leq \mathbb{E}\left(Y_{i}-Y_{j}\right)^{2} . \tag{58}
\end{equation*}
$$

Consider a twice-differentiable function $f: \mathbb{R}^{n} \rightarrow \mathbb{R}$ such that

$$
\frac{\partial^{2} f}{\partial x_{i} \partial x_{j}} \geq 0 \quad \text { for all } i \neq j
$$

Then

$$
\mathbb{E} f(X) \geq \mathbb{E} f(Y)
$$

Proof. The assumptions (58) imply that the entries of the covariance matrices $\Sigma^{X}$ and $\Sigma^{Y}$ of $X$ and $Y$ satisfy

$$
\Sigma_{i i}^{X}=\Sigma_{i i}^{Y} \quad \text { and } \quad \Sigma_{i j}^{X} \geq \Sigma_{i j}^{Y}, \quad \text { for all } i, j=1, \cdots, n,
$$

where for the second relation we used $a b=\left(a^{2}+b^{2}-(a-b)^{2}\right) / 2$. Applying Lemma 5.17 with our assumptions gives

$$
\frac{d}{d u} \mathbb{E} f(Z(u))=\frac{1}{2} \sum_{i, j=1}^{n}\left(\Sigma_{i j}^{X}-\Sigma_{i j}^{Y}\right) \mathbb{E} \frac{\partial^{2} f}{\partial x_{i} \partial x_{j}}(Z(u)) \geq 0
$$

by assumptions, where $Z(u)$ is the Gaussian interpolation we defined in (56). This means $\mathbb{E} f(Z(u))$ increases in $u$. But $Z(0)=Y$ and $Z(1)=X$ by the way we constructed the interpolation, hence $\mathbb{E} f(X)=\mathbb{E} f(Z(1)) \geq \mathbb{E} f(Z(0))=\mathbb{E} f(Y)$ as desired.

Eventually, now we are ready to prove Slepian's inequality (53) in it's equivalent form which is in terms of random vectors rather than random processes.

Theorem 5.19 (Slepian's inequality: random vector). Let $X$ and $Y$ be two mean zero independent Gaussian random vectors in $\mathbb{R}^{n}$. Suppose that for all $s, t \in T$ it holds that

$$
\mathbb{E} X_{i}^{2}=\mathbb{E} Y_{i}^{2}, \quad \mathbb{E}\left(X_{i}-X_{j}\right)^{2} \leq \mathbb{E}\left(Y_{i}-Y_{j}\right)^{2}
$$

Then for every $\tau \in \mathbb{R}$ we have

$$
\begin{equation*}
\mathbb{P}\left[\max _{i \leq n} X_{i} \geq \tau\right] \leq \mathbb{P}\left[\max _{i \leq n} Y_{i} \geq \tau\right] . \tag{59}
\end{equation*}
$$

Consequently,

$$
\begin{equation*}
\mathbb{E} \max _{i \leq n} X_{i} \leq \mathbb{E} \max _{i \leq n} Y_{i} \tag{60}
\end{equation*}
$$

Proof. Let $h: \mathbb{R} \rightarrow[0,1]$ be a twice-differentiable non-increasing approximation to the indicator function of the interval $(-\infty, \tau)$ satisfying

$$
h(x) \approx \mathbf{1}_{(-\infty, \tau)} .
$$

Define the function $f: \mathbb{R}^{n} \rightarrow \mathbb{R}$ by $f(x)=h\left(x_{1}\right) \cdots h\left(x_{n}\right)$ for any $x=\left[x_{1}, \cdots, x_{n}\right] \in \mathbb{R}^{n}$, then

$$
f(x) \approx \mathbf{1}_{\left\{\max _{i} x_{i}<\tau\right\}}
$$

To apply the functional form of Slepian's inequality, we need to check the assumption for $f$. Note that, for every $i \neq j$, we have

$$
\frac{\partial^{2} f}{\partial x_{i} \partial x_{j}}=h^{\prime}\left(x_{i}\right) h^{\prime}\left(x_{j}\right) \prod_{k \notin\{i, j\}} h\left(h_{k}\right) .
$$

The first two factors are non-positive and the other are non-negative by assumption. So the second mixture derivative of $f$ is always non-negative. Hence we can apply the key Lemma above to conclude that

$$
\mathbb{E} f(X) \geq \mathbb{E} f(Y)
$$

Now, by approximation, we have
$\mathbb{P}\left(\max _{i \leq n} X_{i}<\tau\right)=\mathbb{E} \mathbf{1}_{\left\{\max _{i \leq n} X_{i}<\tau\right\}} \approx \mathbb{E} f(X) \geq \mathbb{E} f(Y) \approx \mathbb{E} \mathbf{1}_{\left\{\max _{i \leq n} Y_{i}<\tau\right\}}=\mathbb{P}\left(\max _{i \leq n} Y_{i}<\tau\right)$
which implies (59). And relation (60) follows from (59) as we discussed before.

### 5.2.3 The Sudakov-Fernique Inequality

In theorem 5.11, Slepian's inequality has two assumptions on the random processes $\left\{X_{t}\right.$ : $t \in T\}$ and $\left\{Y_{t}: t \in T\right\}$ : the equality of variances and the dominance of increments. It turns out that, even if we drop the assumption on the equality of variances, we will still be able to obtain the inequality on expectations. This more pratically useful result is due to Sudakov and Fernique.

Theorem 5.20 (Sudakov-Fernique Inequality). Let $\left\{X_{t}: t \in T\right\}$ and $\left\{Y_{t}: t \in T\right\}$ be two mean zero Gaussian processes. Assume that, for all $t, s \in T$, we have

$$
\mathbb{E}\left(X_{t}-X_{s}\right)^{2} \leq \mathbb{E}\left(Y_{t}-Y_{s}\right)^{2}
$$

Then

$$
\mathbb{E} \sup _{t \in T} X_{t} \leq \mathbb{E} \sup _{t \in T} Y_{t} .
$$

Proof. It is enough to prove this theorem for Gaussian random vectors $X$ and $Y$ in $\mathbb{R}^{n}$, just as we did for Slepian's inequality. We again deduce the result form the Gaussian Interpolation lemma 5.17. But this time, instead of choosing a function $f(x)$ that approximates the indicator function of $\left\{\max _{i} x_{i}<\tau\right\}$, we want $f(x)$ to approximate $\max _{i} x_{i}$.

To this end, let $\beta>0$ be a parameter and define the function

$$
f_{\beta}(x):=\frac{1}{\beta} \log \sum_{i=1}^{n} e^{\beta x_{i}}
$$

One can check (Homework!) that this function is twice differentiable and

$$
f_{\beta}(x) \rightarrow \max _{i \leq n} x_{i} \quad \text { as } \beta \rightarrow \infty
$$

Substitute $f(x)$ into the Gaussian interpolation formula and simplifying the expression shows that (Homework!)

$$
\frac{d}{d u} \mathbb{E} f(Z(u)) \leq 0, \quad \text { for all } u
$$

The proof can then be completed in just the same way as the proof of Slepian's inequality.

### 5.2.4 Application of Sudakov-Fernique Inequality to Gaussian Random Matrices

Theorem 5.21. Let $A$ be an $m \times n$ matrix with entries $\left(a_{i j}\right)$ such that $a_{i j}$ are independent and $a_{i j} \sim \mathcal{N}(0,1)$. Then we have

$$
\mathbb{E}\|A\| \leq \sqrt{m}+\sqrt{n}
$$

Remark 5.22. Note $\|\cdot\|_{o p}=\sigma_{1}(A)$. This is consistent with our past results as we have proved that sub-Gaussian ensures $\mathbb{P}(\|A\| \geq t(\sqrt{m}+\sqrt{n})) \leq e^{-c m}$ with $m \geq n$. But here we see that the constant is 1 , i.e., $\mathbb{E}\|A\| \leq C(\sqrt{m}+\sqrt{n})$ with $C=1$.

