Four lectures on probabilistic methods for data science

Roman Vershynin

Abstract. Methods of high-dimensional probability play a central role in applications for statistics, signal processing, theoretical computer science and related fields. These lectures present a sample of particularly useful tools of high-dimensional probability, focusing on the classical and matrix Bernstein's inequality and the uniform matrix deviation inequality. We illustrate these tools with applications for dimension reduction, network analysis, covariance estimation, matrix completion and sparse signal recovery. The lectures are geared towards beginning graduate students who have taken a rigorous course in probability but may not have any experience in data science applications.

Contents

1. Lecture 1: Concentration of sums of independent random variables

These lectures present a sample of modern methods of high dimensional probability and illustrate these methods with applications in data science. This sample is not comprehensive by any means, but it could serve as a point of entry into a branch of modern probability that is motivated by a variety of data-related problems.

To get most out of these lectures, you should have taken a graduate course in probability, have a good command of linear algebra (including singular value decomposition) and be familiar with basic concepts in normed spaces (including L^p spaces).

All of the material of these lectures is covered more systematically, at a slower pace, and with a wider range of applications, in my forthcoming textbook [53]. You may also be interested in two similar tutorials: [51] is focused on random matrices, and a more advanced text [52] discusses high-dimensional inference problems.

It should be possible to use these lectures for a self-study or group study. You will find here many places where you are invited to do some work (marked in the text e.g. by "check this!"), and you are encouraged to do it to get a better grasp of the material. Each lecture ends with a section called "Notes" where you will find references of the results just discussed, as well as some improvements and extensions.

We are now ready to start.

Probabilistic reasoning has a major impact on modern data science. There are roughly two ways in which this happens.

- *Radnomized algorithms*, which perform some operations at random, have long been developed in computer science and remain very popular. Randomized algorithms are among the most effective methods – and sometimes the only known ones – for many data problems.
- *Random models of data* form the usual premise of statistical analysis. Even when the data at hand is deterministic, it is often helpful to think of it as a random sample drawn from some unknown distribution ("population").

In this lectures, we will encounter both randomized algorithms and random models of data.

1.1. Sub-gaussian distributions   Before we start discussing probabilistic methods, we will introduce an important class of probability distributions that forms a natural "habitat" for random variables in many theoretical and applied problems. These are *sub-gaussian* distributions. As the name suggests, we will be looking at an extension of the most fundamental distribution in probability theory – the gaussian, or normal, distribution $N(\mu, \sigma)$.

It is a good exercise to check that the standard normal random variable $X \sim N(0, 1)$ satisfies the following basic properties:

Tails: $\mathbb{P}\left\{|X| \geqslant t\right\} \leqslant 2 \exp(-t^2/2)$ for all $t \geqslant 0$.
Moments: $\|X\|_p := (\mathbb{E}\,|X|^p)^{1/p} = O(\sqrt{p})$ as $p \to \infty$.
MGF of square: [1] $\mathbb{E}\exp(cX^2) \leqslant 2$ for some $c > 0$.
MGF: $\mathbb{E}\exp(\lambda X) = \exp(\lambda^2)$ for all $\lambda \in \mathbb{R}$.

All these properties tell the same story from four different perspectives. It is not very difficult to show (although we will not do it here) that for any random variable X, not necessarily Gaussian, these four properties are essentially equivalent.

Proposition 1.1.1 (Sub-gaussian properties). *For a random variable X, the following properties are equivalent.*[2]

Tails: $\mathbb{P}\left\{|X| \geqslant t\right\} \leqslant 2 \exp(-t^2/K_1^2)$ *for all* $t \geqslant 0$.
Moments: $\|X\|_p \leqslant K_2\sqrt{p}$ *for all* $p \geqslant 1$.
MGF of square: $\mathbb{E}\exp(X^2/K_3^2) \leqslant 2$.

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

MGF: $\mathbb{E}\exp(\lambda X) \leqslant \exp(\lambda^2 K_4^2)$ *for all* $\lambda \in \mathbb{R}$.

Random variables that satisfy one of the first three properties (and thus all of them) are called *sub-gaussian*. The best K_3 is called the *sub-gaussian norm* of X, and is usually denoted $\|X\|_{\psi_2}$, that is

$$\|X\|_{\psi_2} := \inf\left\{t > 0 : \ \mathbb{E}\exp(X^2/t^2) \leqslant 2\right\}.$$

One can check that $\|\cdot\|_{\psi_2}$ indeed defines a norm; it is an example of the general concept of the *Orlicz norm*. Proposition 1.1.1 states that the numbers K_i in all four properties are equivalent to $\|X\|_{\psi_2}$ up to absolute constant factors.

Example 1.1.2. As we already noted, the standard normal random variable $X \sim N(0, 1)$ is sub-gaussian. Similarly, arbitrary *normal* random variables $X \sim N(\mu, \sigma)$ are sub-gaussian. Another example is a *Bernoulli* random variable X that takes values 0 and 1 with probabilities $1/2$ each. More generally, any *bounded* random variable X is sub-gaussian. On the contrary, Poisson, exponential, Pareto and

---

[1]MGF stands for moment generation function.

[2]The parameters $K_i > 0$ appearing in these properties can be different. However, they may differ from each other by at most an absolute constant factor. This means that there exists an absolute constant C such that property implies property j with parameter $K_2 \leqslant CK_1$, and similarly for every other pair or properties.

Cauchy distributions are not sub-gaussian. (Verify all these claims; this is not difficult.)

1.2. Hoeffding's inequality   You may remember from a basic course in probability that the normal distribution $N(\mu, \sigma)$ has a remarkable property: the sum of independent normal random variables is also normal. Here is a version of this property for sub-gaussian distributions.

Proposition 1.2.1 (Sums of sub-gaussians). *Let $X_1, \ldots, X_N$ be independent, mean zero, sub-gaussian random variables. Then $\sum_{i=1}^N X_i$ is a sub-gaussian, and*

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

where C is an absolute constant.[3]

Proof. Let us bound the moment generating function of the sum for any $\lambda \in \mathbb{R}$:

$$\mathbb{E} \exp \Big(\lambda \sum_{i=1}^N X_i \Big) = \prod_{i=1}^N \mathbb{E} \exp(\lambda X_i) \quad \text{(using independence)}$$

$$\leqslant \prod_{i=1}^N \exp(C\lambda^2 \|X_i\|_{\psi_2}^2) \quad \text{(by last property in Proposition 1.1.1)}$$

$$= \exp(\lambda^2 K^2) \quad \text{where } K^2 := C \sum_{i=1}^N \|X_i\|_{\psi_2}^2.$$

Using again the last property in Proposition 1.1.1, we conclude that the sum $S = \sum_{i=1}^N X_i$ is sub-gaussian, and $\|S\|_{\psi_2} \leqslant C_1 K$ where C_1 is an absolute constant. The proof is complete. □

Let us rewrite Proposition 1.2.1 in a form that is often more useful in applications, namely as a *concentration inequality*. To do this, we simply use the first property in Proposition 1.1.1 for the sum $\sum_{i=1}^N X_i$. We immediately get the following.

Theorem 1.2.2 (General Hoeffding's inequality). *Let $X_1, \ldots, X_N$ be independent, mean zero, sub-gaussian random variables. Then, for every $t \geqslant 0$ we have*

$$\mathbb{P}\Big\{ \Big| \sum_{i=1}^N X_i \Big| \geqslant t \Big\} \leqslant 2 \exp \Big(- \frac{ct^2}{\sum_{i=1}^N \|X_i\|_{\psi_2}^2} \Big).$$

Hoeffding's inequality controls how far and with what probability can a sum of independent random variables deviate from its mean, which is zero.

---

[3]In the future, we will always denote positive absolute constants by C, c, C_1, etc. These numbers do not depend on anything. In most cases, one can get good bounds on these constants from the proof, but the optimal constants for each result are rarely known.

1.3. Sub-exponential distributions Sub-gaussian distributions form a sufficiently wide class of distributions. Many results in probability and data science are proved nowadays in the for sub-gaussian random variables. Still, as we noted, there are some natural random variables that are not sub-gaussian. For example, the square X^2 of a normal random variable $X \sim N(0,1)$ is not sub-gaussian. (Check!) To cover examples like this, we will introduce a similar but weaker notion of *sub-exponential distributions*.

Proposition 1.3.1 (Sub-exponential properties). *For a random variable X, the following properties are equivalent, in the same sense as in Proposition 1.1.1.*

Tails: $\mathbb{P}\left\{|X| \geqslant t\right\} \leqslant 2\exp(-t/K_1)$ *for all* $t \geqslant 0$.
Moments: $\|X\|_p \leqslant K_2 p$ *for all* $p \geqslant 1$.
MGF of square: $\mathbb{E}\exp(|X|/K_3) \leqslant 2$.

Moreover, if $\mathbb{E} X = 0$ then these properties imply the following one:

MGF: $\mathbb{E}\exp(\lambda X) \leqslant \exp(\lambda^2 K_4^2)$ *for* $|\lambda| \leqslant 1/K_4$.

Just like we did for sub-gaussian distributions, we call the best K_3 the *sub-exponential norm* of X and denote it by $\|X\|_{\psi_2}$, that is

$$\|X\|_{\psi_1} := \inf\{t > 0: \ \mathbb{E}\exp(|X|/t) \leqslant 2\}.$$

All sub-exponential random variables are squares of sub-gaussian random variables. Indeed, inspecting the definitions you will quickly see that

(1.3.2) $$\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2.$$

(Check!)

1.4. Bernstein's inequality A version of Hoeffding's inequality for sub-exponential random variables is called Bernstein's inequality. You may naturally expect to see a sub-exponential tail bound in this result. So it may come as a surprise that Bernstein's inequality actually has a mixture of *two* tails – sub-gaussian and sub-exponential. Let us state and prove the inequality first, and then we will comment on the mixture of the two tails.

Theorem 1.4.1 (Bernstein's inequality). *Let $X_1, \ldots, X_N$ be independent, mean zero, sub-exponential random variables. Then, for every $t \geqslant 0$ we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^N X_i\right| \geqslant t\right\} \leqslant 2\exp\left[-c\min\left(\frac{t^2}{\sum_{i=1}^N \|X_i\|_{\psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}}\right)\right].$$

Proof. For simplicity, we will assume that $K = 1$ and only prove the one-sided bound (without absolute value); the general case is not much harder. Our approach will be based on bounding the *moment generating function* of the sum $S := \sum_{i=1}^N X_i$. To see how MGF can be helpful here, choose $\lambda \geqslant 0$ and use Markov's inequality to get

(1.4.2) $$\mathbb{P}\left\{S \geqslant t\right\} = \mathbb{P}\left\{\exp(\lambda S) \geqslant \exp(\lambda t)\right\} \leqslant e^{-\lambda t}\,\mathbb{E}\exp(\lambda S).$$

Recall that $S = \sum_{i=1}^{N} X_i$ and use independence to express the right side of (1.4.2) as

$$e^{-\lambda t} \prod_{i=1}^{N} \mathbb{E} \exp(\lambda X_i).$$

(Check!) It remains to bound the MGF of each term X_i, and this is a much simpler task. If we choose λ small enough so that

(1.4.3) $$0 < \lambda \leqslant \frac{c}{\max_i \|X_i\|_{\psi_1}},$$

then we can use the last property in Proposition 1.3.1 to get

$$\mathbb{E} \exp(\lambda X_i) \leqslant \exp\left(C\lambda^2 \|X_i\|_{\psi_1}^2\right).$$

Substitute into (1.4.2) and conclude that

$$\mathbb{P}\{S \geqslant t\} \leqslant \exp\left(-\lambda t + C\lambda^2 \sigma^2\right)$$

where $\sigma^2 = \sum_{i=1}^{N} \|X_i\|_{\psi_1}^2$. The left side does not depend on λ while the right side does. So we can choose λ that minimizes the right side subject to the constraint (1.4.3). When this is done carefully, we obtain the tail bound stated in Bernstein's inequality. (Do this!) □

Now, why does Bernstein's inequality has a mixture of two tails? The sub-exponential tail should of course be there. Indeed, even if the entire sum consisted of a single term X_i, the best bound we could hope for would be of the form $\exp(-ct/\|X_i\|_{\psi_1})$. The sub-gaussian term could be explained by the central limit theorem, which states that the sum should becomes approximately *normal* as the number of terms N increases to infinity.

Remark 1.4.4 (Bernstein's inequality for bounded random variables). Suppose the random variables X_i are *uniformly bounded*, which is a stronger assumption than being sub-gaussian. Then there is a useful version of Bernstein's inequality, which unlike Theorem 1.4.1 is sensitive to the *variances* of X_i's. It states that if $K > 0$ is such that $|X_i| \leqslant K$ almost surely for all i, then, for every $t \geqslant 0$, we have

(1.4.5) $$\mathbb{P}\left\{ \left| \sum_{i=1}^{N} X_i \right| \geqslant t \right\} \leqslant 2 \exp\left(-\frac{t^2/2}{\sigma^2 + CKt} \right).$$

Here $\sigma^2 = \sum_{i=1}^{N} \mathbb{E} X_i^2$ is the variance of the sum. This version of Bernstein's inequality can be proved in essentially the same way as Theorem 1.4.1. We will not do it here, but a stronger Theorem 2.2.1, which is valid for matrix-valued random variables X_i, will be proved in Lecture 2.

To compare this with Theorem 1.4.1, note that $\sigma^2 + CKt \leqslant 2 \max(\sigma^2, CKt)$. So we can state this the probability bound (1.4.5) as

$$2 \exp\left[-c \min\left(\frac{t^2}{\sigma^2}, \frac{t}{K} \right) \right].$$

Just like before, here we also have a mixture of two tails, sub-gaussian and sub-exponential. The sub-gaussian tail is a bit sharper than in Theorem 1.4.1, since it depends on the *variances* rather than sub-gaussian norms of X_i. The sub-exponential tail, on the other hand, is weaker, since it depends on the sup-norms rather than the sub-exponential norms of X_i.

1.5. Sub-gaussian random vectors The concept of sub-gaussian distributions can be extended to higher dimensions. Consider a random vector X taking values in $\mathbb{R}^n$. We call X a *sub-gaussian random vector* if all one-dimensional *marginals* of X, i.e. the random variables $\langle X, x \rangle$ for $x \in \mathbb{R}^n$, are sub-gaussian. The sub-gaussian norm of X is defined as

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

where S^{n-1} denotes the unit Euclidean sphere in $\mathbb{R}^n$.

Example 1.5.1. Examples of sub-gaussian random distributions in $\mathbb{R}^n$ include the standard normal distribution $N(0, I_n)$ (why?), the uniform distribution on the centered Euclidean sphere of radius $\sqrt{n}$, the uniform distribution on the cube $\{-1, 1\}^n$, and many others. The last example can be generalized: a random vector $X = (X_1, \ldots, X_n)$ with independent and sub-gaussian coordinates is sub-gaussian, with $\|X\|_{\psi_2} \leqslant C \max_i \|X_i\|_{\psi_2}$.

1.6. Johnson-Lindenstrauss Lemma Concentration inequalities like Hoeffding's and Bernstein's are successfully used in the analysis of algorithms. Let us give one example for the problem of *dimension reduction*. Suppose we have some data that is represented as a set of N points in $\mathbb{R}^n$. (Think, for example, of n gene expressions of N patients.)

We would like to compress the data by representing it in a lower dimensional space $\mathbb{R}^m$ instead of $\mathbb{R}^n$ with $m \ll n$. By how much can we reduce the dimension without loosing the important features of the data?

The basic result in this direction is Johnson-Lindenstrauss Lemma. It states that a remarkably simple dimension reduction method works – a *random linear map* from $\mathbb{R}^n$ to $\mathbb{R}^m$ with

$$m \sim \log N,$$

see Figure 1.6.3. The logarithmic function grows very slowly, so we can usually reduce the dimension dramatically.

What exactly is a random linear map? Several models are possible to use. Here we will model such a map using a *Gaussian random matrix* – an $m \times n$ matrix A with independent $N(0, 1)$ entries. More generally, we can consider an $m \times n$ matrix A whose rows are independent, mean zero, isotropic[4] and sub-gaussian random vectors in $\mathbb{R}^n$. For example, the entries of A can be independent *Rademacher* entries – those taking values ± 1 with equal probabilities.

---

[4]A random vector $X \in \mathbb{R}^n$ is called *isotropic* if $\mathbb{E} X X^\mathsf{T} = I_n$.

Theorem 1.6.1 (Johnson-Lindenstrauss Lemma). *Let $\mathcal{X}$ be a set of N points in $\mathbb{R}^n$ and $\varepsilon \in (0,1)$. Consider an $m \times n$ matrix A whose rows are independent, mean zero, isotropic and sub-gaussian random vectors in $\mathbb{R}^n$. Rescale A by defining the "Gaussian random projection"*[5]

$$P := \frac{1}{\sqrt{m}} G.$$

Assume that

$$m \geqslant C\varepsilon^{-2} \log N,$$

where C is an appropriately large constant that depends only on the sub-gaussian norms of the vectors X_i. Then, with high probability (say, 0.99), the map P preserves the distances between all points in $\mathcal{X}$ with error ε, that is

(1.6.2) $(1-\varepsilon)\|x-y\|_2 \leqslant \|Px - Py\|_2 \leqslant (1+\varepsilon)\|x-y\|_2$ *for all $x, y \in \mathcal{X}$.*
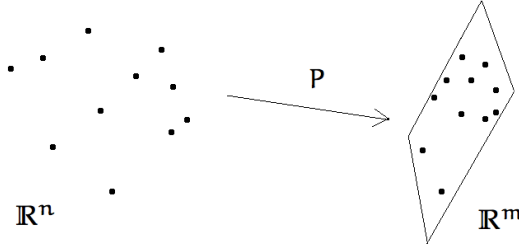

Figure 1.6.3. Johnson-Lindenstrauss Lemma states that a random projection of N data points from dimension n to dimension $m \sim \log N$ preserves the geometry of the data.

Proof. Take a closer look at the desired conclusion (1.6.2). By linearity, $Px - Py = P(x-y)$. So, dividing the inequality by $\|x-y\|_2$, we can rewrite (1.6.2) in the following way:

(1.6.4) $1-\varepsilon \leqslant \|Pz\|_2 \leqslant 1+\varepsilon$ for all $z \in T$

where

$$T := \left\{ \frac{x-y}{\|x-y\|_2} : x, y \in \mathcal{X} \text{ distinct points} \right\}.$$

It will be convenient to square the inequality (1.6.4). Using that $1 + \varepsilon \leqslant (1+\varepsilon)^2$ and $1 - \varepsilon \geqslant (1-\varepsilon)^2$, we see that it is enough to show that

(1.6.5) $1-\varepsilon \leqslant \|Pz\|_2^2 \leqslant 1+\varepsilon$ for all $z \in T$.

By construction, the coordinates of the vector $Pz = \frac{1}{\sqrt{m}}Az$ are $\frac{1}{\sqrt{m}}\langle X_i, z \rangle$. Thus we can restate (1.6.5) as

(1.6.6) $\left| \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \langle X_i, z \rangle^2 - 1 \right| \leqslant \varepsilon$ for all $z \in T$.

---

[5]Strictly speaking, this P is not a projection since it maps $\mathbb{R}^n$ to a different space $\mathbb{R}^m$.

Results like (1.6.6) are often proved by combining *concentration* and a *union bound*. In order to use concentration, we first fix $z \in T$. By assumption, the random variables $\langle X_i, z \rangle^2 - 1$ are independent; they have zero mean (use isotropy to check this!), and they are sub-exponential (use (1.3.2) to check this). Then Bernstein's inequality (Theorem 1.4.1) gives

$$\mathbb{P}\left\{\left|\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\langle X_i, z\rangle^2 - 1\right| > \varepsilon\right\} \leqslant 2\exp(-c\varepsilon^2 m).$$

(Check!)

Finally, we can unfix z by taking a union bound over all possible $z \in T$:

$$\mathbb{P}\left\{\max_{z\in T}\left|\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\langle X_i, z\rangle^2 - 1\right| > \varepsilon\right\} \leqslant \sum_{z\in T}\mathbb{P}\left\{\left|\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\langle X_i, z\rangle^2 - 1\right| > \varepsilon\right\}$$

$$(1.6.7) \qquad\qquad\qquad\qquad\qquad \leqslant |T| \cdot 2\exp(-c\varepsilon^2 m).$$

By definition of T, we have $|T| \leqslant N^2$. So, if we choose $m \geqslant C\varepsilon^{-2}\log N$ with appropriately large constant C, we can make (1.6.7) bounded by 0.01. The proof is complete. □

1.7. Notes   The material presented in Sections 1.1–1.5 is basic and can be found e.g. in [51] and [53] with all the proofs. Bernstein's and Hoeffding's inequalities that we covered here are two basic examples of *concentration inequalities*. There are many other useful concentration inequalities for sums of independent random variables (e.g. Chernoff's and Bennett's) and for more general objects. The textbook [53] is an elementary introduction into concentration; the books [10, 33, 34] offer more comprehensive and more advanced accounts of this area.

The original version of Johnson-Lindenstrauss Lemma was proved in [26]. The version we gave here, Theorem 1.6.1, was stated with probability of success 0.99, but an inspection of the proof gives probability $1 - 2\exp(-c\varepsilon^2 m)$ which is much better for large m. A great variety of ramifications and applications of Johnson-Lindenstrauss lemma are known, see e.g. [2, 4, 7, 10, 29, 37].

2. Lecture 2: Concentration of sums of independent random matrices

In the previous lecture we proved Bernstein's inequality, which quantifies how a sum of independent random variables concentrates about its mean. We will now study an extension of Bernstein's inequality to higher dimensions, which holds for sums of independent *random matrices*.

2.1. Matrix calculus   The key idea of developing a matrix Bernstein's inequality will be to use *matrix calculus*, which allows us to operate with matrices as with scalars – adding and multiplying them of course, but also comparing matrices and applying functions to matrices. Let us explain this.

We can compare matrices to each other using the notion of being *positive semi-definite*. Let us focus here on $n \times n$ symmetric matrices. If $A - B$ is a positive semidefinite matrix, which we denote $A - B \succeq 0$, then we say that $A \succeq B$ (and, of course, $B \preceq A$). This defines a *partial order* on the set of $n \times n$ symmetric matrices. The ream "partial" indicates that, unlike the real numbers, there exist $n \times n$ symmetric matrices A and B that can not be compared. (Give an example where neither $A \preceq B$ nor $B \succeq A$!)

Next, let us guess how to measure the *magnitude* of a matrix A. The magnitude of a scalar $a \in \mathbb{R}$ is measured by the absolute value $|a|$; it is the smallest non-negative number t such that

$$-t \leqslant a \leqslant t.$$

Extending this reasoning to matrices, we can measure the magnitude of of an $n \times n$ symmetric matrix A by the smallest non-negative number t such that[6]

$$-tI_n \preceq A \preceq tI_n.$$

The smallest t is called the *operator norm* of A and is denoted $\|A\|$. Diagonalizing A, we can see that

(2.1.1) $$\|A\| = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } A\}.$$

With a little more work (do it!), we can see that $\|A\|$ is the norm of A acting as a linear operator on $\mathbb{R}^n$ equipped with the Euclidean norm $\|\cdot\|_2$; this is why $\|A\|$ is called the operator norm. Thus $\|A\|$ is the smallest non-negative number M such that

$$\|Ax\|_2 \leqslant M\|x\|_2 \quad \text{for all } x \in \mathbb{R}^n.$$

Finally, we will need to be able to take *functions of matrices*. Let $f : \mathbb{R} \to \mathbb{R}$ be a function and X be an $n \times n$ symmetric matrix. We can define $f(X)$ in two equivalent ways. The spectral theorem allows us to represent X as

$$X = \sum_{i=1}^{n} \lambda_i u_i u_i^{\mathsf{T}}$$

where λ_i are the eigenvalues of X and u_i are the corresponding eigenvectors. Then we can simply define

$$f(X) := \sum_{i=1}^{n} f(\lambda_i) u_i u_i^{\mathsf{T}}.$$

Note that $f(X)$ has the same eigenvectors as X, but the eigenvalues change under the action of f. An equivalent way to define $f(X)$ is using power series. Suppose the function f has a convergent power series expansion about some point $x \in \mathbb{R}$, i.e.

$$f(x) = \sum_{k=1}^{\infty} a_k (x - x_0)^k.$$

---

[6]Here and later, I_n denotes the $n \times n$ identity matrix.

Then one can check that the following matrix series converges[7] and defines $f(X)$:

$$f(X) = \sum_{k=1}^{\infty} a_k (X - X_0)^k.$$

(Check!)

2.2. Matrix Bernstein's inequality   We are now ready to state and prove a remarkable generalization of Bernstein's inequality for random matrices.

Theorem 2.2.1 (Matrix Bernstein's inequality). *Let* $X_1, \ldots, X_N$ *be independent, mean zero,* $n \times n$ *symmetric random matrices, such that* $\|X_i\| \leqslant K$ *almost surely for all* i. *Then, for every* $t \geqslant 0$ *we have*

$$\mathbb{P}\left\{\left\|\sum_{i=1}^{N} X_i\right\| \geqslant t\right\} \leqslant 2n \cdot \exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right).$$

Here $\sigma^2 = \left\|\sum_{i=1}^{N} \mathbb{E} X_i^2\right\|$ *is the norm of the "matrix variance" of the sum.*

The scalar case, where $n = 1$, is the classical Bernstein's inequality we stated in (1.4.5). A remarkable feature of matrix Bernstein's inequality, which makes it especially powerful, is that it *does not require any independence of the entries* (or the rows or columns) of X_i; all is needed is that the random matrices X_i be independent *from each other*.

We will prove matrix Bernstein's inequality and give a few applications in this and next lecture.

Our proof will be based on bounding the *moment generating function* (MGF) $\mathbb{E}\exp(\lambda S)$ of the sum $S = \sum_{i=1}^{N} X_i$. Note that to exponentiate the matrix λS in order to define the matrix MGF, we rely on matrix calculus that we introduced in Section 2.1.

If the terms X_i were *scalars*, independence would yield the classical fact that MGF of a product is a product of MGF's, i.e.

$$(2.2.2) \qquad \mathbb{E}\exp(\lambda S) = \mathbb{E}\prod_{i=1}^{N}\exp(\lambda X_i) = \prod_{i=1}^{N}\mathbb{E}\exp(\lambda X_i).$$

But for *matrices*, this reasoning breaks down badly, for in general

$$e^{X+Y} \neq e^X e^Y$$

even for 2×2 symmetric matrices X and Y. (Give a counterexample!)

Fortunately, there are some trace inequalities that can often serve as proxies for the missing inequality $e^{X+Y} = e^X e^Y$. One of such proxies is *Golden-Thompson inequality*, which states that

$$(2.2.3) \qquad \qquad \mathrm{tr}(e^{X+Y}) \leqslant \mathrm{tr}(e^X e^Y)$$

---

[7]The convergence holds in any given metric on the set of matrices, for example in the metric given by the operator norm. In this series, the terms $(X - X_0)^k$ are defined by the usual matrix product.

for any $n \times n$ symmetric matrices X and Y. Another result, which we will actually use in the proof of matrix Bernstein's inequality, is *Lieb's inequality*.

Theorem 2.2.4 (Lieb's inequality). *Let H be an $n \times n$ symmetric matrix. Then the function*

$$f(X) = \operatorname{tr}\exp(H + \log X)$$

is concave[8] on the space on $n \times n$ symmetric matrices.

Note that in the scalar case, where $n = 1$, the function f in Lieb's inequality is linear and the result is trivial.