Proof of Theorem 5.21. Let $A$ be an $m \times n$ matrix with entries $\left(a_{i j}\right)$ such that $a_{i j}$ are independent and $a_{i j} \sim \mathcal{N}(0,1)$. Then

$$
\|A\|=\sup _{x \in \mathbb{S}^{n-1}}|A x|=\sup _{\substack{y \in \mathbb{S}^{m-1} \\ x \in \mathbb{S}^{n-1}}}\langle A x, y\rangle=\max _{\substack{t=x, y) \\ t \in T \\ T=\mathbb{S}^{n-1} \times \mathbb{S}^{m-1}}} X_{t},
$$

since if we have $a \in \mathbb{R}^{m}$, then $|a|=\sup _{y \in \mathbb{S}^{m-1}}\langle a, y\rangle$. Here we have $X_{t}$ as a random Gaussian process, indeed, if we fix $x$ and $y$, then $\langle A x, y\rangle$ is a Gaussian random variable. If $A$ has all the Gaussian entries, then all rows of $A$ are Gaussian processes, i.e.,

$$
A x=\left(\begin{array}{c}
\left\langle A^{\top} e_{1}, x\right\rangle \\
\vdots \\
\left\langle A^{\top} e_{m}, x\right\rangle
\end{array}\right)
$$

with $A^{\top} e_{i} \sim \mathcal{N}(0, I d)$ and $A^{\top} e_{i}$ are all independent. Thus, $A x$ is a vector with independent coordinates and each of them is Gaussian, so $A x$ is Gaussian, and $\langle A x, y\rangle$ is a Gaussian random variable. The idea of the proof is that we can apply Sudakov-Fernique Inequality 5.20 to find $Y_{t}$ indexed by $\mathbb{S}^{m-1} \times \mathbb{S}^{n-1}$ which is also Gaussian, then compare $X_{t}$ to $Y_{t}$ to get the upper bound of $\mathbb{E}\|A\|=\mathbb{E} \sup X_{t} \leq \mathbb{E} \sup Y_{t}$ and find the increment by estimating from above of $\mathbb{E}\left(X_{t}-X_{s}\right)^{2}$.

Let $t=(u, v) \in \mathbb{S}^{n-1} \times \mathbb{S}^{m-1}$ and let $s=(w, z) \in \mathbb{S}^{n-1} \times \mathbb{S}^{m-1}$. Then

$$
\begin{aligned}
\mathbb{E}\left(X_{t}-X_{s}\right)^{2} & =\mathbb{E}(\langle A u, v\rangle-\langle A w, z\rangle)^{2} \\
& =\mathbb{E}\left(\sum_{i, j} a_{i j}\left(u_{i} v_{j}-w_{i} z_{j}\right)\right)^{2} \\
& =\sum_{i, j} \mathbb{E} a_{i j}^{2}\left(u_{i} v_{j}-w_{i} z_{j}\right)^{2} \\
& =\sum_{i, j}\left(u_{i} v_{j}-w_{i} z_{j}\right)^{2} \\
& =\|u \otimes v-w \otimes z\|_{H S}^{2} \\
& \stackrel{H W}{\leq}|u-w|^{2}+|v-z|^{2} .
\end{aligned}
$$

Recall that for independent and mean zero $\xi_{1}, \ldots, \xi_{k}$, we have $\mathbb{E}\left(\sum \xi_{l}\right)^{2}=\sum \mathbb{E} \xi_{l}^{2}$ so here $a_{i j}\left(u_{i} v_{j}-w_{i} z_{j}\right)$ are independent and mean zero. We can conclude that

$$
\mathbb{E}\left|X_{u v}-X_{w z}\right|^{2} \leq|u-w|^{2}+|v-z|^{2} .
$$

In fact, we can construct $Y_{t}$ with $t \in \mathbb{S}^{n-1} \times \mathbb{S}^{m-1}$ such that

$$
\mathbb{E}\left|Y_{u v}-Y_{w z}\right|^{2} \leq|u-w|^{2}+|v-z|^{2} .
$$

Indeed, consider $Y_{u v}=\langle g, u\rangle+\langle h, v\rangle$ where $h \sim \mathcal{N}\left(0, I d_{m}\right), g \sim \mathcal{N}\left(0, I d_{n}\right), h, g$ are independent, and $(u, v) \in \mathbb{S}^{n-1} \times \mathbb{S}^{m-1}$. Note that here $Y_{u v}$ is a Gaussian random process, then by definition

$$
\begin{aligned}
\mathbb{E}\left|Y_{u v}-Y_{w z}\right|^{2} & =\mathbb{E}|\langle g, u\rangle+\langle h, v\rangle-\langle g, w\rangle-\langle h, z\rangle|^{2} \\
& =\mathbb{E}\langle g, u-w\rangle^{2}+\mathbb{E}\langle h, v-z\rangle^{2} \quad \text { (by independence and mean zero) } \\
& =|u-w|^{2}+|v-z|^{2}
\end{aligned}
$$

since if $X \sim \mathcal{N}(0, I d)$, we have $\langle x, \theta\rangle \sim \mathcal{N}\left(0,|\theta|^{2}\right)$.
By the results above, we can conclude that $\mathbb{E}\left|X_{t}-X_{s}\right|^{2} \leq \mathbb{E}\left|Y_{t}-Y_{s}\right|^{2}$. Since $X_{t}$ and $Y_{t}$ are Gaussian, by Sudakov-Fernique Inequality 5.20 , we have

$$
\begin{aligned}
& \mathbb{E}\|A\|_{o p}^{2}=\mathbb{E} \sup _{t \in \mathbb{S}^{n-1} \times \mathbb{S}^{m-1}} X_{t} \leq \mathbb{E} \sup _{t \in \mathbb{S}^{n-1} \times \mathbb{S}^{m-1}} Y_{t} \\
&=\mathbb{E} \sup _{\substack{u \in \mathbb{S}^{n-1} \\
v \in \mathbb{S}^{m-1}}}(\langle g, u\rangle+\langle h, v\rangle) \\
&=\mathbb{E} \sup _{u \in \mathbb{S}^{n-1}}\langle g, u\rangle+\mathbb{E} \sup _{v \in \mathbb{S}^{m-1}}\langle h, v\rangle \\
&=\mathbb{E}|g|+\mathbb{E}|h| \\
& \text { Cauchy } \\
& \leq \sqrt{\mathbb{E}|g|^{2}}+\sqrt{\mathbb{E}|h|^{2}} \\
&=\sqrt{n}+\sqrt{m}
\end{aligned}
$$

as $\mathbb{E}|g|^{2}=\mathbb{E}\left(\sum_{i=1}^{n} g_{i}^{2}\right)=n$ and $g \sim \mathcal{N}(0, I d)$.
Definition 5.23 (Sub-Gaussian Random Process). For $X_{t}$ with $t \in T$ and metric $d$ on T , we say that the random process $X_{t}$ is sub-Gaussian if for some constant $K \geq 0$,

$$
\left\|X_{t}-X_{s}\right\|_{\psi_{2}} \leq K \cdot d(t, s)
$$

i.e., all increments are sub-Gaussian random vectors.

Note that Gaussian random process are sub-Gaussian.
Theorem 5.24 (Dudley's Inequality). Suppose $X_{t}$ with $t \in T$ is mean zero random process on the metric $(T, d)$ and it is sub-Gaussian with some constant $K$. Then

$$
\mathbb{E} \sup _{t \in T} X_{t} \leq C K \sum_{j \in \mathbb{Z}} 2^{-j} \sqrt{\log N\left(T, d, 2^{-j}\right)},
$$

where $\log N\left(T, d, 2^{-j}\right)$ is the metric entropy, i.e., $N\left(T, d, 2^{-j}\right)$ is the smallest number of balls of radius $2^{-j}$ required to cover $T$ in metric $d$.

Remark 5.25. If $T$ is compact, there is $K \in \mathbb{Z}$ such that for all $j \leq K$, we have $N\left(T, d, 2^{-j}\right)=$ 1 such that all summands are zero.

Example 5.26. Let $T=\mathbb{S}^{n-1} \subset \mathbb{R}^{n}$ and let $d$ be the Euclidean metric. For any $\epsilon>0$, we have $N\left(\mathbb{S}^{n-1}, d, \epsilon\right) \leq\left(\frac{3}{\epsilon}\right)^{n-1}$. Then

$$
\begin{aligned}
C \sum_{j \in \mathbb{Z}} 2^{-j} \sqrt{\log N\left(\mathbb{S}^{n-1}, d, 2^{-j}\right)} & =C \sum_{j \geq 0} 2^{-j} \sqrt{(n-1)\left(\log 2^{j} \cdot 3\right)} \\
& =c^{\prime} \sqrt{n} \sum_{j \geq 0} \sqrt{j}^{2^{-j}}=c^{\prime \prime} \sqrt{n}
\end{aligned}
$$

In conclusion, we have $\mathbb{E} \sup X_{t} \leq c \sqrt{n}$ if $X_{t}$ is 1-sub-Gaussian. Dudley's Inequality 5.24 is applicable to random walks on Hamming cube.

## 6 The Semigroup method

The semigroup method is a powerful method to prove interesting "isoperimetric-type" inequalities.