To use Lieb's inequality in a probabilistic context, we will combine it with the classical Jensen's inequality. It states that for any concave function f and a random matrix X, one has[9]

(2.2.5) $\mathbb{E} f(X) \leqslant f(\mathbb{E} X).$

Using this for the function f in Lieb's inequality, we get

$$\mathbb{E} \operatorname{tr}\exp(H + \log X) \leqslant \operatorname{tr}\exp(H + \log \mathbb{E} X).$$

And changing variables to $X = e^Z$, we get the following:

Lemma 2.2.6 (Lieb's inequality for random matrices). *Let H be a fixed $n \times n$ symmetric matrix and Z be an $n \times n$ symmetric random matrix. Then*

$$\mathbb{E} \operatorname{tr}\exp(H + Z) \leqslant \operatorname{tr}\exp(H + \log \mathbb{E} e^Z).$$

Lieb's inequality is a perfect tool for bounding the MGF of a sum of independent random variables $S = \sum_{i=1}^{N} X_i$. To do this, let us condition on the random variables $X_1, \ldots, X_{N-1}$. Apply Lemma 2.2.6 for the fixed matrix $H := \sum_{i=1}^{N-1} \lambda X_i$ and the random matrix $Z := \lambda X_i$, and afterwards take expectation with respect to $X_1, \ldots, X_{N-1}$. By the law of total expectation, we get

$$\mathbb{E} \operatorname{tr}\exp(\lambda S) \leqslant \mathbb{E} \operatorname{tr}\exp\left[\sum_{i=1}^{N-1} \lambda X_i + \log \mathbb{E} e^{\lambda X_N} \right].$$

Next, apply Lemma 2.2.6 in a similar manner for $H := \sum_{i=1}^{N-2} \lambda X_i + \log \mathbb{E} e^{\lambda X_N}$ and $Z := \lambda X_{N-1}$, and so on. After N times, we obtain:

Lemma 2.2.7 (MGF of a sum of independent random matrices). *Let $X_1, \ldots, X_N$ be independent $n \times n$ symmetric random matrices. Then the sum $S = \sum_{i=1}^{N} X_i$ satisfies*

$$\mathbb{E} \operatorname{tr}\exp(\lambda S) \leqslant \operatorname{tr}\exp\left[\sum_{i=1}^{N} \log \mathbb{E} e^{\lambda X_i} \right].$$

---

[8]Formally, *concavity* of f means that $f(\lambda X + (1 - \lambda)Y) \geqslant \lambda f(X) + (1 - \lambda)f(Y)$ for all symmetric matrices X and Y and all $\lambda \in [0, 1]$.

[9]Jensen's inequality is usually stated for a *convex* function g and a *scalar* random variable X, and it reads $g(\mathbb{E} X) \leqslant \mathbb{E} g(X)$. From this, inequality (2.2.5) for concave functions and random matrices easily follows (Check!).

Think of this inequality is a matrix version of the scalar identity (2.2.2). The main difference is that it bounds the trace of the MGF[10] rather the MGF itself.

You may recall from a course in probability theory that the quantity $\log \mathbb{E}\, e^{\lambda X_i}$ that appears in this bound is called the *cumulant generating function* of X_i. Lemma 2.2.7 reduces the complexity of our task significantly, for it is much easier to bound the cumulant generating function of each *single* random variable X_i than to say something about their sum. Here is a simple bound.

Lemma 2.2.8 (Moment generating function)**.** *Let X be an $n \times n$ symmetric random matrix. Assume that $\mathbb{E}\, X = 0$ and $\|X\| \leqslant K$ almost surely. Then, for all $0 < \lambda < 3/K$ we have*

$$\mathbb{E} \exp(\lambda X) \preceq \exp\left(g(\lambda)\, \mathbb{E}\, X^2\right) \quad \text{where} \quad g(\lambda) = \frac{\lambda^2/2}{1 - \lambda K/3}.$$

Proof. First, check that the following scalar inequality holds for $0 < \lambda < 3/K$ and $|x| \leqslant K$:

$$e^{\lambda x} \leqslant 1 + \lambda x + g(\lambda)x^2.$$

Then extend it to matrices using matrix calculus: if $0 < \lambda < 3/K$ and $\|X\| \leqslant K$ then

$$e^{\lambda X} \preceq I + \lambda X + g(\lambda)X^2.$$

(Do these two steps carefully!) Finally, take expectation and recall $\mathbb{E}\, X = 0$ to obtain

$$\mathbb{E}\, e^{\lambda X} \preceq I + g(\lambda)\, \mathbb{E}\, X^2 \preceq \exp\left(g(\lambda)\, \mathbb{E}\, X^2\right).$$

In the last inequality, we use the matrix version of the scalar inequality $1 + z \leqslant e^z$ that holds for all $z \in \mathbb{R}$. The lemma is proved. $\qquad\square$

Proof of Matrix Bernstein's inequality. We would like to bound the operator norm of the random matrix $S = \sum_{i=1}^{N} X_i$, which, as we know from (2.1.1), is the largest eigenvalue of S *by magnitude*. For simplicity of exposition, let us drop the absolute value from (2.1.1) and just bound the maximal eigenvalue of S, which we denote $\lambda_{\max}(S)$. (Once this is done, we can repeat the argument for $-S$ to reinstate the absolute value. Do this!) So, we are to bound

$$
\begin{aligned}
\mathbb{P}\left\{\lambda_{\max}(S) \geqslant t\right\} &= \mathbb{P}\left\{e^{\lambda \cdot \lambda_{\max}(S)} \geqslant e^{\lambda t}\right\} \quad \text{(multiply by } \lambda > 0 \text{ and exponentiate)} \\
&\leqslant e^{-\lambda t}\, \mathbb{E}\, e^{\lambda \cdot \lambda_{\max}(S)} \quad \text{(by Markov's inequality)} \\
&= e^{-\lambda t}\, \mathbb{E}\, \lambda_{\max}(e^{\lambda S}) \quad \text{(check!)} \\
&\leqslant e^{-\lambda t}\, \mathbb{E}\, \mathrm{tr}\, e^{\lambda S} \quad \text{(max of eigenvalues is bounded by the sum)} \\
&\leqslant e^{-\lambda t}\, \mathrm{tr} \exp\left[\sum_{i=1}^{N} \log \mathbb{E}\, e^{\lambda X_i}\right] \quad \text{(use Lemma 2.2.7)} \\
&\leqslant \mathrm{tr} \exp\left[-\lambda t + g(\lambda)Z\right] \quad \text{(by Lemma 2.2.8)}
\end{aligned}
$$

---

[10]Note that the order of expectation and trace can be swapped using linearity.

where

$$Z := \sum_{i=1}^{N-1} \mathbb{E}\, X_i^2.$$

It remains to optimize this bound in λ. The minimum is attained for $\lambda = t/(\sigma^2 + Kt/3)$. (Check!) Substituting this value for λ, we conclude

$$\mathbb{P}\left\{\lambda_{\max}(S) \geqslant t\right\} \leqslant n \cdot \exp\left(-\frac{-t^2/2}{\sigma^2 + Kt/3}\right).$$

This completes the proof of Theorem 2.2.1. $\square$

Bernstein's inequality gives a powerful *tail bound* for $\|\sum_{i=1}^{N} X_i\|$. This easily implies a useful bound on the *expectation*:

Corollary 2.2.9 (Expected norm of sum of random matrices)**.** *Let* $X_1, \ldots, X_N$ *be independent, mean zero,* $n \times n$ *symmetric random matrices, such that* $\|X_i\| \leqslant K$ *almost surely for all* i. *Then*

$$\mathbb{E}\left\|\sum_{i=1}^{N} X_i\right\| \lesssim \sigma\sqrt{\log n} + K \log n$$

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

Proof. The link from tail bounds to expectation is provided by the basic identity

$$(2.2.10) \qquad\qquad \mathbb{E}\, Z = \int_0^\infty \mathbb{P}\left\{Z > t\right\} dt$$

which is valid for any non-negative random variable Z. (Check it!) Integrating the tail bound given by matrix Bernstein's inequality, you will arrive at the expectation bound we claimed. (Check!) $\square$

Notice in this corollary a mild, logarithmic, dependence on the ambient dimension n. As we will see shortly, this can be an important feature in some applications.

2.3. Community recovery in networks   Matrix Bernstein's inequality has many applications. The one we are going to discuss first is for the analysis of *networks*. A network can be mathematically represented by graph, a set of n vertices with edges connecting some of them. For simplicity, we will consider undirected graphs where the edges do not have arrows. Real world networks often tend to have clusters, or *communities* – subsets of vertices that are connected by unusually many edges. (Think, for example, about a friendship network where communities form around some common interests.) An important problem in data science is to recover communities from a given network.

We are going to explain one of the simplest methods for community recovery, which is called *spectral clustering*. But before we introduce it, we will first of all place a probabilistic model on the networks we consider. In other words, it will be convenient for us to view networks as *random graphs* whose edges are formed at random. Although not all real-world networks are truly random, this simplistic

model can motivate us to develop algorithms that would empirically succeed also for real-world networks.

The basic probabilistic model of random graphs is the *Erdös-Rényi model*.

Definition 2.3.1 (Erdös-Rényi model). *Consider a set of n vertices and connect every pair of vertices independently and with fixed probability p. The resulting random graph is said to follow the* Erdös-Rényi model $G(n, p)$.

Erdös-Rényi random model is very simple. But is not a good choice if we want to model a network with communities, for every pair of vertices has the same chance to be connected. So let us introduce a natural generalization of Erdös-Rényi random model that does allow for community structure:

Definition 2.3.2 (Stochastic block model). *Partition a set of n vertices into two subsets ("communities") with $n/2$ vertices each, and connect every pair vertices independently with probability p if they belong to the same community and $q < p$ if not. The resulting random graph is said to follow the* stochastic block model $G(n, p, q)$.

Figure 2.3.3 illustrate a simulation of a stochastic block model.

Figure 2.3.3. A network generated according to the stochastic block model $G(n, p, q)$ with $n = 200$ nodes and connection probabilities $p = 1/20$ and $q = 1/200$.

Suppose we are shown one instance of a random graph generated according to a stochastic block model $G(n, p, q)$. How can we find which vertices belong to which community?

The *spectral clustering* algorithm we are going to explain will do precisely this. It will be based on the spectrum of the *adjacency matrix* A of the graph, which is the $n \times n$ symmetric matrix whose entries A_{ij} equal 1 if the vertices i and j are connected by an edge, and 0 otherwise.[11]

The adjacency matrix A is a random matrix. Let us compute its expectation first. This is easy, since the entires of A are Bernoulli random variables. If i and j

---

[11]For convenience, we call the vertices of the graph $1, 2, \dots, n$.

belong to the same community then $\mathbb{E}\, A_{ij} = p$ and otherwise $\mathbb{E}\, A_{ij} = q$. Thus A has block structure: for example, if $n = 4$ then A looks like this:

$$
\mathbb{E}\, A =
\left[
\begin{array}{cc|cc}
p & p & q & q \\
p & p & q & q \\
\hline
q & q & p & p \\
q & q & p & p
\end{array}
\right]
$$

(For illustration purposes, we grouped the vertices from each community together. In reality, we do not know in advance how to group them, but we do not need to.)

You will easily check that A has rank 2, and the non-zero eigenvalues and the corresponding eigenvectors are

(2.3.4)

$$
\lambda_1(\mathbb{E}\, A) = \left(\frac{p+q}{2}\right)n, \quad
v_1(\mathbb{E}\, A) =
\begin{bmatrix}
1 \\ 1 \\ 1 \\ 1
\end{bmatrix}
; \quad
\lambda_2(\mathbb{E}\, A) = \left(\frac{p-q}{2}\right)n, \quad
v_2(\mathbb{E}\, A) =
\begin{bmatrix}
1 \\ 1 \\ -1 \\ -1
\end{bmatrix}.
$$

(Check!)

The eigenvalues and eigenvectors of $\mathbb{E}\, A$ tell us a lot about the community structure of the underlying graph. Indeed, the first (larger) eigenvalue,

$$
d := \left(\frac{p+q}{2}\right)n,
$$

is the *expected degree* of any vertex of the graph.[12] The second eigenvalue tells us whether there is any community structure at all (which happens when $p \neq q$ and thus $\lambda_2(\mathbb{E}\, A) \neq 0$). The first eigenvector v_1 is not informative of the structure of the network at all. It is the second eigenvector v_2 that tells us exactly how to separate the vertices into the two communities: the signs of the coefficients of v_2 can be used for this purpose.

Thus if we know $\mathbb{E}\, A$, we can recover the community structure of the network from the signs of the second eigenvector. The problem is that we do not know $\mathbb{E}\, A$. Instead, we know the adjacency matrix A. And if, by some chance, A is not far from $\mathbb{E}\, A$, we may hope to use the A to approximately recover the community structure. So is it true that $A \approx \mathbb{E}\, A$? The answer is yes, and we can prove it using matrix Bernstein's inequality.

Theorem 2.3.5 (Concentration of the stochastic block model)**.** *Let* A *be the adjacency matrix of a* $G(n, p, q)$ *random graph. Then*

$$
\mathbb{E}\, \|A - \mathbb{E}\, A\| \lesssim \sqrt{d \log n} + \log n.
$$

Here $d = (p+q)n/2$ *is the expected degree.*

---

[12]The degree of the vertex is the number of edges connected to it.

Proof. Let us sketch the argument. To use matrix Bernstein's inequality, let us break A into a sum of independent random matrices

$$A = \sum_{i,j:\, i \leqslant j} X_{ij},$$

where each matrix X_{ij} contains a pair of symmetric entries of A, or one diagonal entry.[13] Matrix Bernstein's inequality obviously applies for the sum

$$A - \mathbb{E}\,A = \sum_{i \leqslant j} (X_{ij} - \mathbb{E}\,X_{ij}).$$

Corollary 2.2.9 gives[14]

(2.3.6) $$\mathbb{E}\,\|A - \mathbb{E}\,A\| \lesssim \sigma\sqrt{\log n} + K\log n$$

where $\sigma^2 = \left\| \sum_{i \leqslant j} \mathbb{E}(X_{ij} - \mathbb{E}\,X_{ij})^2 \right\|$ and $K = \max_{ij} \|X_{ij} - \mathbb{E}\,X_{ij}\|$. It is a good exercise to check that

$$\sigma^2 \lesssim d \quad \text{and} \quad K \leqslant 2.$$

(Do it!) Substituting into (2.3.6), we complete the proof.   □

How useful is Theorem 2.3.5 for community recovery? Suppose that the network is not too sparse, namely

$$d \gg \log n.$$

Then

$$\|A - \mathbb{E}\,A\| \lesssim \sqrt{d \log n} \quad \text{while} \quad \|\mathbb{E}\,A\| = \lambda_1(\mathbb{E}\,A) = d,$$

which implies that

$$\|A - \mathbb{E}\,A\| \ll \|\mathbb{E}\,A\|.$$

In other words, A nicely approximates $\mathbb{E}\,A$: the relative error or approximation is small in the operator norm.

At this point one can classical results from the *perturbation theory* for matrices, which state that since A and $\mathbb{E}\,A$ are close, their eigenvalues and eigenvectors must also be close. The relevant perturbation results are *Weyl's inequality* for eigenvalues and *Davis-Kahan's inequality* for eigenvectors, which we will not reproduce here. Heuristically, what they give us is

$$v_2(A) \approx v_2(\mathbb{E}\,A) = \frac{1}{\sqrt{n}}\begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}.$$

---

[13]Precisely, if $i \neq j$, then X_{ij} has all zero entries except the (i,j) and (j,i) entries that equal 1. If $i = j$, the only non-zero entry of X_{ij} is the (i,i).

[14]We will liberally use the notation $\lesssim$ to hide constant factors appearing in the inequalities. Thus, $a \lesssim b$ means that $a \leqslant C\,b$ for some constant C.

Then we should expect that most of the coefficients of $v_2(A)$ be positive on one community and negative on the other. So we can use $v_2(A)$ to approximately recover the communities. This method is called *spectral clustering*:

Spectral Clustering Algorithm. *Compute $v_2(A)$, the eigenvector corresponding to the second largest eigenvalue of the adjacency matrix A of the network. Use the signs of the coefficients of $v_2(A)$ to predict the community membership of the vertices.*

We saw that spectral clustering should perform well for stochastic block model $G(n, p, q)$ if it is not too sparse, namely if the expected degrees satisfy $d = (p + q)n/2 \gg \log n$.

A more careful analysis along these lines, which you should be able to do yourself with some work, leads to the following more rigorous result.

Theorem 2.3.7 (Guarantees of spectral clustering)**.** *Consider a random graph generated according to the stochastic block model $G(n, p, q)$ with $p > q$, and set $a = pn$, $b = qn$. Suppose that*

(2.3.8) $$(a - b)^2 \gg \log(n)(a + b).$$

Then, with high probability, the spectral clustering algorithm recovers the communities up to $o(n)$ misclassified vertices.

2.4. Notes   The idea to extend concentration inequalities like Bernstein's to matrices goes back to R. Ahlswede and A. Winter [3]. They used Golden-Thompson inequality (2.2.3) and proved a slightly weaker form of matrix Bernstein's inequality than we gave in Section 2.2. R. Oliveira [42, 43] found a way to improve this argument and gave a result similar to Theorem 2.2.1. The version of matrix Bernstein's inequality we gave here (Theorem 2.2.1) and a proof based on Lieb's inequality is due to J. Tropp [45].

The survey [46] contains a comprehensive introduction of matrix calculus, a proof of Lieb's inequality (Theorem 2.2.4), a detailed proof of matrix Bernstein's inequality (Theorem 2.2.1) and a variety of applications. A proof of Golden-Thompson inequality (2.2.3) can be found in [8, Theorem 9.3.7].

In Section 2.3 we scratched the surface of an interdisciplinary area of *network analysis*. For a systematic introduction into networks, refer to the book [41]. Stochastic block models (Definition 2.3.2) were introduced in [28]. The community recovery problem in stochastic block models, sometimes also called community detection problem, has been in the spotlight in the last few years. A vast and still growing body of literature exists on algorithms and theoretical results for community recovery, see the book [41], the survey [21], papers such as [9, 24, 25, 27, 32, 40, 54] and the references therein.

A concentration result similar to Theorem 2.3.5 can be found in [42]; the argument there is also is based on matrix concentration. This theorem is not quite optimal. For dense networks, where with the expected degree d satisfies $d \gtrsim \log n$,

the concentration inequality in Theorem 2.3.5 can be improved to

$$\text{(2.4.1)} \qquad \mathbb{E}\,\|A - \mathbb{E}\,A\| \lesssim \sqrt{d}.$$

This improved bound goes back to the original paper [20] which studies the simpler Erdös-Rényi model but the results extend to stochastic block models [16]; it can also be deduced from [6,27,32].

If the network is relatively dense, i.e. $d \gtrsim \log n$, one can improve the guarantee (2.3.8) of spectral clustering in Theorem 2.3.7 to

$$(a - b)^2 \gg (a + b).$$

All one has to do is to use the improved concentration inequality (2.4.1) instead of Theorem 2.3.5. Furthermore, in this case there exist algorithms that can recover the communities *exactly*, i.e. without any misclassified vertices, and with high probability, see e.g. [1,16,27,38].

For sparser networks, where $d \ll \log n$ and possibly even $d = O(1)$, relatively little algorithms had been known until recently, but now there exist many approaches that provably recover communities in sparse stochastic block models, see e.g. [9,16,24,25,32,40,54].

3. Lecture 3: Covariance estimation and matrix completion

In the last lecture, we proved matrix Bernstein's inequality and gave an application for network analysis. We will spend this lecture discussing a couple of other interesting applications of matrix Bernstein's inequality. In Section 3.1 we will work on *covariance estimation*, a basic problem in high-dimensional statistics. In Section 3.2, we will derive a useful bound on norms random matrices, which unlike Bernstein's inequality does not require any boundedness assumptions on the distribution. We will apply this bound in Section 3.3 for a problem of *matrix completion*, where we are shown a small sample of the entries of a matrix and asked to guess the missing entries.

3.1. Covariance estimation Covariance estimation is a problem of fundamental importance in high-dimensional statistics. Suppose we have a sample of data points $X_1, \ldots, X_N$ in $\mathbb{R}^n$. It is often reasonable to assume that these points are independently sampled from the same probability distribution (or "population") which is unknown. We would like to learn something useful about this distribution.

Denote by X a random vector that has this (unknown) distribution. The most basic parameter of the distribution is the *mean* $\mathbb{E}\,X$. One can estimate $\mathbb{E}\,X$ from the sample by computing the *sample mean* $\frac{1}{N}\sum_{i=1}^{N} X_i$. The law of large numbers guarantees that the estimate becomes tight as the sample size N grows to infinity, i.e.

$$\frac{1}{N}\sum_{i=1}^{N} X_i \to \mathbb{E}\,X \quad \text{as } N \to \infty.$$

The next most basic parameter of the distribution is the *covariance matrix*

$$\Sigma := \mathbb{E}(X - \mathbb{E}\,X)(X - \mathbb{E}\,X)^{\mathsf{T}}.$$

This is a higher-dimensional version of the usual notion of *variance* of a random variable Z, which is

$$\mathrm{Var}(Z) = \mathbb{E}(Z - \mathbb{E}\,Z)^2.$$

The eigenvectors of the covariance matrix of Σ are called the *principal components*. Principal components that correspond to large eigenvalues of Σ are the directions in which the distribution of X is most extended, see Figure 3.1.1. These are often the most interesting directions in the data. Practitioners often visualize the high-dimensional data by projecting it onto the span of a few (maybe two or three) of such principal components; the projection may reveal some hidden structure of the data. This method is called *Principal Component Analysis* (PCA).
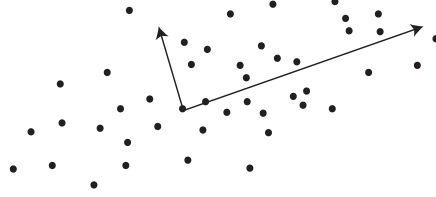

Figure 3.1.1. Data points $X_1, \ldots, X_N$ sampled from a distribution in $\mathbb{R}^n$ and the principal components of the covariance matrix.

One can estimate the covariance matrix Σ from sample by computing the *sample covariance*

$$\Sigma_N := \frac{1}{N} \sum_{i=1}^{N} (X_i - \mathbb{E}\,X_i)(X_i - \mathbb{E}\,X_i)^{\mathsf{T}}.$$

Again, the law of large numbers guarantees that the estimate becomes tight as the sample size N grows to infinity, i.e.

$$\Sigma_N \to \Sigma \quad \text{as } N \to \infty.$$

But how large should the sample size N for covariance estimation? Generally, one can not have $N < n$ for dimension reasons. (Why?) We are going to show that

$$N \sim n \log n$$

is enough. In other words, covariance estimation is possible with just *logarithmic oversampling*.

For simplicity, we shall state the covariance estimation bound for *mean zero* distributions. (If the mean is not zero, we can estimate it from the sample and subtract.)

Theorem 3.1.2 (Covariance estimation). *Let X be a random vector in $\mathbb{R}^n$ with covariance matrix Σ. Suppose that*

(3.1.3) $\|X\|_2^2 \lesssim \mathbb{E}\,\|X\|_2^2 = \mathrm{tr}\,\Sigma$ *almost surely.*

Then, for every $N \geqslant 1$, *we have*

$$\mathbb{E} \|\Sigma_N - \Sigma\| \leqslant \|\Sigma\| \left(\sqrt{\frac{n \log n}{N}} + \frac{n \log n}{N} \right).$$

Before we pass to the proof, let us note that Theorem 3.1.2 yields the covariance estimation result we promised. Let $\varepsilon \in (0, 1)$. If we take a sample of size

$$N \sim \varepsilon^{-2} n \log n,$$

then we are guaranteed covariance estimation with a good relative error:

$$\mathbb{E} \|\Sigma_N - \Sigma\| \leqslant \varepsilon \|\Sigma\|.$$

Proof. Apply matrix Bernstein's inequality (Corollary 2.2.9) for the sum of independent random matrices $X_i X_i^\mathsf{T} - \Sigma$ and get

(3.1.4) $\qquad \mathbb{E} \|\Sigma_N - \Sigma\| = \frac{1}{N} \mathbb{E} \left\| \sum_{i=1}^{N} (X_i X_i^\mathsf{T} - \Sigma) \right\| \lesssim \frac{1}{N} \left(\sigma \sqrt{\log n} + K \log n \right)$

where

$$\sigma^2 = \left\| \sum_{i=1}^{N} \mathbb{E}(X_i X_i^\mathsf{T} - \Sigma)^2 \right\| = N \left\| \mathbb{E}(XX^\mathsf{T} - \Sigma)^2 \right\|$$

and K is chosen so that

$$\|XX^\mathsf{T} - \Sigma\| \leqslant K \quad \text{almost surely.}$$

It remains to bound σ and K. Let us start with σ. We have

$$\begin{aligned}
\mathbb{E}(XX^\mathsf{T} - \Sigma)^2 &= \mathbb{E} \|X\|_2^2 XX^\mathsf{T} - \Sigma^2 \quad \text{(check by expanding the square)} \\
&\precsim \operatorname{tr}(\Sigma) \cdot \mathbb{E} XX^\mathsf{T} \quad \text{(drop } \Sigma^2 \text{ and use (3.1.3))} \\
&= \operatorname{tr}(\Sigma) \cdot \Sigma.
\end{aligned}$$

Thus

$$\sigma^2 \lesssim N \operatorname{tr}(\Sigma) \|\Sigma\|.$$

Next, to bound K, we have

$$\begin{aligned}
\|XX^\mathsf{T} - \Sigma\| &\leqslant \|X\|_2^2 + \|\Sigma\| \quad \text{(by triangle inequality)} \\
&\lesssim \operatorname{tr} \Sigma + \|\Sigma\| \quad \text{(using (3.1.3))} \\
&\leqslant 2 \operatorname{tr} \Sigma =: K.
\end{aligned}$$

Substitute the bounds on σ and K into (3.1.4) and get

$$\mathbb{E} \|\Sigma_N - \Sigma\| \lesssim \frac{1}{N} \left(\sqrt{N \operatorname{tr}(\Sigma) \|\Sigma\| \log n} + \operatorname{tr}(\Sigma) \log n \right)$$

To complete the proof, use that $\operatorname{tr} \Sigma \leqslant n \|\Sigma\|$ (check this!) and simplify the bound. $\qquad \square$

Remark 3.1.5 (Low-dimensional distributions). Much fewer samples are needed for covariance estimation for low-dimensional, or approximately low-dimensional, distributions. To measure approximate low-dimensionality we can use the notion

of the *stable rank* of Σ^2. The stable rank of a matrix A is defined as the square of the ratio of Frobenius to operator norms:[15]

$$r(A) := \frac{\|A\|_F^2}{\|A\|^2}.$$

The stable rank is always bounded by the usual, linear algebraic rank,

$$r(A) \leqslant \operatorname{rank}(A),$$

and it can be much smaller. (Check both claims.)

Our proof of Theorem 3.1.2 actually gives

$$\mathbb{E} \|\Sigma_N - \Sigma\| \leqslant \|\Sigma\| \left(\sqrt{\frac{r \log n}{N}} + \frac{r \log n}{N} \right).$$

where

$$r = r(\Sigma^{1/2}) = \frac{\operatorname{tr} \Sigma}{\|\Sigma\|}.$$

(Check this!) Therefore, covariance estimation is possible with

$$N \sim r \log n$$

samples.