### 6.1 Basic definitions and set up

Definition 6.1 (Markov Process). Consider $X_{t}$ as a random process on time $T \subset \mathbb{R}$. A Markov process is a stochastic process with the property that

$$
\mathbb{P}\left(X_{t_{n}} \leq x_{n} \mid X_{t_{n-1}}, X_{t_{n-2}}, \ldots, X_{t_{1}}\right)=\mathbb{P}\left(X_{t_{n}} \leq x_{n} \mid X_{t_{n-1}}\right)
$$

where $x_{n} \in \mathbb{R}$ and $X_{t} \in \mathbb{R}$, i.e., this stochastic process "does not see the past".
Recall the conditional probability is $\mathbb{P}(A \mid B)=\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$, and the conditional expectation $\mathbb{E}(X \mid Y)$ of random variables $X$ and $Y$ is also a random variable which is the best prediction of $X$ given some behavior of $Y$. Suppose $X$ has density function $f_{X}$ and $Y$ has density function $f_{Y}$. Then $(X, Y)$ has joint density $f_{X Y}$, we can define the conditional density as $f_{X \mid Y}(x \mid y)=\frac{f_{X Y}(x, y)}{f_{Y}(y)}$, so $\mathbb{E}(X \mid Y=y)=\int_{-\infty}^{\infty} x \cdot f_{X \mid Y}(x \mid y) d x$ which is a number that depends on $y$. If let $y$ be vary, then we get a random variable.

Definition 6.2 (Conditional Expectation). Consider $X$ on $\sigma$-algebra $\mathcal{F}_{0}$. Consider $\mathcal{F} \subset \mathcal{F}_{0}$ as another $\sigma$-algebra. The conditional expectation of $X$ with respect to $\mathcal{F}, \mathbb{E}(X \mid \mathcal{F})$, is such an $L_{1}$ random variable/vector that $\mathbb{E}(X \mid \mathcal{F}) \in \mathcal{F}$, so that all events that relates to the random variable are sets of $\sigma$-algebra $\mathcal{F}$, i.e., $\{\mathbb{E}(X \mid \mathcal{F})<t\} \in \mathcal{F}$. And for any event $A \in \mathcal{F}$,

$$
\mathbb{E}\left(X \cdot \mathbb{1}_{A}\right)=\mathbb{E}\left(\mathbb{E}(X \mid \mathcal{F}) \cdot \mathbb{1}_{A}\right), \quad \text { i.e., } \int_{A} X d P=\int_{A} \mathbb{E}(X \mid \mathcal{F}) d P
$$

If $Y$ is another random vector, then the conditional expectation of $X$ with respect to $Y$ is

$$
\mathbb{E}(X \mid Y):=\mathbb{E}(X \mid \sigma(Y))
$$

where $\sigma(Y)$ is the $\sigma$-algebra generated by $Y$.
Definition 6.3 (Alternative Definition of Markov Process). Consider $X_{t} \in \mathbb{R}^{n}$ with $t \geq 0$ as a random process. Assume for any bounded measurable function $f: \mathbb{R}^{n} \rightarrow \mathbb{R}$ and for any times $t, s>0$, there is a bounded measurable function $P_{s} f: \mathbb{R}^{n} \rightarrow \mathbb{R}$ such that

$$
\begin{equation*}
\mathbb{E}\left(f\left(X_{t+s}\right) \mid\left\{X_{z}\right\}_{z \leq t}\right)=P_{s} f\left(X_{t}\right), \tag{61}
\end{equation*}
$$

i.e., the behavior of $X_{t+s}$ only depends on $X_{t}$ and no earlier times.

Homework: find an example of a non-Markov process.
Definition 6.4 (Markov Semigroup). Suppose $P_{s}$ is an operator on bounded measurable functions such that

$$
f: \mathbb{R}^{n} \rightarrow \mathbb{R} \quad \xrightarrow{P_{s}} \quad P_{s} f: \mathbb{R}^{n} \rightarrow \mathbb{R}
$$

as defined in (61). $P_{s}$ is what we called a Markov semigroup.

Definition 6.5 (Stationary measure of a Markov Process). Consider $X_{t}$ on $\mathbb{R}^{n}$ as a Markov process indexed by $T \subset \mathbb{R}^{+}$. A measure $\mu$ on $\mathbb{R}^{n}$ is called a stationary measure of $X_{t}$ if for any bounded measurable function $f: \mathbb{R}^{n} \rightarrow \mathbb{R}$, we have

$$
\int f d \mu=\int P_{t} \cdot f d \mu
$$

Here the semigroup $\left\{P_{t} f\right\}_{t \geq 0}$ is a collection of functions that describes some evolution of a function in time. We say the measure is stationary whatever the evolution it is, the average of the function does not change.

Remark 6.6. Consider initial random vector $X_{0} \sim \mu$ where $\mu$ is a stationary measure. Then

$$
\begin{aligned}
\mathbb{E} f\left(X_{t}\right) & =\mathbb{E}\left(\mathbb{E}\left(f\left(X_{t}\right) \mid X_{0}\right)\right) \\
& =\mathbb{E} P_{t} f\left(X_{0}\right) \\
& =\int P_{t} f d \mu \quad\left(\text { as } X_{0} \sim \mu\right) \\
& =\int f d \mu \quad(\text { by stationary }) \\
& =\mathbb{E} f\left(X_{0}\right) .
\end{aligned}
$$

Hence, for any function $f$ that has a stationary measure, we have $\mathbb{E} f\left(X_{t}\right)=\mathbb{E} f\left(X_{0}\right)$. If $f=\mathbb{1}_{\Omega}$, then $\mathbb{P}\left(X_{t} \in \Omega\right)=\mathbb{P}\left(X_{0} \in \Omega\right)$. In other words, if $X_{0} \sim \mu$ with stationary $\mu$, then for any $t \geq 0, X_{t} \sim \mu$.

### 6.2 Properties

Lemma 6.7. Let $\mu$ be a stationary measure of a Markov process $X_{t}$ indexed by $t \geq 0$. Then the following hold for all $p \geq 1, \alpha, \beta \in \mathbb{R}$, bounded measurable function $f, g$ :

1. $\left\|P_{t} f\right\|_{L^{p}(\mu)} \leq\|f\|_{L^{p}(\mu)}$ (contraction).
2. $P_{t}$ is a linear operator, i.e. $P_{t}(\alpha f+\beta g)=\alpha P_{t} f+\beta P_{t} g$ (linearity).
3. $P_{t+s} f=P_{t} P_{s} f \quad \mu$-a.s. (semigroup property).
4. $P_{t} 1=1 \quad \mu$-a.s. (conservativeness).

Proof. 1. Suppose $X_{0} \sim \mu$, we have

$$
\left.\left.\int\left(P_{t} f\right)^{p} d \mu=\mathbb{E}_{\mu}\left(\mathbb{E}\left(f\left(X_{t}\right) \mid X_{0}\right)^{p}\right)\right) \leq \mathbb{E}_{\mu}\left(\mathbb{E}\left(f\left(X_{t}\right)^{p} \mid X_{0}\right)\right)\right)=\int(f)^{p} d \mu
$$

where we have used Jensen's inequality.
2. Linearity follows similarly as
$P_{t}(\alpha f+\beta g)=\mathbb{E}\left(\alpha f\left(X_{t}\right)+\beta g\left(X_{t}\right) \mid X_{0}\right)=\alpha \mathbb{E}\left(f\left(X_{t}\right) \mid X_{0}\right)+\beta \mathbb{E}\left(g\left(X_{t}\right) \mid X_{0}\right)=\alpha P_{t} f+\beta P_{t} g$,
where we have used the linearity of conditional expectation.
3. For the semigroup property, we have

$$
P_{t+s} f=\mathbb{E}\left(f\left(X_{t+s}\right) \mid X_{0}\right)=\mathbb{E}\left(\mathbb{E}\left(f\left(X_{t+s}\right) \mid\left\{X_{r}\right\}_{r \leq t}\right) \mid X 0\right)=\mathbb{E}\left(P_{s} f\left(X_{t}\right) \mid X_{0}\right)=P_{t} P_{s} f,
$$

where the third equality is based on markovianity.
4. Conservativeness follows $P_{s} 1=\mathbb{E}\left(1 \mid X_{0}\right)=1$.

Remark 6.8. If $X_{t}$ has a stationary measure $\mu$, then the above lemma is true for all $f \in$ $L^{1}(\mu)$, not only for bounded measurable functions $f$. From now on, we will assume the $P_{t} f$ is defined in this manner for every $f \in L^{1}(\mu)$.

Remark 6.9. Not every Markov process has a stationary measure; the questions of existence are complicated. We will consider concrete examples when things work well, the stationary measure exists and has nice properties, and we will explain the existence in these examples.