Remark 3.1.6 (The boundedness condition). It is a good exercise to check that if we remove the boundedness condition (3.1.3), a nontrivial covariance estimation is impossible in general. (Show this!) But how do we know whether the boundedness condition holds for data at hand? We may not, but we can enforce this condition by *truncation*. All we have to do is to discard 1% of data points with largest norms. (Check this accurately, assuming that such truncation does not change the covariance significantly.)

3.2. Norms of random matrices We have worked a lot with the *operator norm* of matrices, denoted $\|A\|$. One may ask if is there a formula that expresses $\|A\|$ in terms of the entires A_{ij}. Unfortunately, there is no such formula. The operator norm is a more difficult quantity in this respect than the *Frobenius norm*, which as we know can be easily expressed in terms of entries: $\|A\|_F = (\sum_{i,j} A_{ij}^2)^{1/2}$.

If we can not express of $\|A\|$ in terms of the entires, can we at least get a good estimate? Let us consider $n \times n$ symmetric matrices for simplicity. In one direction, $\|A\|$ is always bounded *below* by the largest Euclidean norm of the rows A_i:

(3.2.1) $$\|A\| \geqslant \max_i \|A_i\|_2 = \max_i \left(\sum_j A_{ij}^2 \right)^{1/2}.$$

---

[15]The Frobenius norm of an $n \times m$ matrix, sometimes also called *Hilbert-Schmidt norm*, is defined as $\|A\|_F^2 = (\sum_{i=1}^n \sum_{j=1}^m A_{ij}^2)^{1/2}$. Equivalently, for an $n \times n$ symmetric matrix, $\|A\|_2 = (\sum_{i=1}^n \lambda_i(A)^2)^{1/2}$, where $\lambda_i(A)$ are the eigenvalues of A. Thus the stable rank of A can be expressed as $r(A) = \sum_{i=1}^n \lambda_i(A)^2 / \max_i \lambda_i(A)^2$.

(Check!) Unfortunately, this bound is sometimes very loose, and the best possible upper bound is

(3.2.2) $$\|A\| \leqslant \sqrt{n} \cdot \max_i \|A_i\|_2.$$

(Show this bound, and give an example where it is sharp.)

Fortunately, for *random* matrices with independent entries the bound (3.2.2) can be improved to the point where the upper and lower bounds almost match.

Theorem 3.2.3 (Norms of random matrices without boundedness assumptions)**.** *Let A be an $n \times n$ symmetric random matrix whose entries on and above the diagonal are independent, mean zero random variables. Then*

$$\mathbb{E} \max_i \|A_i\|_2 \leqslant \mathbb{E} \|A\| \leqslant C \log n \cdot \mathbb{E} \max_i \|A_i\|_2,$$

where A_i denote the rows of A.

In words, the operator norm of a random matrix is almost determined by the norm of the rows.

Our proof of this result will be based on matrix Bernstein's inequality – more precisely, Corollary 2.2.9. There is one surprising point. How can we use matrix Bernstein's inequality, which applies only for bounded distributions, to prove a result like Theorem 3.2.3 that does not have any boundedness assumptions? We will do this using a trick based on conditioning and symmetrization. Let us introduce this technique first.

Lemma 3.2.4 (Symmetrization)**.** *Let $X_1, \ldots, X_N$ be independent, mean zero random vectors in a normed space. Then*

$$\frac{1}{2} \, \mathbb{E} \, \Big\| \sum_{i=1}^{N} \varepsilon_i X_i \Big\| \leqslant \mathbb{E} \, \Big\| \sum_{i=1}^{N} X_i \Big\| \leqslant 2 \, \mathbb{E} \, \Big\| \sum_{i=1}^{N} \varepsilon_i X_i \Big\|.$$

Proof. To prove the upper bound, let (X_i') be an independent copy of the random vectors (X_i). Then

$$\mathbb{E} \, \Big\| \sum_i X_i \Big\| = \mathbb{E} \, \Big\| \sum_i X_i - \mathbb{E} \Big(\sum_i X_i' \Big) \Big\| \quad \text{(since } \mathbb{E} \sum_i X_i' = 0 \text{ by assumption)}$$

$$\leqslant \mathbb{E} \, \Big\| \sum_i X_i - \sum_i X_i' \Big\| \quad \text{(by Jensen's inequality)}$$

$$= \mathbb{E} \, \Big\| \sum_i (X_i - X_i') \Big\|.$$

The distribution of the random vectors $Y_i := X_i - X_i'$ is *symmetric*, which means that the distributions of Y_i and $-Y_i'$ are the same. (Why?) Thus the distribution of the random vectors Y_i and $\varepsilon_i Y_i$ is also the same, for all we do is change the signs of these vectors at random and independently of the values of the vectors. Summarizing, we can replace $X_i - X_i'$ in the sum above with $\varepsilon_i (X_i - X_i')$. Thus

$$\mathbb{E} \, \Big\| \sum_i X_i \Big\| \leqslant \mathbb{E} \, \Big\| \sum_i \varepsilon_i (X_i - X_i') \Big\|$$

$$\leqslant \mathbb{E}\left\|\sum_i \varepsilon_i X_i\right\| + \mathbb{E}\left\|\sum_i \varepsilon_i X_i'\right\| \quad \text{(using triangle inequality)}$$

$$= 2\,\mathbb{E}\left\|\sum_i \varepsilon_i X_i\right\| \quad \text{(the two sums have the same distribution).}$$

This proves the upper bound in the symmetrization inequality. The lower bound can be proved by a similar argument. (Do this!)   □

Proof. We already discussed the lower bound in Theorem 3.2.3. The proof of the upper bound will be based on matrix Bernstein's inequality.

First, we decompose A in the same way as we did in the proof of Theorem 2.3.5. Thus we represent A as a sum of independent, mean zero, symmetric random matrices X_{ij} each of which contains a pair of symmetric entries of A (or one diagonal entry):

$$A = \sum_{i,j:\,i\leqslant j} Z_{ij}.$$

Apply the symmetrization inequality (Lemma 3.2.4) for the random matrices Z_{ij} and get

$$(3.2.5) \qquad \mathbb{E}\,\|A\| = \mathbb{E}\left\|\sum_{i\leqslant j} Z_{ij}\right\| \leqslant 2\,\mathbb{E}\left\|\sum_{i\leqslant j} X_{ij}\right\|$$

where we set

$$X_{ij} := \varepsilon_{ij} Z_{ij}$$

and ε_{ij} are independent Rademacher random variables.

Now we condition on A. The random variables Z_{ij} become *fixed values* and all randomness remains in the Rademacher random variables ε_{ij}. Note that X_{ij} are (conditionally) *bounded* almost surely, and this is exactly what we have lacked to apply matrix Bernstein's inequality. Now we can do it. Corollary 2.2.9 gives[16]

$$(3.2.6) \qquad \mathbb{E}_\varepsilon\left\|\sum_{i\leqslant j} X_{ij}\right\| \lesssim \sigma\sqrt{\log n} + K\log n,$$

where $\sigma^2 = \left\|\sum_{i\leqslant j} \mathbb{E}_\varepsilon X_{ij}^2\right\|$ and $K = \max_{i\leqslant j}\|X_{ij}\|$.

A good exercise is to check that

$$\sigma \lesssim \max_i \|A_i\|_2 \quad \text{and} \quad K \lesssim \max_i \|A_i\|_2.$$

(Do it!) Substituting into (3.2.6), we get

$$\mathbb{E}_\varepsilon\left\|\sum_{i\leqslant j} X_{ij}\right\| \lesssim \log n \cdot \max_i \|A_i\|_2.$$

Finally, we unfix A by taking expectation of both sides of this inequality with respect to A and using the law of total expectation. The proof is complete.   □

---

[16]We stick a subscript ε to the expected value to remember that this is a conditional expectation, i.e. we average only with respect to ε_i.

We stated Theorem 3.2.3 for symmetric matrices, but it is simple to extend it to general $m \times n$ random matrices A. The bound in this case becomes

(3.2.7) $$\mathbb{E} \|A\| \leqslant C \log(m+n) \cdot \left(\mathbb{E} \max_i \|A_i\|_2 + \mathbb{E} \max_i \|A^j\|_2 \right)$$

where A_i and A^j denote the rows and columns of A. To see this, apply Theorem 3.2.3 to the $(m+n) \times (m+n)$ symmetric random matrix

$$\begin{bmatrix} 0 & A \\ A^\mathsf{T} & 0 \end{bmatrix}.$$

(Do this!)

3.3. Matrix completion   Consider a fixed, unknown $n \times n$ matrix X. Suppose we are shown m randomly chosen entries of X. Can we guess all the missing entries? This important problem is called *matrix completion*. We will analyze it using the bounds on the norms on random matrices we just obtained.

Obviously, there is no way to guess the missing entries unless we know something extra about the matrix X. So let us assume that X has *low rank*:

$$\mathrm{rank}(X) =: r \ll n.$$

The number of degrees of freedom of an $n \times n$ matrix with rank r is $O(rn)$. (Why?) So we may hope that

(3.3.1) $$m \sim rn$$

observed entries of X will be enough to determine X completely. But how?

Here we will analyze what is probably the simplest method for matrix completion. Take the matrix Y that consists of the observed entries of X while all unobserved entries are set to zero. Unlike X, the matrix Y may not have small rank. Compute the best rank r approximation[17] of Y. The result, as we will show, will be a good approximation to X.

But before we show this, let us define sampling of entries more rigorously. Assume each entry of X is shown or hidden independently of others with fixed probability p. Which entries are shown is decided by independent Bernoulli random variables

$$\delta_{ij} \sim \mathrm{Ber}(p) \quad \text{with} \quad p := \frac{m}{n^2}$$

which are often called *selectors* in this context. The value of p is chosen so that among n^2 entries of X, the expected number of selected (known) entries is m. Define the $n \times n$ matrix Y with entries

$$Y_{ij} := \delta_{ij} X_{ij}.$$

---

[17]The best rank r approximation of an $n \times n$ matrix A is a matrix B that minimizes the operator norm $\|A - B\|$ or, alternatively, the Frobenius norm $\|A - B\|_F$ (the minimizer turns out to be the same). One can compute B by truncating the singular value decomposition $A = \sum_{i=1}^n s_i u_i v_i^\mathsf{T}$ of A as follows: $B = \sum_{i=1}^r s_i u_i v_i^\mathsf{T}$, where we assume that the singular values s_i are arranged in the non-increasing order.

We can assume that we are shown Y, for it is a matrix that contains the observed entries of X while all unobserved entries are replaced with zeros. The following result shows how to estimate X based on Y.

Theorem 3.3.2 (Matrix completion)**.** *Let $\hat{X}$ be a best rank r approximation to $p^{-1}Y$. Then*

$$(3.3.3) \qquad \mathbb{E} \frac{1}{n} \|\hat{X} - X\|_F \leqslant C \log(n) \sqrt{\frac{rn}{m}} \, \|X\|_\infty,$$

Here $\|X\|_\infty = \max_{i,j} |X_{ij}|$ denotes the maximum magnitude of the entries of X.

Before we prove this result, let us understand what this bound says about the quality of matrix completion. The recovery error is measured in the Frobenius norm, and the left side of (3.3.3) is

$$\frac{1}{n} \|\hat{X} - X\|_F = \Big(\frac{1}{n^2} \sum_{i,j=1}^n |\hat{X}_{ij} - X_{ij}|^2 \Big)^{1/2}.$$

Thus Theorem 3.3.2 controls *the average error per entry* in the mean-squared sense. To make the error small, let us assume that we have a sample of size

$$m \gg rn \log^2 n,$$

which is slightly larger than the ideal size we discussed in (3.3.1). This makes $C \log(n) \sqrt{rn/m} = o(1)$ and forces the recovery error to be bounded by $o(1)\|X\|_\infty$. Summarizing, Theorem 3.3.2 says that *the expected average error per entry is much smaller than the maximal magnitude of the entry of X.* This is true for a sample of almost optimal size m. The smaller the rank r of the matrix X, the fewer entires of X we need to see in order to do matrix completion.

Proof of Theorem 3.3.2. **Step 1: The error in the operator norm.** Let us first bound the recovery error in the *operator* norm. Decompose the error into two parts using triangle inequality:

$$\|\hat{X} - X\| \leqslant \|\hat{X} - p^{-1}Y\| + \|p^{-1}Y - X\|.$$

Recall that $\hat{X}$ is a best approximation to $p^{-1}Y$. Then the first part of the error is smaller than the second part, i.e. $\|\hat{X} - p^{-1}Y\| \leqslant \|p^{-1}Y - X\|$, and we have

$$(3.3.4) \qquad \|\hat{X} - X\| \leqslant 2\|p^{-1}Y - X\| = \frac{2}{p}\|Y - pX\|.$$

The entries of the matrix $Y - pX$,

$$(Y - pX)_{ij} = (\delta_{ij} - p)X_{ij},$$

are independent and mean zero random variables. Thus we can apply the bound (3.2.7) on the norms of random matrices and get

$$(3.3.5) \qquad \mathbb{E}\,\|Y - pX\| \leqslant C \log n \cdot \Big(\mathbb{E} \max_{i \in [n]} \|(Y - pX)_i\|_2 + \mathbb{E} \max_{i \in [n]} \|(Y - pX)^j\|_2 \Big).$$

All that remains is to bound the norms of the rows and columns of $Y - pX$. This is not difficult if we note that they can be expressed as sums of independent

random variables:

$$\|(Y-pX)_i\|_2^2 = \sum_{j=1}^{n}(\delta_{ij}-p)^2 X_{ij}^2 \leqslant \sum_{j=1}^{n}(\delta_{ij}-p)^2 \cdot \|X\|_\infty^2,$$

and similarly for columns. Taking expectation and noting that $\mathbb{E}(\delta_{ij}-p)^2 = \mathrm{Var}(\delta_{ij}) = p(1-p)$, we get[18]

(3.3.6) $$\mathbb{E}\max_{i\in[n]}\|(Y-pX)_i\|_2 \leqslant (\mathbb{E}\|(Y-pX)_i\|_2^2)^{1/2} \leqslant \sqrt{pn}\,\|X\|_\infty.$$

This is a good bound, but we need something stronger in (3.3.5). Since the maximum appears inside the expectation, we need a *uniform* bound, which will say that all rows are bounded simultaneously with high probability.

Such uniform bounds are usually proved by applying concentration inequalities followed by a union bound. Bernstein's inequality (1.4.5) yields

$$\mathbb{P}\left\{\sum_{j=1}^{n}(\delta_{ij}-p)^2 > tpn\right\} \leqslant \exp(-ctpn) \quad \text{for } t \geqslant 3.$$

(Check!) This probability can be further bounded by n^{-ct} using the assumption that $m = pn^2 \geqslant n\log n$. A union bound over n rows leads to

$$\mathbb{P}\left\{\max_{i\in[n]}\sum_{j=1}^{n}(\delta_{ij}-p)^2 > tpn\right\} \leqslant n \cdot n^{-ct} \quad \text{for } t \geqslant 3.$$

Integrating this tail, we conclude using (2.2.10) that

$$\mathbb{E}\max_{i\in[n]}\sum_{j=1}^{n}(\delta_{ij}-p)^2 \lesssim pn.$$

(Check!) And this yields the desired bound on the rows,

$$\mathbb{E}\max_{i\in[n]}\|(Y-pX)_i\|_2 \lesssim \sqrt{pn},$$

which is an improvement of (3.3.6) we wanted. We can do similarly for the columns. Substituting into (3.3.5), this gives

$$\mathbb{E}\|Y-pX\| \lesssim \log(n)\sqrt{pn}\,\|X\|_\infty.$$

Then, by (3.3.4), we get

(3.3.7) $$\mathbb{E}\|\hat{X}-X\| \lesssim \log(n)\sqrt{\frac{n}{p}}\,\|X\|_\infty.$$

Step 2: Passing to Frobenius norm. Now we will need to pass from the operator to Frobenius norm. This is where we will use for the first (and only) time the rank of X. We know that $\mathrm{rank}(X) \leqslant r$ by assumption and $\mathrm{rank}(\hat{X}) \leqslant r$ by construction, so $\mathrm{rank}(\hat{X}-X) \leqslant 2r$. There is a simple relationship between the

---

[18]The first bound below that compares the L^1 and L^2 averages follows from Hölder's inequality.

operator and Frobenius norms:

$$\|\hat{X} - X\|_F \leqslant \sqrt{2r}\|\hat{X} - X\|.$$

(Check it!) Take expectation of both sides and use (3.3.7); we get

$$\mathbb{E}\,\|\hat{X} - X\|_F \leqslant \sqrt{2r}\,\mathbb{E}\,\|\hat{X} - X\| \lesssim \log(n)\sqrt{\frac{rn}{p}}\;\|X\|_\infty.$$

Dive both sides by n, we can rewrite this bound as

$$\mathbb{E}\,\frac{1}{n}\|\hat{X} - X\|_F \lesssim \log(n)\sqrt{\frac{rn}{pn^2}}\;\|X\|_\infty.$$

But $pn^2 = m$ by definition of the sampling probability p. This yields the desired bound (3.3.3). $\square$

3.4. Notes   Theorem 3.1.2 on covariance estimation is a version of [51, Corollary 5.52], see also [31]. The logarithmic factor is in general necessary. This theorem is a general-purpose result. If one knows some additional structural information about the covariance matrix (such as sparsity), then fewer samples may be needed, see e.g. [11, 15, 35].

Although the logarithmic factor in Theorem 3.2.3 can not be completely removed in general, it can be improved. Our argument actually gives

$$\mathbb{E}\,\|A\| \leqslant C\sqrt{\log n} \cdot \mathbb{E}\max_i \|A_i\|_2 + C\log n \cdot \mathbb{E}\max_{ij}|A_{ij}|,$$

Using different methods, one can save an extra $\sqrt{\log n}$ factor and show that

$$\mathbb{E}\,\|A\| \leqslant C\,\mathbb{E}\max_i \|A_i\|_2 + C\sqrt{\log n} \cdot \mathbb{E}\max_{ij}|A_{ij}|$$

(see [6]) and

$$\mathbb{E}\,\|A\| \leqslant C\sqrt{\log n \cdot \log\log n} \cdot \mathbb{E}\max_i \|A_i\|_2,$$

see [48]. (The results in [6, 48] are stated for Gaussian random matrices; the two bounds above can be deduced by using conditioning and symmetrization.) The surveys [6, 51] and the textbook [53] present several other useful techniques to bound the operator norm of random matrices.

The matrix completion problem, which we discussed in Section 3.3, has attracted a lot of recent attention. E. Candes and B. Recht [13] showed that one can often achieve *exact* matrix completion, thus computing the precise values of all missing values of a matrix, from $m \sim rn\log^2(n)$ randomly sampled entries. For exact matrix completion, one needs an extra *incoherence* assumption that is not present in Theorem 3.3.2. This assumption basically excludes matrices that are simultaneously sparse and low rank (such as a matrix whose all but one entries are zero – it would be extremely hard to complete it, since sampling will likely miss the non-zero entry). Many further results on exact matrix completion are known, e.g. [14, 17, 23, 49].

Theorem 3.3.2 with a simple proof is borrowed from [44]; see also the tutorial [52]. This result only guarantees approximate matrix completion, but it does not have any incoherence assumptions on the matrix.

4. Lecture 4: Matrix deviation inequality

In this last lecture we will study a new uniform deviation inequality for random matrices. This result will be a far reaching generalization of Johnson-Lindenstrauss Lemma we proved in Lecture 1.

Consider the same setup as in Theorem 1.6.1, where A is an $m \times n$ random matrix whose rows are independent, mean zero, isotropic and sub-gaussian random vectors in $\mathbb{R}^n$. (If you find it helpful to think in terms of concrete examples, let the entries of A be independent $N(0,1)$ random variables.) Like in Johnson-Lindenstrauss Lemma, we will be looking at A as a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$, and we will be interested in what A does to points in some set in $\mathbb{R}^n$. This time, however, we will allow for *infinite* sets $T \subset \mathbb{R}^n$.

Let us start by analyzing what A does to a single fixed vector $x \in \mathbb{R}^n$. We have

$$\mathbb{E}\,\|Ax\|_2^2 = \mathbb{E} \sum_{j=1}^m \left\langle A_j, x \right\rangle^2 \quad \text{(where } A_j^{\mathsf{T}} \text{ denote the rows of } A)$$

$$= \sum_{j=1}^m \mathbb{E} \left\langle A_j, x \right\rangle^2 \quad \text{(by linearity)}$$

$$= m\|x\|_2^2 \quad \text{(using isotropy of } A_j).$$

Further, if we believe that concentration about the mean holds here (and in fact, it does), we should expect that

$$(4.0.1) \qquad\qquad\qquad \|Ax\|_2 \approx \sqrt{m}\,\|x\|_2$$

with high probability.

Similarly to Johnson-Lindenstrauss Lemma, our next goal is to make (4.0.1) hold simultaneously over all vectors x in some fixed set $T \subset \mathbb{R}^n$. Precisely, we may ask – how large is the average *uniform deviation*:

$$(4.0.2) \qquad\qquad\qquad \mathbb{E} \sup_{x \in T} \left| \|Ax\|_2 - \sqrt{m}\,\|x\| \right| ?$$

This quantity should clearly depend on some notion of the size of T: the larger T, the larger should the uniform deviation be. So, how can we quantify the size of T for this problem? In the next section we will do precisely this – introduce a convenient, geometric measure of the sizes of sets in $\mathbb{R}^n$, which is called *Gaussian width*.

4.1. Gaussian width

Definition 4.1.1. *Let* $T \subset \mathbb{R}^n$ *be a bounded set, and* g *be a standard normal random vector in* $\mathbb{R}^n$, *i.e.* $g \sim N(0, I_n)$. *Then the quantities*

$$w(T) := \mathbb{E} \sup_{x \in T} \langle g, x \rangle \quad and \quad \gamma(T) := \mathbb{E} \sup_{x \in T} |\langle g, x \rangle|$$

are called the Gaussian width *and of* T *and the* Gaussian complexity *of* T, *respectively.*

Gaussian width and Gaussian complexity are closely related. Indeed,[19]

(4.1.2)
$$w(T) = \frac{1}{2}w(T - T) = \frac{1}{2} \mathbb{E} \sup_{x,y \in T} \langle g, x - y \rangle = \frac{1}{2} \mathbb{E} \sup_{x,y \in T} |\langle g, x - y \rangle| = \frac{1}{2}\gamma(T - T).$$

(Check these identities!)

Gaussian width has a natural geometric interpretation. Suppose g is a unit vector in $\mathbb{R}^n$. Then a moment's thought reveals that $\sup_{x,y \in T} \langle g, x - y \rangle$ is simply the *width of* T *in the direction of* g, i.e. the distance between the two hyperplanes with normal g that touch T on both sides as shown in Figure 4.1.3. Then $2w(T)$ can be obtained by averaging the width of T over all directions g in $\mathbb{R}^n$.
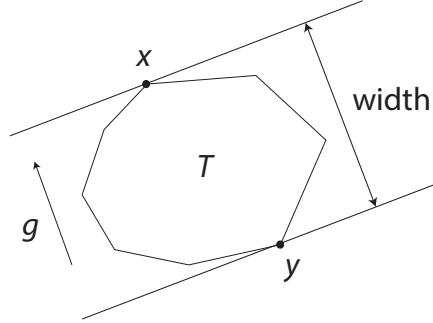

Figure 4.1.3. The width of a set T in the direction of g.

This reasoning is valid except where we assumed that g is a unit vector. Instead, for $g \sim N(0, I_n)$ we have $\mathbb{E} \|g\|_2^2 = n$ and

$$\|g\|_2 \approx \sqrt{n} \quad \text{with high probability.}$$

(Check both these claims using Bernstein's inequality.) Thus, we need to scale by the factor $\sqrt{n}$. Ultimately, the geometric interpretation of the Gaussian width becomes the following: $w(T)$ *is approximately* $2\sqrt{n}$ *larger than the usual, geometric width of* T *averaged over all directions.*

A good exercise is to compute the Gaussian width and complexity for some simple sets, such as the unit balls of the ℓ_p norms in $\mathbb{R}^n$, which we denote $B_p^n = \{x \in \mathbb{R}^n : \|x\|_p \leqslant 1\}$. In particular, we have

(4.1.4) $$\gamma(B_2^n) \sim \sqrt{n}, \quad \gamma(B_1^n) \sim \sqrt{\log n}.$$

---

[19]The set $T - T$ is defined as $\{x - y : x, y \in T\}$. More generally, given two sets A and B in the same vector space, the *Minkowski sum* of A and B is defined as $A + B = \{a - b : a \in A, b \in B\}$.

For any finite set $T \subset B_2^n$, we have

$$(4.1.5) \qquad\qquad \gamma(T) \lesssim \sqrt{\log |T|}.$$

The same holds for Gaussian width $w(T)$. (Check these facts!)

A look a these examples reveals that the Gaussian width captures some non-obvious geometric qualities of sets. Of course, the fact that the Gaussian width of the unit Euclidean ball B_2^n is or order $\sqrt{n}$ is not surprising: the usual, geometric width in all directions is 2 and the Gaussian width is about $\sqrt{n}$ times that. But it may be surprising that the Gaussian width of the ℓ_1 ball B_1^n is much smaller, and so is the width of any finite set T (unless the set has exponentially large cardinality). As we will see later, Gaussian width nicely captures the geometric size of "the bulk" of a set.

4.2. Matrix deviation inequality   Now we are ready to answer the question we asked in the beginning of this lecture: what is the magnitude of the uniform deviation (4.0.2)? The answer is surprisingly simple: it is bounded by the Gaussian complexity of T. The proof is not too simple however, and we will skip it.

Theorem 4.2.1 (Matrix deviation inequality). *Let A be an $m \times n$ matrix whose rows A_i are independent, isotropic and sub-gaussian random vectors in $\mathbb{R}^n$. Let $T \subset \mathbb{R}^n$ be a fixed bounded set. Then*

$$\mathbb{E} \sup_{x \in T} \left| \|Ax\|_2 - \sqrt{m}\|x\|_2 \right| \leqslant CK^2 \gamma(T)$$

where $K = \max_i \|A_i\|_{\psi_2}$ is the maximal sub-gaussian norm[20] of the rows of A.

Remark 4.2.2 (Tail bound). It is often useful to have results that hold *with high probability* rather than in expectation. There exists a high-probability version of matrix deviation inequality, and it states the following. Let $u \geqslant 0$. Then the event

$$(4.2.3) \qquad \sup_{x \in T} \left| \|Ax\|_2 - \sqrt{m}\|x\|_2 \right| \leqslant CK^2 \left[\gamma(T) + u \cdot \mathrm{rad}(T) \right]$$

holds with probability at least $1 - 2\exp(-u^2)$. Here $\mathrm{rad}(T)$ is the *radius* of T, defined as

$$\mathrm{rad}(T) := \sup_{x \in T} \|x\|_2.$$

Since $\mathrm{rad}(T) \lesssim \gamma(T)$ (check!) we can continue the bound (4.2.3) by

$$\lesssim K^2 u \gamma(T)$$

for all $u \geqslant 1$. This is a weaker but still a useful inequality. For example, we can use it to bound all higher moments of the deviation:

$$(4.2.4) \qquad \left(\mathbb{E} \sup_{x \in T} \left| \|Ax\|_2 - \sqrt{m}\|x\|_2 \right|^p \right)^{1/p} \leqslant C_p K^2 \gamma(T)$$

where $C_p \leqslant C\sqrt{p}$ for $p \geqslant 1$. (Check this using Proposition 1.1.1.)