Definition 6.10 (Variance). Let $\mu$ be a stationary measure and $f \in L^{2}(\mu)$. The variance is defined as

$$
\begin{equation*}
\operatorname{Var}_{\mu}(f):=\int f^{2} d \mu-\left(\int f d \mu\right)^{2}=\mathbb{E}_{\mu} f^{2}-\left(\mathbb{E}_{\mu} f\right)^{2} \tag{62}
\end{equation*}
$$

Note, by Cauchy's inequality, that $\operatorname{Var}_{\mu}(f) \geq 0$.
Lemma 6.11. Let $\mu$ be a stationary measure of a Markov process $X_{t}$ and $f \in L^{2}(\mu)$. Then $\operatorname{Var}_{\mu}(f)$ decreases in $t \geq 0$

Proof. Note that

$$
\begin{align*}
\operatorname{Var}_{\mu}\left(P_{t} f\right) & =\int\left(P_{t} f-\int P_{t} f d \mu\right)^{2} d \mu=\int\left(P_{t} f-\int f d \mu\right)^{2} d \mu=\int\left(P_{t}\left(f-\int f d \mu\right)\right)^{2} d \mu  \tag{63}\\
& \leq \int\left(f-\int f d \mu\right)^{2} d \mu=\operatorname{Var}_{\mu}(f) \tag{64}
\end{align*}
$$

Here we use the definition of stationary measure for the second equality, linearity and conservativeness for the third equality and contraction for the fourth. Also we have

$$
\begin{equation*}
\int\left(P_{t}\left(f-\int f d \mu\right)\right)^{2} d \mu=\int\left(P_{t-s} P_{s}\left(f-\int f d \mu\right)\right)^{2} d \mu \leq \operatorname{Var}_{\mu}\left(P_{s} f\right) \tag{65}
\end{equation*}
$$

Hence, we get $\operatorname{Var}_{\mu}\left(P_{t} f\right) \leq \operatorname{Var}_{\mu}\left(P_{s} f\right)$ for all $t \geq s \geq 0$.

Remark 6.12. $\operatorname{Var}_{\mu}(f)$ measures how far is $f$ from a constant, as

$$
\operatorname{Var}_{\mu}(C)=\int C^{2} d \mu-\left(\int C d \mu\right)^{2}=C^{2}-C^{2}=0
$$

Therefore, variance is a measure of the distance of the function to a constant function. The fact that it decreases along the semi-group means that the function $P_{t} f$ becomes closer and closer to a constant function - namely, to the function $\int f d \mu$ (since this integral is preserved). Soon we will see that in some nice situations, not only does the variance decrease, but it decreases all the way to zero, and $P_{t} f \underset{t \rightarrow \infty}{\rightarrow} \int f d \mu$; however, this is not necessarily the case for an arbitrary Markov process.

Definition 6.13 (Generator of Markov process). Given a Markov process $X_{t}$ with stationary measure $\mu$ on $\mathbb{R}^{n}$. For every $f \in L^{2}(\mu)$, the generator $\mathscr{L}$ is defined as

$$
\begin{equation*}
\mathscr{L} f:=\lim _{t \rightarrow 0} \frac{P_{t} f-f}{t} \tag{66}
\end{equation*}
$$

Here, $\mathscr{L}$ is an operator on functions from $L^{2}(\mu)$ for which this limit makes sense. $\mathscr{L}$ is called the Generator associated with $X_{t}$.

Remark 6.14. $\mathscr{L}$ is a linear operator, since $P_{t} f$ is linear.
Remark 6.15 (Important). One can in fact define the Markov semigroup using a given linear operator $\mathscr{L}$. Indeed, $\frac{d}{d t} P_{t} f=\lim _{\delta \rightarrow 0} \frac{P_{t+\delta} f-P_{t} f}{\delta}=\lim _{\delta \rightarrow 0} P_{t}\left(\frac{P_{\delta} f-f}{\delta}\right)=P_{t} \mathscr{L}$ f. Equivalently, $\frac{d}{d t} P_{t} f=\lim _{\delta \rightarrow 0} \frac{P_{s} P_{t} f-P_{t} f}{\delta}=\mathscr{L} P_{t} f$.
Corollary 6.16.

$$
\begin{equation*}
\mathscr{L} P_{t} f=P_{t} \mathscr{L} f \tag{67}
\end{equation*}
$$

Consider the following PDE:

$$
\left\{\begin{array}{l}
\frac{d}{d t}\left(P_{t} f\right)=\mathscr{L}\left(P_{t} f\right) \\
P_{0} f=f
\end{array}\right.
$$

when $\mathscr{L}$ is fixed, $P_{t} f$ can then be defined as a solution of the above PDE. When the linear operator $\mathscr{L}$ is such that the PDE methods allow to conclude existence, probabilistic methods also provide a way of solving this PDE.

Example 6.17 (Finite state space). Let $\left(X_{t}\right)_{t \in \mathbb{R}_{+}}$be a Markov process with values in a finite state space $X_{t} \in\{1, \ldots, d\}$. Such processes are typically described in terms of their transition rates $\lambda_{i j} \geq 0$ for $i \neq j$ :

$$
\mathbb{P}\left[X_{t+\delta}=j \mid X_{t}=i\right]=\lambda_{i j} \delta+o(\delta) \quad \text { for } i \neq j
$$

Evidently, the transition rates $\lambda_{i j}$ describe the infinitesimal rate of growth of the probability of jumping from state $i$ to state $j$ (informally if $X_{t}=i$ then the probability that $X_{t+d t}=j$ is $\lambda_{i j} d t$ ).

Let us organize the transition probabilities $q_{t i j}=\mathbb{P}\left[X_{t}=j \mid X_{0}=i\right]$ and rates $\lambda_{i j}$ into matrices $Q_{t}=\left(q_{t i j}\right)_{1 \leq i, j \leq d}$ and $\Lambda=\left(\lambda_{i j}\right)_{1 \leq i, j \leq d}$ respectively where we define the diagonal entries of $\Lambda$ as $\lambda_{i i}=-\sum_{j \neq i} \lambda_{i j} \leq 0$. Then

$$
\lim _{t \rightarrow 0} \frac{q_{t i j}-q_{0 i j}}{t}=\lambda_{i j}
$$

for every $1 \leq i, j \leq d$ (the diagonal entries $\lambda_{i i}$ were chosen precisely to enforce the law of total probability $\sum_{j} q_{t i j}=1$ ). In particular, we have

$$
\mathscr{L} f(i)=\lim _{t \rightarrow 0} \frac{1}{t} \sum_{j=1}^{d} \lambda_{i j} f(j)=(\Lambda f)_{i}
$$

where we identify the function $f$ with the vector $(f(1), \ldots, f(d)) \in \mathbb{R}^{d}$. We therefore conclude that the generator of a Markov process in a finite state space corresponds precisely to the matrix of transition rates. The Kolmogorov equation now reduces to the matrix differential equation (semigroup)

$$
\frac{d}{d t} Q_{t}=Q_{t} \Lambda, \quad Q_{0}=I
$$

This differential equation is the basic tool for computing probabilities of finite state space Markov processes. The solution is in fact easily obtained as $Q_{t}=e^{t \Lambda}$ from which we readily see why $P_{t}$ and $L$ must commute.

Example 6.18 (Heat semigroup). Suppose the generator $\mathscr{L}=\Delta$ be the Laplace operator and $u: \mathbb{R}^{n} \rightarrow R$.

$$
\begin{equation*}
\Delta u=\sum_{i=1}^{n} \partial_{i i} u=\operatorname{tr}\left(\nabla^{2} u\right) \tag{68}
\end{equation*}
$$

Here, $\partial_{i i} u$ is the partial derivative and $\nabla^{2} u$ is the Hessian matrix. We have Dom $=$ $\overline{L^{2}\left(\mathbb{R}^{n}\right) \cap C^{2}\left(\mathbb{R}^{n}\right)}$. Pick $f \in \operatorname{Dom}$, the heat semigroup is defined as

$$
\left\{\begin{array}{l}
\partial_{t}\left(P_{t} f\right)=\Delta\left(P_{t} f\right), \\
P_{0} f=f
\end{array}\right.
$$

There exists a solution and this defines a Markov process. Then, what is the invariant measure? We have the condition $\int_{\mathbb{R}^{n}} P_{t} f d \mu=\int_{\mathbb{R}^{n}} f d \mu$, i.e., $\frac{d}{d t}\left(\int_{\mathbb{R}^{n}} P_{t} f d \mu\right)=0=\int_{\mathbb{R}^{n}} \frac{d}{d t}\left(\int_{\mathbb{R}^{n}} P_{t} f d \mu=\right.$ $\int_{\mathbb{R}^{n}} \Delta\left(P_{t} f\right) d \mu$.

For Lebesgue measure on $\mathbb{R}^{n}$, for all $g \in C^{2}\left(\mathbb{R}^{n}\right)$ such that $\Delta g \in L^{1}\left(\mathbb{R}^{n}\right)$, we have $\int_{\mathbb{R}^{n}} \Delta g d x=0$. Indeed, recall Green's formula: for measurable $u$, $v$, we have

$$
\int u \Delta v d x=-\int\langle\nabla u, \nabla v\rangle d x
$$

By taking $u=1$ and $\nabla u=0$, we got $\int \nabla v d x=0$. The conclusion is: Lebesgue measure is stationary for the heat semigroup.

Example 6.19 (Heat semigroup on the circle/torus). Consider a nice enough one dimensional function $f:[-\pi, \pi] \rightarrow \mathbb{R}$ (2 $\pi$ periodic). Then we have the PDE

$$
\left\{\begin{array}{l}
\partial_{t}\left(P_{t} f\right)=\Delta\left(P_{t} f\right) \\
P_{0} f=f
\end{array}\right.
$$

Then the stationary measure is uniform on circle.
Example 6.20 (Main example: Ornstein-Uhlenbeck semigroup). Firstly, Ornstein-Uhlenbeck operator is defined as a second order linear operator on nice enough function on $\mathbb{R}^{n}$ following:

$$
\mathscr{L} u=\Delta u-\langle\nabla u, x\rangle
$$

Ornstein-Uhlenbeck semigroup is defined as: $\frac{d}{d t}\left(P_{t} f\right)=\mathscr{L}\left(P_{t} f\right), P_{0} f=f$. The question here is: What is the stationary measure for this process? Observe that

$$
\begin{align*}
& \int P_{t} f d \mu=\int f d \mu,  \tag{69}\\
& \frac{d}{d t} \int P_{t} f d \mu=0,  \tag{70}\\
& \int \frac{d}{d t}\left(P_{t} f\right) d \mu=\int \mathscr{L}\left(P_{t} f\right) d \mu=0 . \tag{71}
\end{align*}
$$

The answer is the Gaussian measure!!!

$$
\begin{align*}
& d \gamma=\frac{1}{\sqrt{2 \pi}^{n}} \exp -\frac{|x|^{2}}{2} d x  \tag{72}\\
& \int u \mathscr{L} v d \gamma=-\int\langle\nabla u, \nabla v\rangle d \gamma \tag{73}
\end{align*}
$$

for nice enough $u, v$. Plug in $u=1$ and $\int \mathscr{L} v d \gamma=0$, we have $\left.\int \mathscr{L}\left(P_{t} f\right)\right) d \mu=0$. Hence we got the conclusion: Gaussian measure is stationary for the Ornstein-Uhlenbeck process. The proof of 73 is left as homework (use first order gaussian integration by parts twice). It could also be proved by using Green's formula.
Definition 6.21 (Reversibility of semigroups). A Markov semigroup $P_{t}$ with stationary measure $\mu$ is called reversible if:

$$
\int f \cdot P_{t} g d \mu=\int P_{t} f \cdot g d \mu
$$

for all $f, g \in \operatorname{Dom}\left(P_{t}\right)$.
The name of reversibility indicates that if we assume our Markov process $X_{t}$ is such that $X_{0} \sim \mu$ as we usually do, then

$$
\mathbb{E}\left(f\left(X_{0}\right) \mathbb{E}\left(g\left(X_{t}\right) \mid X_{0}\right)\right)=\mathbb{E}\left(g\left(X_{t}\right) \mathbb{E}\left(f\left(X_{0}\right) \mid X_{t}\right)\right)
$$

where $\mathbb{E}\left(f\left(X_{0}\right) \mathbb{E}\left(g\left(X_{t}\right) \mid X_{0}\right)\right)=\int f P_{t} g d \mu$, and $\mathbb{E}\left(g\left(X_{t}\right) \mathbb{E}\left(f\left(X_{0}\right) \mid X_{t}\right)\right)=\int P_{t} f g d \mu$. One can show that, left as homework, $P_{t} f(x)=\mathbb{E}\left(f\left(X_{t}\right) \mid X_{0}=x\right)=\mathbb{E}\left(f\left(X_{0}\right) \mid X_{t}=x\right)$. That is to say, time goes in both directions in the same way.

Definition 6.22 (Ergodicity). A Markov semigroup semigroup $P_{t}$ is called ergodic if for all $f \in \operatorname{Dom}\left(P_{t}\right)$,

$$
P_{t} f \underset{t \rightarrow \infty}{\stackrel{L_{2}}{\rightarrow}} \mathbb{E}_{\mu} f=\int f d \mu
$$

Here, $P_{t} f \underset{t \rightarrow \infty}{\stackrel{L_{2}}{\rightarrow}}=C=\mathbb{E}_{\mu} f$, since $\int P_{t} f d \mu=\int f d \mu$ for all t. In particular, $\int c d \mu=$ $\int P_{\infty} f d \mu=\int f d \mu$.
Remark 6.23. Recall that for all Markov semigroup, the $\operatorname{Var}\left(P_{t} f\right)$ decreases as $t \rightarrow \infty$. If $P_{t}$ is ergodic, that is equivalent to $\operatorname{Var}\left(P_{t} f\right) \rightarrow 0$.

Definition 6.24 (Dirichlet form of semigroups). Assume $X_{t}$ is a Markov process with a stationary measure $\mu$ and generator $\mathscr{L}$, i.e., $\partial_{t}\left(P_{t} f\right)=\mathscr{L} P_{t} f$, and $P_{0} f=f$. The Dirichlet form $\mathcal{E}$ is defined by:

$$
\mathcal{E}(f, g):=-\int f \mathscr{L} g d \mu
$$

Example 6.25 (Dirichlet form of Heat semigroup). Recall that Heat semigroup is $\mathscr{L}=\nabla$ on $\mathbb{R}^{n}$, then $\partial_{t}\left(P_{t} f\right)=\nabla\left(P_{t} f\right)$ and $P_{0} f=f$, and the stationary measure $\mu$ is the Lebesgue measure because it satisfies the integral by parts: $\int u \Delta v d x=-\int\langle\nabla u, \nabla v\rangle d x$, in particular, $\int \nabla v d x=0$, then we have:

$$
\mathcal{E}(f, g)=-\int f \nabla g d x=\int\langle\nabla u, \nabla v\rangle d x
$$

Remark 6.26. Notice that one can show the Heat semigroup is reversible: given $f$ and $g$ : $\int P_{t} f g d \mu=\int g P_{t} f d \mu$.

Proof left as homework.
Theorem 6.27 (Abstract Theorem). $P_{t}$ is a reversible and ergodic markov semigroup with stationary measure $\mu$, fix a constant $c \geq 0$, the followings are all equivalent:

1. For all $f \in \operatorname{Dom}\left(P_{t}\right), \operatorname{Var}_{\mu}(f) \leq c \mathcal{E}(f, f)$ (Poincare inequality).
2. $P_{t}$ is "Hypercontractive". That is,

$$
\int_{\mathbb{R}^{n}}\left(P_{t} f-\int_{\mathbb{R}^{n}} f d \mu\right)^{2} d \mu \leq e^{-\frac{t}{c}} \int_{\mathbb{R}^{n}}\left(f-\int_{\mathbb{R}^{n}} f d \mu\right)^{2} d \mu
$$

for all $t \geq 0$ and $f \in \operatorname{Dom}\left(P_{t}\right)$. This is the same as $\operatorname{Var}_{\mu}\left(P_{t} f\right) \leq e^{\frac{t}{c}} \operatorname{Var}_{\mu}(f)$.
3. $\mathcal{E}\left(P_{t} f, P_{t} f\right) \leq e^{-\frac{2 t}{c}} \mathcal{E}(f, f)$.
4. For all $f$, there exists a constant $\kappa(f)$ such that $\sqrt{\operatorname{Var}_{\mu}\left(P_{t} f\right)} \leq \kappa(f) e^{-\frac{t}{c}}$.
5. for all $f$, there exists a constant $k(f)$ such that $\mathcal{E}\left(P_{t} f, P_{t} f\right) \leq k(f) e^{-\frac{2 t}{c}}$

Before the proof, we need to cover the following example.
Example 6.28 (Ornstein-Uhlenbeck semigroup). Recall $\mathscr{L} u=\Delta u-\langle\nabla u, x\rangle$, where $u: \mathbb{R}^{n} \rightarrow$ $\mathbb{R}$ is an appropriate function here. The semigroup $P_{t}$ is defined as: $\partial_{t}\left(P_{t} f\right)=\mathscr{L}\left(P_{t} f\right)$, $P_{0} f=f$, given $f \in \operatorname{Dom}\left(P_{t}\right)$.