---

[20]A definition of the sub-gaussian norm of a random vector was given in Section 1.5. Foe example, if A is a Gaussian random matrix with independent $N(0,1)$ entries, then K is an absolute constant.

Remark 4.2.5 (Deviation of squares). It is sometimes helpful to bound the deviation of the *square* $\|Ax\|_2^2$ rather than $\|Ax\|_2$ itself. We can easily deduce the deviation of squares by using the identity $a^2 - b^2 = (a-b)^2 + 2b(a-b)$ for $a = \|Ax\|_2$ and $b = \sqrt{m}\|x\|_2$. Doing this, we conclude that

$$(4.2.6) \qquad \mathbb{E} \sup_{x \in T} \left| \|Ax\|_2^2 - m\|x\|_2^2 \right| \leqslant CK^4 \gamma(T)^2 + CK^2 \sqrt{m}\, \mathrm{rad}(T)\gamma(T).$$

(Do this calculation using (4.2.4) for $p = 2$.) We will use this bound in Section 4.4.

Matrix deviation inequality has many consequences. We will explore some of then now.

4.3. Deriving Johnson-Lindenstrauss Lemma   We started this lecture by promising a result that is more general than Johnson-Lindenstrauss Lemma. So let us show how to quickly derive Johnson-Lindenstrauss from the matrix deviation inequality. Theorem 1.6.1 from Theorem 4.2.1.

Assume we are in the situation of Johnson-Lindenstrauss Lemma (Theorem 1.6.1). Given a set $\mathcal{X} \subset \mathbb{R}$, consider the normalized difference set

$$T := \left\{ \frac{x-y}{\|x-y\|_2} \; : \; x, y \in \mathcal{X} \text{ distinct vectors} \right\}.$$

Then T is a finite subset of the unit sphere of $\mathbb{R}^n$, and thus (4.1.5) gives

$$\gamma(T) \lesssim \sqrt{\log |T|} \leqslant \sqrt{\log |\mathcal{X}|^2} \lesssim \sqrt{\log |\mathcal{X}|}.$$

Matrix deviation inequality (Theorem 4.2.1) then yields

$$\sup_{x,y \in \mathcal{X}} \left| \frac{\|A(x-y)\|_2}{\|x-y\|_2} - \sqrt{m} \right| \lesssim \sqrt{\log N} \leqslant \varepsilon \sqrt{m}.$$

with high probability, say 0.99. (To pass from expectation to high probability, we can use Markov's inequality. To get the last bound, we use the assumption on m in Johnson-Lindenstrauss Lemma.)

Multiplying both sides by $\|x-y\|_2/\sqrt{m}$, we can write the last bound as follows. With probability at least 0.99, we have

$$(1-\varepsilon)\|x-y\|_2 \leqslant \frac{1}{\sqrt{m}}\|Ax - Ay\|_2 \leqslant (1+\varepsilon)\|x-y\|_2 \quad \text{for all } x, y \in \mathcal{X}.$$

This is exactly the consequence of Johnson-Lindenstrauss lemma.

The argument based on matrix deviation inequality, which we just gave, can be easily extended for infinite sets. It allows one to state a version of Johnson-Lindenstrauss lemma for general, possibly infinite, sets, which depends on the Gaussian complexity of T rather than cardinality. (Try to do this!)

4.4. Covariance estimation   In Section 3.1, we introduced the problem of covariance estimation, and we showed that

$$N \sim n \log n$$

samples are enough to estimate the covariance matrix of a general distribution in $\mathbb{R}^n$. We will now show how to do better if the distribution is sub-gaussian.

(Recall Section 1.5 for the definition of sub-gaussian random vectors.) In this case, we can get rid of the logarithmic oversampling and the boundedness condition (3.1.3).

Theorem 4.4.1 (Covariance estimation for sub-gaussian distributions). *Let X be a random vector in $\mathbb{R}^n$ with covariance matrix Σ. Suppose X is sub-gaussian, and more specifically*

(4.4.2) $\qquad \| \langle X, x \rangle \|_{\psi_2} \lesssim \| \langle X, x \rangle \|_{L^2} = \| \Sigma^{1/2} x \|_2 \quad$ *for any $x \in \mathbb{R}^n$.*

Then, for every $N \geqslant 1$, we have

$$\mathbb{E} \, \| \Sigma_N - \Sigma \| \lesssim \| \Sigma \| \left(\sqrt{\frac{n}{N}} + \frac{n}{N} \right).$$

This result implies that if, for $\varepsilon \in (0, 1)$, we take a sample of size

$$N \sim \varepsilon^{-2} n,$$

then we are guaranteed covariance estimation with a good relative error:

$$\mathbb{E} \, \| \Sigma_N - \Sigma \| \leqslant \varepsilon \| \Sigma \|.$$

Proof. Since we are going to use Theorem 4.2.1, we will need to first bring the random vectors $X, X_1, \ldots, X_N$ to the isotropic position. This can be done by a suitable linear transformation. You will easily check that there exists an *isotropic* random vector Z such that

$$X = \Sigma^{1/2} Z.$$

(For example, Σ has full rank, set $Z := \Sigma^{-1/2} X$. Check the general case.) Similarly, we can find independent and isotropic random vectors Z_i such that

$$X_i = \Sigma^{1/2} Z_i, \quad i = 1, \ldots, N.$$

The sub-gaussian assumption (4.4.2) then implies that

$$\| Z \|_{\psi_2} \lesssim 1.$$

(Check!) Then

$$\| \Sigma_N - \Sigma \| = \| \Sigma^{1/2} R_N \Sigma^{1/2} \| \quad \text{where} \quad R_N := \frac{1}{N} \sum_{i=1}^N Z_i Z_i^\mathsf{T} - I_n.$$

The operator norm of a symmetric $n \times n$ matrix A can be computed by maximizing the quadratic form over the unit sphere: $\| A \| = \max_{x \in S^{n-1}} | \langle Ax, x \rangle |$. (To see this, recall that the operator norm is the biggest eigenvalue of A in magnitude.) Then

$$\| \Sigma_N - \Sigma \| = \max_{x \in S^{n-1}} \left\langle \Sigma^{1/2} R_N \Sigma^{1/2} x, x \right\rangle = \max_{x \in T} \langle R_N x, x \rangle$$

where T is the ellipsoid

$$T := \Sigma^{1/2} S^{n-1}.$$

Recalling the definition of R_N, we can rewrite this as

$$\|\Sigma_N - \Sigma\| = \max_{x \in T} \left| \frac{1}{N} \sum_{i=1}^{N} \langle Z_i, x \rangle^2 - \|x\|_2^2 \right| = \frac{1}{N} \max_{x \in T} \left| \|Ax\|_2^2 - N\|x\|_2^2 \right|.$$

Now we apply the matrix deviation inequality for squares (4.2.6) and conclude that

$$\|\Sigma_N - \Sigma\| \lesssim \frac{1}{N} \left(\gamma(T)^2 + \sqrt{N}\, \mathrm{rad}(T)\gamma(T) \right).$$

(Do this calculation!) The radius and Gaussian width of the ellipsoid T are easy to compute:

$$\mathrm{rad}(T) = \|\Sigma\|^{1/2} \quad \text{and} \quad \gamma(T) \leqslant (\mathrm{tr}\, \Sigma)^{1/2}.$$

Substituting, we get

$$\|\Sigma_N - \Sigma\| \lesssim \frac{1}{N} \left(\mathrm{tr}\, \Sigma + \sqrt{N\|\Sigma\|\,\mathrm{tr}\,\Sigma} \right).$$

To complete the proof, use that $\mathrm{tr}\, \Sigma \leqslant n\|\Sigma\|$ (check this!) and simplify the bound. $\qquad\square$

Remark 4.4.3 (Low-dimensional distributions). Similarly to Section 3.1, we can show that much fewer samples are needed for covariance estimation of *low-dimensional* sub-gaussian distributions. Indeed, the proof of Theorem 4.4.1 actually yields

(4.4.4) $$\mathbb{E}\,\|\Sigma_N - \Sigma\| \leqslant \|\Sigma\| \left(\sqrt{\frac{r}{N}} + \frac{r}{N} \right)$$

where

$$r = r(\Sigma^{1/2}) = \frac{\mathrm{tr}\, \Sigma}{\|\Sigma\|}$$

is the *stable rank* of $\Sigma^{1/2}$. This means that covariance estimation for is possible with

$$N \sim r$$

samples.

4.5. Underdetermined linear equations We will give one more application of the matrix deviation inequality – this time, to the area of high dimensional inference. Suppose we need to solve a severely undetermined system of linear equations: say, we have m equations in $n \gg m$ variables. Let us write it in the matrix form as

$$y = Ax$$

where A is a given $m \times n$ matrix, $y \in \mathbb{R}^m$ is a given vector and $x \in \mathbb{R}^n$ is an unknown vector. We would like to compute x from A and y.

When the linear system is underdetermined, we can not find x with any accuracy, unless we know something extra about x. So, let us assume that we do have some a-priori information. We can describe this situation mathematically by assuming that

$$x \in K$$

where $K \subset \mathbb{R}^n$ is some known set in $\mathbb{R}^n$ that describes anything that we know about x a-priori. (Admittedly, we are operating on a high level of generality here. If you need a concrete example, we will consider it in Section 4.6.)

Summarizing, here is the problem we are trying to solve. Determine a solution $x = x(A, y, K)$ to the undetermined linear equation $y = Ax$ as accurately as possible, assuming that $x \in K$.

A variety of approaches to this and similar problems were proposed during the last decade. The one we will describe here is based on optimization. To do this, we will be convenient to convert the *set* K into a *function* on $\mathbb{R}^n$ which is called *Minkowski functional* of K. This is basically a function whose level sets are multiples of K. To define it formally, assume that K is *star-shaped*, which means that together with any point x, the set K contains the entire interval that connects x with the origin; see Figure 4.5.1 for illustration. The Minkowski functional of K is defined as

$$\|x\|_K := \inf \{t > 0 : x \in tK\}, \quad x \in \mathbb{R}^n.$$

If the set K is convex and symmetric about the origin, $\|x\|_K$ is actually a *norm* on $\mathbb{R}^n$. (Check this!)
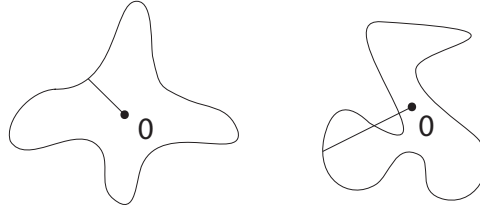

Figure 4.5.1. The set on the left (whose boundary is shown) is star-shaped, the set on the right is not.

Now we propose the following way to solve the recovery problem: solve the optimization program

(4.5.2) $\min \|x'\|_K \quad \text{subject to} \quad y = Ax'.$

Note that this is a very natural program: it looks at all solutions to the equation $y = Ax'$ and tries to "shrink" the solution x' toward K. (This is what minimization of Minkowski functional is about.)

Also note that if K is convex, this is a *convex optimization* program, and thus can be solved effectively by one of the many available numeric algorithms.

The main question we should now be asking is – would the solution to this program approximate the original vector x? The following result bounds the approximation error for a *probabilistic model* of linear equations. Assume that A is a random matrix as in Theorem 4.2.1.

Theorem 4.5.3 (Recovery by optimization). *The solution $\hat{x}$ of the optimization program* (4.5.2) *satisfies*[21]

$$\mathbb{E}\,\|\hat{x} - x\|_2 \lesssim \frac{w(K)}{\sqrt{m}},$$

where $w(K)$ is the Gaussian width of K.

Proof. Both the original vector x and the solution $\hat{x}$ are feasible vectors for the optimization program (4.5.2). Then

$$\|\hat{x}\|_K \leqslant \|x\|_K \quad \text{(since } \hat{x} \text{ minimizes the Minkowski functional)}$$
$$\leqslant 1 \quad \text{(since } x \in K\text{)}.$$

Thus both $\hat{x}, x \in K$.

We also know that $A\hat{x} = Ax = y$, which yields

$$(4.5.4) \qquad\qquad\qquad\qquad A(\hat{x} - x) = 0.$$

Let us apply matrix deviation inequality (Theorem 4.2.1) for $T := K - K$. It gives

$$\mathbb{E}\sup_{u,v \in K}\left|\,\|A(u-v)\|_2 - \sqrt{m}\,\|u-v\|_2\right| \lesssim \gamma(K-K) = 2w(K),$$

where we used (4.1.2) in the last identity. Substitute $u = \hat{x}$ and $v = x$ here. We may do this since, as we noted above, both these vectors belong to K. But then the term $\|A(u-v)\|_2$ will be equal zero by (4.5.4). It disappears from the bound, and we get

$$\mathbb{E}\,\sqrt{m}\,\|\hat{x} - x\|_2 \lesssim w(K).$$

Dividing both sides by $\sqrt{m}$ we complete the proof. $\square$

Theorem 4.5.3 says that a signal $x \in K$ can be efficiently recovered from

$$m \sim w(K)^2$$

random linear measurements.

4.6. Sparse recovery   Let us illustrate Theorem 4.5.3 with an important specific example of the feasible set K.

Suppose we know that the signal x is *sparse*, which means that only few coordinates of x are nonzero. As before, our task is to recover x from the random linear measurements given by the vector

$$y = Ax,$$

where A is an $m \times n$ random matrix. This is a basic example of *sparse recovery problems*, which are ubiquitous in various disciplines.

The number of nonzero coefficients of a vector $x \in \mathbb{R}^n$, or the sparsity of x, is often denoted $\|x\|_0$. This is similar to the notation for the ℓ_p norm $\|x\|_p =$

---

[21]Here and in other similar results, the notation $\lesssim$ will hide possible dependence on the sub-gaussian norms of the rows of A.