Lemma 6.29. A very nice and concrete representation for the Ornstein-Uhlenbeck semigroup:

1. $P_{t} f(x)=\mathbb{E} f\left(e^{-t} x+\sqrt{1-e^{-2 t}} Z\right)$, where $Z \sim \mathcal{N}(0, I d)$ is the Ornstein-Uhlenbeck semigroup, such that it satisfies $\partial_{t}\left(P_{t} f\right)=\mathscr{L}\left(P_{t} f\right), P_{0} f=f$, which is true because we can check: when $t=0, P_{0} f=\mathbb{E} f(x)$, when $t \rightarrow \infty, P_{\infty} f=\mathbb{E} f(z)=\int f d \gamma$.
2. $P_{t}$ is ergodic.
3. $P_{t}$ is reversible.
4. $\gamma$ is the stationary measure.

Claim 6.30. This claim of second order integration by parts is used in the following proof. For all $f, g$,

$$
\int g \mathscr{L} f d \gamma=-\int\langle\nabla f, \nabla g\rangle d \gamma
$$

In particular, the Dirichlet form is given by $\mathcal{E}(f, g)=-\int\langle\nabla f, \nabla g\rangle d \gamma$.
Proof of Claim 6.30. By Green's formula, we have

$$
\begin{aligned}
\int_{\mathbb{R}^{n}} g \Delta f d \gamma & =c_{n} \int_{\mathbb{R}^{n}} \Delta f \cdot g e^{\frac{-|x|^{2}}{2}} d x \\
& =-c_{n} \int_{\mathbb{R}^{n}}\left\langle\nabla f, \nabla\left(g e^{\frac{-|x|^{2}}{2}}\right)\right\rangle d x \\
& =-\int_{\mathbb{R}^{n}}\langle\nabla f, \nabla g\rangle c_{n} e^{\frac{-|x|^{2}}{2}} d x+\int_{\mathbb{R}^{n}} g c_{n}\langle\nabla f, x\rangle e^{\frac{-|x|^{2}}{2}} d x \\
& =-\int_{\mathbb{R}^{n}}\langle\nabla f, \nabla g\rangle d \gamma+\int_{\mathbb{R}^{n}} g\langle\nabla f, x\rangle d \gamma .
\end{aligned}
$$

Therefore,

$$
\int_{\mathbb{R}^{n}} g \mathscr{L} f d \gamma=\int_{\mathbb{R}^{n}} g \Delta f d \gamma-\int_{\mathbb{R}^{n}} g\langle\nabla f, x\rangle d \gamma=-\int_{\mathbb{R}^{n}}\langle\nabla f, \nabla g\rangle d \gamma .
$$

Proof of Lemma 6.29. Proof of Property 4. Let $f \in \operatorname{Dom}\left(P_{t}\right)$. Then,

$$
\frac{d}{d t} \int P_{t} f d \gamma=\int \partial_{t}\left(P_{t} f\right) d \gamma=\int \mathscr{L}\left(P_{t} f\right) d \gamma
$$

Here, if we let $g=1$, then

$$
\int \mathscr{L} f d \gamma=-\int\langle\nabla f, 0\rangle d \gamma=0
$$

Thus, $\frac{d}{d t} \int P_{t} f d \gamma=0$, implying that $\int P_{t} f d \gamma=\int f d \gamma$ and $P_{t}$ is constant in time.
Proof of Property 1. We want to show that $P_{t} f(x)=\mathbb{E} f\left(e^{-t} x+\sqrt{1-e^{-2 t}} z\right)$ satisfies the following

$$
\left\{\begin{array}{l}
\partial_{t}\left(P_{t} f\right)=\mathscr{L}\left(P_{t} f\right) \\
P_{0} t=f
\end{array}\right.
$$

By direct calculation, we see that $P_{0} f(x)=\mathbb{E} f(x)=f(x)$. Next, by the chain rule, we have

$$
\begin{aligned}
\partial_{t}\left(P_{t} f\right) & =\partial_{t} \mathbb{E} f\left(e^{-t} x+\sqrt{1-e^{-2 t}} Z\right) \\
& =\mathbb{E} \partial_{t} f\left(e^{-t} x+\sqrt{1-e^{-2 t}} Z\right) \\
& =\mathbb{E}\left\langle\nabla f\left(e^{-t} x+\sqrt{1-e^{-2 t}} Z\right),-e^{-t} x+\left(1-e^{-2 t}\right)^{-1 / 2} e^{-2 t} Z\right\rangle \\
& =\int_{\mathbb{R}^{n}}\left\langle\nabla f\left(e^{-t} x+\sqrt{1-e^{-2 t}} z\right),-e^{-t} x+\left(1-e^{-2 t}\right)^{-1 / 2} e^{-2 t} z\right\rangle d \gamma(z) .
\end{aligned}
$$

Using Claim 6.30, we see that

$$
\begin{aligned}
& \int_{\mathbb{R}^{n}}\left\langle\nabla f\left(e^{-t} x+\sqrt{1-e^{-2 t}} z\right),-e^{-t} x+\left(1-e^{-2 t}\right)^{-1 / 2} e^{-2 t} z\right\rangle d \gamma(z) \\
& =\int_{\mathbb{R}^{n}}\left\langle\nabla f\left(e^{-t} x+\sqrt{1-e^{-2 t}} z\right),-e^{-t} x\right\rangle d \gamma(z)+\int_{\mathbb{R}^{n}} e^{-2 t} \Delta f\left(e^{-t} x+\sqrt{1-e^{-2 t}} z\right) d \gamma(z) \\
& =-\left\langle\nabla_{x}\left[\int_{\mathbb{R}^{n}} f\left(e^{-t} x+\sqrt{1-e^{-2 t}} z\right) d \gamma(z)\right], x\right\rangle+\Delta_{x}\left[\int_{\mathbb{R}^{n}} f\left(e^{-t} x+\sqrt{1-e^{2 t}} z\right) d \gamma(z)\right] \\
& =\mathscr{L} \mathbb{E} f\left(e^{-t} x+\sqrt{1-e^{-2 t}} Z\right) .
\end{aligned}
$$

Proof of Property 3. For $t=0$, note that

$$
\int_{\mathbb{R}^{n}} f\left(P_{0} g\right) d \gamma=\int_{\mathbb{R}^{n}} f g d \gamma=\int_{\mathbb{R}^{n}} g\left(P_{0} f\right) d \gamma
$$

Furthermore, applying Gaussian integration by parts, we see that

$$
\frac{d}{d t} \int_{\mathbb{R}^{n}} f P_{t} g d \gamma=\int_{\mathbb{R}^{n}} f \mathscr{L}\left(P_{t} g\right) d \gamma=-\int_{\mathbb{R}^{n}}\left\langle\nabla f, \nabla\left(P_{t} g\right)\right\rangle d \gamma
$$

Similarly,

$$
\frac{d}{d t} \int g P_{t} f d \gamma=-\int\left\langle\nabla g, \nabla P_{t} f\right\rangle d \gamma
$$

Now, via a change of variable, we see that

$$
\begin{aligned}
\int_{\mathbb{R}^{n}}\left\langle\nabla f, \nabla\left(P_{t} g\right)\right\rangle d \gamma & =\int_{\mathbb{R}^{n}} \int_{\mathbb{R}^{n}} e^{-t}\left\langle\nabla f(x), \nabla g\left(e^{-t} x+\sqrt{1-e^{-2 t}} z\right)\right\rangle d \gamma(z) d \gamma(x) \\
& =\int_{\mathbb{R}^{n}} \int_{\mathbb{R}^{n}} c_{n} \frac{e^{-t}}{\sqrt{1-e^{-2 t}}} e^{-\frac{\left(\xi-e^{-t} x x^{2}\right.}{2\left(1-e^{-2 t}\right)}}\langle\nabla f(x), \nabla g(\xi)\rangle d \xi d \gamma(x) \\
& =\int_{\mathbb{R}^{n}} \int_{\mathbb{R}^{n}} e^{-t}\left\langle\nabla g(x), \nabla f\left(e^{-t} x+\sqrt{1-e^{-2 t}} z\right)\right\rangle d \gamma(z) d \gamma(x) \\
& =\int_{\mathbb{R}^{n}}\left\langle\nabla g, \nabla\left(P_{t} f\right)\right\rangle d \gamma
\end{aligned}
$$

Hence,

$$
\left\langle f, P_{t} g\right\rangle_{L^{2}(\gamma)}=\int_{0}^{t} \frac{d}{d \tau}\left[\left\langle f, P_{\tau} g\right\rangle_{L^{2}(\gamma)}\right] d \tau=\int_{0}^{t} \frac{d}{d \tau}\left[\left\langle g, P_{\tau} f\right\rangle_{L^{2}(\gamma)}\right] d \tau=\left\langle P_{t} f, g\right\rangle_{L^{2}(\gamma)}
$$

Therefore, $P_{t}$ follows accordingly.
Proof of Property 2. Indeed, by the Bounded Convergence Theorem, note that

$$
P_{t} f=\mathbb{E} f\left(e^{-t}+\sqrt{1-e^{-2 t}} Z\right) \underset{t \rightarrow \infty}{\stackrel{L_{2}}{\longrightarrow}} \mathbb{E} f(Z) .
$$

Remark 6.31. For the students who have background in stochastic partial differential equation, when $X_{t}$ is a Markov process, $d X_{t}=-X_{t} d t+\sqrt{2} d B_{t}$ where $B_{t}$ is Brownian motion. Then $X_{t}$ is the Ornstein-Uhlenbeck semigroup.
Lemma 6.32 (Hypercontactivity). Let $\left\{P_{t}\right\}$ be the Ornstein-Ulhenbeck semigroup. Then, the following hold:

1. $\nabla P_{t} f=e^{-t} P_{t} \nabla f$, where $P_{t} \nabla f=\left(P_{t} \partial_{1} f, \ldots, P_{t} \partial_{n} f\right)$. Here the $\partial_{i} f$ denote the partial derivatives of $f$ and $\nabla f=\left(\partial_{1} f, \ldots, \partial_{n} f\right)$.
2. $\int\left|\nabla P_{t} f\right|^{2} d \gamma \leq e^{-2 t} \int|\nabla f|^{-2} d \gamma$.

Proof. 1. Recall that $P_{t} f=\mathbb{E} f\left(e^{-t} x+\sqrt{1-e^{-2 t}} Z\right)$. Then, the chain rule yields that

$$
\partial_{i}\left(P_{t} f\right)=\mathbb{E} \partial_{x^{i}} f\left(e^{-t} x+\sqrt{1-e^{-2 t}} Z\right)=e^{-t} \mathbb{E} \partial_{i} f\left(e^{-t} x+\sqrt{1-e^{-2 t}} Z\right)=e^{-t} P\left(\partial_{i} f\right)
$$

2. Note that

$$
\begin{aligned}
\int_{\mathbb{R}^{n}}\left|\nabla P_{t} f\right|^{2} d \gamma=e^{-2 t} \int_{\mathbb{R}^{n}} \sum_{1 \leq i \leq n}\left|P_{t} \partial_{i} f\right|^{2} d \gamma & =e^{-2 t} \sum_{1 \leq i \leq n} \int\left|P_{t} \partial_{i} f\right|^{2} d \gamma \\
& \leq e^{-2 t} \sum_{1 \leq i \leq n} \int_{\mathbb{R}^{n}}\left|\partial_{i} f\right|^{2} d \gamma \\
& =e^{-2 t} \int_{\mathbb{R}^{n}}|\nabla f|^{2} d \gamma
\end{aligned}
$$

where we have used $\int\left|P_{t} g\right|^{2} d \gamma \leq \int g^{2} d \gamma$.

Theorem 6.33 (Gaussian Poincaré Inequality). Let $f \in \mathscr{L}^{2}\left(\mathbb{R}^{n}\right) \cap \mathcal{C}^{2}\left(\mathbb{R}^{n}\right)$. Then,

$$
\operatorname{Var}_{\gamma}(f)=\int_{\mathbb{R}^{n}} f^{2} d \gamma-\left(\int_{\mathbb{R}^{n}} f d \gamma\right)^{2} \leq \mathbb{E}_{\gamma}|\nabla f|^{2}=\int_{\mathbb{R}^{n}}|\nabla f|^{2} d \gamma,
$$

where $\gamma$ is Gaussian measure.
Proof. Let $P_{t}$ be the Ornstein-Uhlenbeck semigroup. Then, note that $\int_{\mathbb{R}^{n}} f^{2} d \gamma=\int\left(P_{0} f\right)^{2} d \gamma$ and $\left(\int_{\mathbb{R}^{n}} f d \gamma\right)^{2}=\left(P_{\infty} f\right)^{2}$ since ergodicity of $\left\{P_{t}\right\}$ implies that $P_{\infty} f=\int f d \gamma$. Since $\gamma$ is a probability measure, the Fundamental Theorem of Calculus implies that

$$
\begin{aligned}
\operatorname{Var}_{\gamma}(f)=\int_{\mathbb{R}^{n}}\left(P_{0} f\right)^{2} d \gamma-\int_{\mathbb{R}^{n}}\left(P_{\infty} f\right)^{2} d \gamma & =\int_{\mathbb{R}^{n}}\left(P_{0} f\right)^{2}-\left(P_{\infty} f\right)^{2} d \gamma \\
& =-\int_{\mathbb{R}^{n}} \int_{0}^{\infty} \frac{\partial}{\partial t}\left(P_{t} f\right)^{2} d t d \gamma(x) \\
& =-\int_{0}^{\infty} \int_{\mathbb{R}^{n}} \frac{\partial}{\partial t}\left(P_{t} f\right)^{2} d \gamma(x) d t \\
& =-\int_{0}^{\infty} \int_{\mathbb{R}^{n}} 2 P_{t} f \cdot \mathscr{L}\left(P_{t} f\right) d \gamma(x) d t \\
& =-\int_{0}^{\infty} \int_{\mathbb{R}^{n}} 2 P_{t} f \cdot \mathscr{L}\left(P_{t} f\right) d \gamma(x) d t
\end{aligned}
$$

Applying Gaussian integration by parts, we obtain

$$
\begin{aligned}
-2 \int_{0}^{\infty} \int_{\mathbb{R}^{n}} P_{t} f \cdot \mathscr{L}\left(P_{t} f\right) d \gamma(x) d t & =2 \int_{0}^{\infty} \int_{\mathbb{R}^{n}}\left\langle\nabla P_{t} f, \nabla P_{t} f\right\rangle d \gamma d t \\
& =2 \int_{0}^{\infty} \int_{\mathbb{R}^{n}} e^{-2 t}\left|P_{t} \nabla f\right|^{2} d \gamma d t \\
& \leq 2 \int_{0}^{\infty} \int_{\mathbb{R}^{n}} e^{-2 t}|\nabla f|^{2} d \gamma d t \\
& =\int_{\mathbb{R}^{n}}|\nabla f|^{2} d \gamma
\end{aligned}
$$

Remark 6.34. When is the Gaussian Poincaré inequality sharp? - It is when we use the hypercontactivity. The step we used $\int\left|P_{t} g\right|^{2} \leq \int g^{2} d \gamma$ was the only step we used the inequality. In hypercontactivity, the equality is achieved when $g$ is a constant, that is to say, $P_{t} g=g$ is constant. That is, when all of $\partial_{i} f$ in Gaussian Poincaré inequality are constant or $f$ is a linear function. $f(x)=\langle x, \theta\rangle$ where $\theta \in \mathbb{R}^{n}$.