$(\sum_{i=1}^{n} |x_i|^p)^{1/p}$, and for a reason. You can quickly check that

(4.6.1) $$\|x\|_0 = \lim_{p \to 0} \|x\|_p$$

(Do this!) Keep in mind that neither $\|x\|_0$ nor $\|x\|_p$ for $0 < p < 1$ are actually *norms* on $\mathbb{R}^n$, since they fail triangle inequality. (Give an example.)

Let us go back to the sparse recovery problem. Our first attempt to recover x is to try the following optimization problem:

(4.6.2) $$\min \|x'\|_0 \quad \text{subject to} \quad y = Ax'.$$

This is sensible because this program selects the sparsest feasible solution. One can show that the program performs well. But there is an implementation caveat: the function $f(x) = \|x\|_0$ is highly non-convex and even discontinuous. There is simply no known algorithm to solve the optimization problem (4.6.2) efficiently.

To overcome this difficulty, let us turn to the relation (4.6.1) for an inspiration. What if we replace $\|x\|_0$ in the optimization problem (4.6.2) by $\|x\|_p$ with $p > 0$? The smallest p for which $f(x) = \|x\|_p$ is a genuine norm (and thus a convex function on $\mathbb{R}^n$) is $p = 1$. So let us try

(4.6.3) $$\min \|x'\|_1 \quad \text{subject to} \quad y = Ax'.$$

This is a *convexification* of the non-convex program (4.6.2), and a variety of numeric convex optimization methods are available to solve it efficiently.

We will now show that ℓ_1 minimization works nicely for sparse recovery. As before, we assume that A is a random matrix as in Theorem 4.2.1.

Theorem 4.6.4 (Sparse recovery by optimization). *Assume that an unknown vector* $x \in \mathbb{R}^n$ *has at most s non-zero coordinates, i.e.* $\|x\|_0 \leqslant s$. *The solution $\hat{x}$ of the optimization program* (4.6.3) *satisfies*

$$\mathbb{E} \|\hat{x} - x\|_2 \lesssim \sqrt{\frac{s \log n}{m}} \, \|x\|_2.$$

Proof. Since $\|x\|_0 \leqslant s$, Cauchy-Schwarz inequality shows that

(4.6.5) $$\|x\|_1 \leqslant \sqrt{s} \, \|x\|_2.$$

(Check!) Denote the unit ball of the ℓ_1 norm in $\mathbb{R}^n$ by B_1^n, i.e. $B_1^n := \{x \in \mathbb{R}^n : \|x\|_1 \leqslant 1\}$. Then we can rewrite (4.6.5) as the inclusion

$$x \in \sqrt{s} \, \|x\|_2 \cdot B_1^n := K.$$

Apply Theorem 4.5.3 for this set K. We noted the Gaussian width of B_1^n in (4.1.4), so

$$w(K) = \sqrt{s} \, \|x\|_2 \cdot w(B_1^n) \leqslant \sqrt{s} \, \|x\|_2 \cdot \gamma(B_1^n) \leqslant \sqrt{s} \, \|x\|_2 \cdot \sqrt{\log n}.$$

Substitute this is Theorem 4.5.3 and complete the proof. $\qquad\qquad\qquad\square$

Theorem 4.6.4 says that an s-sparse signal $x \in \mathbb{R}^n$ can be efficiently recovered from

$$m \sim s \log n$$

random linear measurements.

4.7. Notes   For a more thorough introduction to Gaussian width and its role in high-dimensional estimation, refer to the tutorial [52] and the textbook [53]; see also [5].

Matrix deviation inequality (Theorem 4.2.1) is borrowed from [36]. In the special case where A is a Gaussian random matrix, this result follows from the work of G. Schechtman [50]. In the general case of sub-gaussian distributions, earlier variants of Theorem 4.2.1 were proved by B. Klartag and S. Mendelson [30], S. Mendelson, A. Pajor and N. Tomczak-Jaegermann [39] and S. Dirksen [19].

Theorem 4.4.1 for covariance estimation can be proved alternatively using more elementary tools (Bernstein's inequality and ε-nets), see [51]. However, no known elementary approach exists for the *low-rank* covariance estimation discussed in Remark 4.4.3. The bound (4.4.4) was proved by V. Koltchinskii and K. Lounici [31] by a different method.

In Section 4.5, we scratched the surface of a recently developed area of *sparse signal recovery*, which is also called *compressed sensing*. Our presentation there essentially follows the tutorial [52]. Theorem 4.6.4 can be improved: if we take

$$m \gtrsim s \log(n/s)$$

measurements, then with high probability the optimization program (4.6.3) recovers the unknown signal x *exactly*, i.e.

$$\hat{x} = x.$$

First results of this kind were proved by J. Romberg, E. Candes and T. Tao [12] and a great number of further developments followed; refer e.g. to the book [22] and the chapter in [18] for an introduction into this reach area.

References

[1] E. Abbe, A. S. Bandeira, G. Hall. *Exact recovery in the stochastic block model*, IEEE Transactions on Information Theory 62 (2016), 471–487. 18

[2] D. Achlioptas, *Database-friendly random projections: Johnson-Lindenstrauss with binary coins*, Journal of Computer and System Sciences, 66 (2003), 671–687. 9

[3] R. Ahlswede, A. Winter, *Strong converse for identification via quantum channels*, IEEE Trans. Inf. Theory 48 (2002), 569–579. 17

[4] N. Ailon, B. Chazelle, *Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform*, Proceedings of the 38th Annual ACM Symposium on Theory of Computing. New York: ACM Press, 2006. pp. 557–563. 9

[5] D. Amelunxen, M. Lotz, M. McCoy, J. Tropp, *Living on the edge: phase transitions in convex programs with random data,* Inf. Inference 3 (2014), 224–294. 37

[6] A. Bandeira, R. van Handel, *Sharp nonasymptotic bounds on the norm of random matrices with independent entries*, Ann. Probab. 44 (2016), 2479–2506. 18, 28

[7] R. Baraniuk, M. Davenport, R. DeVore, M. Wakin, *A simple proof of the restricted isometry property for random matrices,* Constructive Approximation, 28 (2008), 253–263. 9

[8] R. Bhatia, *Matrix Analysis.* Graduate Texts in Mathematics, vol. 169. Springer, Berlin, 1997. 18

[9] C. Bordenave, M. Lelarge, L. Massoulie, ǀem Non-backtracking spectrum of random graphs: community detection and non-regular Ramanujan graphs, Annals of Probability, to appear. 18

[10] S. Boucheron, G. Lugosi, P. Massart, *Concentration inequalities. A nonasymptotic theory of independence.* With a foreword by Michel Ledoux. Oxford University Press, Oxford, 2013. 8, 9

[11] T. Cai, R. Zhao, H. Zhou, *Estimating structured high-dimensional covariance and precision matrices: optimal rates and adaptive estimation,* Electron. J. Stat. 10 (2016), 1–59. 27

[12] E. Candes, J. Romberg, T. Tao, *Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information,* IEEE Trans. Inform. Theory 52 (2006), 489–509. 38

[13] E. Candes, B. Recht, *Exact Matrix Completion via Convex Optimization,* Foundations of Computational Mathematics 9 (2009), 717–772. 28

[14] E. Candes, T. Tao, *The power of convex relaxation: near-optimal matrix completion,* IEEE Trans. Inform. Theory 56 (2010), 2053–2080. 28

[15] R. Chen, A. Gittens, J. Tropp, *The masked sample covariance estimator: an analysis using matrix concentration inequalities,* Inf. Inference 1 (2012), 2–20. 27

[16] P. Chin, A. Rao, and V. Vu, lem Stochastic block model and community detection in the sparse graphs: A spectral algorithm with optimal rate of recovery, preprint, 2015. 18

[17] M. Davenport, Y. Plan, E. van den Berg, M. Wootters, *1-bit matrix completion,* Inf. Inference 3 (2014), 189–223. 28

[18] M. Davenport, M. Duarte, Yonina C. Eldar, Gitta Kutyniok, *Introduction to compressed sensing,* in: *Compressed sensing.* Edited by Yonina C. Eldar and Gitta Kutyniok. Cambridge University Press, Cambridge, 2012. 38

[19] S. Dirksen, *Tail bounds via generic chaining,* Electron. J. Probab. 20 (2015), 1–29. 37

[20] U. Feige, E. Ofek, *Spectral techniques applied to sparse random graphs,* Random Structures Algorithms 27 (2005), 251–275. 18

[21] S. Fortunato, Santo; D. Hric, *Community detection in networks: A user guide.* Phys. Rep. 659 (2016), 1–44. 18

[22] S. Foucart, H. Rauhut, *A mathematical introduction to compressive sensing.* Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York, 2013. 38

[23] D. Gross, *Recovering low-rank matrices from few coefficients in any basis,* IEEE Trans. Inform. Theory 57 (2011), 1548–1566. 28

[24] O. Guedon, R. Vershynin, *Community detection in sparse networks via Grothendieck's inequality,* Probability Theory and Related Fields 165 (2016), 1025–1049. 18

[25] A. Javanmard, A. Montanari, F. Ricci-Tersenghi, *Phase transitions in semidefinite relaxations,* PNAS, April 19, 2016, vol. 113, no.16, E2218–E2223. 18

[26] W. B. Johnson, J. Lindenstrauss, *Extensions of Lipschitz mappings into a Hilbert space.* In Beals, Richard; Beck, Anatole; Bellow, Alexandra; et al. Conference in modern analysis and probability (New Haven, Conn., 1982). Contemporary Mathematics. 26. Providence, RI: American Mathematical Society, 1984. pp. 189–206. 8

[27] B. Hajek, Y. Wu, J. Xu, *Achieving exact cluster recovery threshold via semidefinite programming,* IEEE Transactions on Information Theory 62 (2016), 2788–2797. 18

[28] P. W. Holland, K. B. Laskey, S. Leinhardt, lem Stochastic blockmodels: first steps, Social Networks 5 (1983), 109–137. 18

[29] D. Kane, J. Nelson, *Sparser Johnson-Lindenstrauss Transforms,* Journal of the ACM 61 (2014): 1. 9

[30] B. Klartag, S. Mendelson, *Empirical processes and random projections,* J. Funct. Anal. 225 (2005), 229–245. 37

[31] V. Koltchinskii, K. Lounici, *Concentration inequalities and moment bounds for sample covariance operators,* Bernoulli 23 (2017), 110–133. 27, 37

[32] C. Le, E. Levina, R. Vershynin, *Concentration and regularization of random graphs,* Random Structures and Algorithms, to appear. 18

[33] M. Ledoux, *The concentration of measure phenomenon.* American Mathematical Society, Providence, RI, 2001. 8

[34] M. Ledoux, M. Talagrand, *Probability in Banach spaces. Isoperimetry and processes.* Springer-Verlag, Berlin, 1991. 8

[35] E. Levina, R. Vershynin, *Partial estimation of covariance matrices,* Probability Theory and Related Fields 153 (2012), 405–419. 27

[36] C. Liaw, A. Mehrabian, Y. Plan, R. Vershynin, *A simple tool for bounding the deviation of random matrices on geometric sets,* Geometric Aspects of Functional Analysis, Lecture Notes in Mathematics, Springer, Berlin, to appear. 37

[37] J. Matouĺek, *Lectures on discrete geometry.* Graduate Texts in Mathematics, 212. Springer-Verlag, New York, 2002. 9

[38] F. McSherry, *Spectral partitioning of random graphs*, Proc. 42nd FOCS (2001), 529–537. 18

[39] S. Mendelson, A. Pajor, N. Tomczak-Jaegermann, *Reconstruction and subgaussian operators in asymptotic geometric analysis.* Geom. Funct. Anal. 17 (2007), 1248–1282. 37

[40] E. Mossel, J. Neeman, A. Sly, *Belief propagation, robust reconstruction and optimal recovery of block models.* Ann. Appl. Probab. 26 (2016), 2211–2256. 18

[41] M. E. Newman, *Networks. An introduction.* Oxford University Press, Oxford, 2010. 18

[42] R. I. Oliveira, *Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges*, unpublished (2010), arXiv:0911.0600. 17, 18

[43] R. I. Oliveira, *Sums of random Hermitian matrices and an inequality by Rudelson*, Electron. Commun. Probab. 15 (2010), 203–212. 17

[44] Y. Plan, R. Vershynin, E. Yudovina, *High-dimensional estimation with geometric constraints,* Information and Inference 0 (2016), 1–40. 28

[45] J. Tropp, *User-friendly tail bounds for sums of random matrices.* Found. Comput. Math. 12 (2012), 389–434. 18

[46] J. Tropp, *An introduction to matrix concentration inequalities.* Found. Trends Mach. Learning 8 (2015), 10-230. 18

[47] R. van Handel, *Structured random matrices.* in: IMA Volume "Discrete Structures: Analysis and Applications", Springer, to appear.

[48] R. van Handel, *On the spectral norm of Gaussian random matrices,* Trans. Amer. Math. Soc., to appear. 28

[49] B. Recht, *A simpler approach to matrix completion,* J. Mach. Learn. Res. 12 (2011), 3413–3430. 28

[50] G. Schechtman, *Two observations regarding embedding subsets of Euclidean spaces in normed spaces,* Adv. Math. 200 (2006), 125–135. 37

[51] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices.* Compressed sensing, 210–268, Cambridge University Press, Cambridge, 2012. 1, 8, 27, 28, 37

[52] R. Vershynin, *Estimation in high dimensions: a geometric perspective.* Sampling Theory, a Renaissance, 3–66, Birkhauser Basel, 2015. 1, 28, 37

[53] R. Vershynin, *High-Dimensional Probability. An Introduction with Applications in Data Science.* Cambridge University Press, to appear. 1, 8, 28, 37

[54] H. Zhou, A. Zhang, *Minimax Rates of Community Detection in Stochastic Block Models,* Annals of Statistics, to appear. 18

Department of Mathematics, University of Michigan, 530 Church Street, Ann Arbor, MI 48109, U.S.A.

E-mail address: romanv@umich.edu