Indeed,

$$
\operatorname{Var}_{\gamma}\langle x, \theta\rangle=\sum \theta_{i}^{2} \operatorname{Var}_{\gamma}\left(x_{i}\right)=|\theta|^{2}=\int_{\mathbb{R}^{n}}|\nabla\langle x, \theta\rangle|^{2} d \gamma
$$

## References

[1] S. Artstein-Avidan, A. Giannopoulos, V. D. Milman, Asymptotic Geometric Analysis, part I, 2015.
[2] N. Alon, B. Klartag, Optimal compression of approximate inner products and dimension reduction, Symposium on Foundations of Computer Science (FOCS 2017), 639-650.
[3] Z. D. Bai and Y. Q. Yin, Necessary and sufficient conditions for almost sure convergence of the largest eigenvalue of a Wigner matrix, Ann. Probab. 16 (1988), no. 4, 1729-1741. MR0958213
[4] Z. D. Bai,Y. Q. Yin, Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix, Ann. Probab. 21 (1993), 1275-1294.
[5] J. Bourgain, V. H. Vu and P. M. Wood, On the singularity probability of discrete random matrices, J. Funct. Anal. 258 (2010), no. 2, 559-603. MR2557947
[6] N. Cook, Lower bounds for the smallest singular value of structured random matrices, Ann. Probab. 46 (2018), no. 6, 3442-3500.
[7] R. Durrett, Probability: Theory and examples, 2019.
[8] A. Edelman, Eigenvalues and condition numbers of random matrices, SIAM J. Matrix Anal. Appl. 9 (1988), 543-560.
[9] C. G. Esseen, On the Kolmogorov-Rogozin inequality for the concentration function, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 5 (1966), 210-216
[10] O. N. Feldheim and S. Sodin, A universality result for the smallest eigenvalues of certain sample covariance matrices, Geom. Funct. Anal. 20 (2010), no. 1, 88-123. MR2647136
[11] Y. Gordon, Some inequalities for Gaussian processes and applications, Israel J. Math. 50 (1985), 265-289.
[12] O. Guedon, A. Litvak, K. Tatarko, Random polytopes obtained by matrices with heavy tailed entries, Commun. Contemp. Math. 22 (2020), no. 4, 1950027, 28 pp.
[13] R. van Handel, R. Latala, P. Youssef, The dimension-free structure of nonhomogeneous random matrices, Invent. Math. 214 (2018), no. 3, 1031-1080.
[14] V. Jain, S. Silwal. A note on the universality of ESDs of inhomogeneous random matrices, arXiv preprint arXiv:2006.05418 (2020).
[15] J. Kahn, J. Komlos, E. Szemeredi, On the probability that a random $\pm 1$ matrix is singular, J. Amer. Math. Soc. 8 (1995), 223-240.
[16] R. Kannan, S. Vempala, Sampling Lattice Points, Proc. 29th ACM Symposium on the Theory of Computing (STOC '97), El Paso, (1997), Invited for publication in Journal of Comp. and System Sciences.
[17] B. Klartag, A central limit theorem for convex sets, Invent. Math., Vol. 168, (2007), 91-131.
[18] B. Klartag, G. V. Livshyts, The lower bound for Koldobsky's slicing inequality via random rounding, Lect. notes GAFA seminar 2019.
[19] B. Klartag, V.D. Milman, Geometry of log-concave functions and measures, Geom. Dedicata 112 (2005) 169-182.
[20] J. E. Littlewood, A. C. Offord, On the number of real roots of a random algebraic equation. III, Rec. Math. [Mat. Sbornik] N.S. 12 (54), (1943), 277-286
[21] A. Litvak, A. Pajor, M. Rudelson, N. Tomczak-Jaegermann, Smallest singular value of random matrices and geometry of random polytopes, Adv. Math. 195 (2005), no. 2, 491-523.
[22] A. E. Litvak and O. Rivasplata, Smallest singular value of sparse random matrices, Studia Math. 212 (2012), no. 3, 195-218.
[23] G. Livshyts, G. Paouris, P. Pivovarov, Small deviations for operator norms, work in progress
[24] A. E. Litvak, S. Spektor, Quantitative version of a Silverstein's result, GAFA, Lecture Notes in Math., 2116 (2014), 335-340.
[25] G. V. Livshyts, The smallest singular value of heavy-tailed not necessarily i.i.d. random matrices via random rounding, Journal d'Analyze Mathematique 2020.
[26] G. V. Livshyts, K. Tikhomirov, R. Vershynin, The smallest singular value of inhomogeneous square random matrices, Ann. Probab. 49 (2021), no. 3, 1286-1309.
[27] A. Lytova, K. Tikhomirov, On delocalization of eigenvectors of random non-Hermitian matrices, Probab. Theory Related Fields 177 (2020), no. 1-2, 465-524.
[28] J. von Neumann, H. H. Goldstine, Numerical inverting of matrices of high order, Bull. Amer. Math. Soc. 53 (1947), 1021-1099.
[29] S. Mendelson, G. Paouris, On the singular values of random matrices, Journal of the European Mathematics Society, 16, 823-834, 2014.
[30] G. Pólya, Szegö, Aufgaben und Lehrsätze aus der Analysis. Band I: Reihen. Integralrechnung. Funktionentheorie. Dritte berichtigte Auflage. Die Grundlehren der Mathematischen Wissenschaften, Band 19 Springer-Verlag, Berlin-New York 1964.
[31] P. Raghavan, C. D. Thompson, Randomized rounding: A technique for provably good algorithms and algorithmic proofs, Combinatorica, 7(4):365-374.
[32] E. Rebrova, K. Tikhomirov, Coverings of random ellipsoids, and invertibility of matrices with i.i.d. heavy-tailed entries, Israel J. Math. 227 (2018), no. 2, 507-544.
[33] B. A. Rogozin, An estimate for the maximum of the convolution of bounded densities, Teor. Veroyatnost. i Primenen. 32 (1987), no. 1, 53-61, English translation:Theory Probab. Appl. 32 (1987), no. 1, 48-56.
[34] M. Rudelson, Invertibility of random matrices: norm of the inverse, Annals of Mathematics 168 (2008), 575-600.
[35] M. Rudelson, R. Vershynin, The Littlewood-Offord problem and invertibility of random matrices, Adv. Math. 218 (2008), no. 2, 600-633.
[36] M. Rudelson, R. Vershynin, The least singular value of a random square matrix is $O\left(\sqrt{n}^{-1}\right)$, Comptes rendus de l'Academie des sciences - Mathematique 346 (2008), 893896.
[37] M. Rudelson, R. Vershynin, Smallest singular value of a random rectangular matrix, Communications on Pure and Applied Mathematics 62 (2009), 1707-1739.
[38] M. Rudelson, R. Vershynin, Non-asymptotic theory of random matrices: extreme singular values, Proceedings of the International Congress of Mathematicians, 2010, pp. 83-120.
[39] M. Rudelson, R. Vershynin, Small ball probabilities for linear images of high dimensional distributions, Int. Math. Res. Not. 19 (2015), 9594-9617.
[40] M. Rudelson, R. Vershynin, Delocalization of eigenvectors of random matrices with independent entries, Duke Math. J. Volume 164, Number 13 (2015), 2507-2538.
[41] C. Schütt, Entropy numbers of diagonal operators between symmetric Banach spaces, J. Approx. Theory 40 (1984), 121-128.
[42] A. Srinivasan, Approximation Algorithms via Randomized Rounding: a Survey, Lectures on Approximation and Randomized Algorithms, Series in Advanced Topics in Mathematics, Polish Scientific Publishers PWN, Warsaw, 9-71, (1999).
[43] S. Szarek, Condition numbers of random matrices, J. Complexity 7 (1991), 131-149.
[44] K. Tatarko, An upper bound on the smallest singular value of a square random matrix, J. Complexity 48 (2018), 119-128.
[45] T. Tao and V. Vu, On random $\pm 1$ matrices: Singularity and Determinant, Random Structures and Algorithms 28 (2006), no 1, 1-23.
[46] T. Tao, V. Vu, On the singularity probability of random Bernoulli matrices, J. Amer. Math. Soc. 20 (2007), 603-628.
[47] T. Tao and V. Vu, Inverse Littlewood-Offord theorems and the condition number of random discrete matrices, Annals of Mathematics 169 (2009), 595-632.
[48] T. Tao and V. Vu, Random matrices: the distribution of the smallest singular values, Geom. Funct. Anal. 20 (2010), no. 1, 260-297. MR2647142
[49] K. Tikhomirov, The limit of the smallest singular value of random matrices with i.i.d. entries, Adv. Math. 284 (2015), 1-20.
[50] K. Tikhomirov, The smallest singular value of random rectangular matrices with no moment assumptions on entries, Israel J. Math. 212 (2016), no. 1, 289-314.
[51] K. Tikhomirov, Invertibility via distance for non-centered random matrices with continuous distributions, Random Structures and Algorithms, to appear.
[52] K. Tikhomirov, Singularity of random Bernoulli matrices, Ann. of Math. (2) 191 (2020), no. 2, 593-634.
[53] R. Vershynin, High-dimensional probability: an introduction with applications in data science, Cambridge University Press, 2018.
[54] R. Vershynin, Spectral norm of products of random and deterministic matrices, Probability Theory and Related Fields 150 (2011), 471-509.
[55] M. S. Viazovska, The sphere packing problem in dimension 8, Annals of mathematics (2017), 991-1015.
[56] H. Cohn, A. Kumar, S. Miller, D. Radchenko, M. S. Viazovska, The sphere packing problem in dimension 24, Annals of mathematics (2017), 185 (3), 1017-1033.
[57] Shiryaev, A. N. Probability, vol. 1, Graduate Texts in Mathematics 95 (1989).
[58] S. Jukna Extremal combinatorics: with applications in computer science, Springer 512 (2011).

