EEN100: Statistics and machine learning in high dimensions

Giuseppe Durisi
Chalmers University of Technology

September 13, 2023

Contents

1 Introduction

1.1 Course description

The explosion in the volume of data collected in all scientific disciplines and in industry requires students interested in statistical analyses and machine-learning and signal-processing algorithms to acquire more sophisticated probability tools than the ones taught in basic probability courses.

This course provides an introduction to the area of high-dimensional statistics, which deals with large scale problems where both the number of parameters and the sample size is large.

The course covers fundamental tools for the analysis of random vectors, random matrices, and random projections, such as tail bounds and concentration inequalities. It further provides practically relevant applications of such tools in the context of sparse linear models, matrix models with rank constraints, community detection, principal component analyses, clustering, and sample complexity in statistical learning theory.

1.2 Literature

This course is mainly based on the following five books:

- Vershynin, High-dimensional probability: an introduction with applications in data science (2019). Available online.
- Wainwright, High-dimensional statistics: a nonasymptotic viewpoint (2019). Available online via Chalmers library.
- Bandeira, Singer, and Strohmer, Mathematics of Data Science (2020) Available online.
- Foucart and Rauhut, A mathematical introduction to compressive sensing (2013). Available online through Chalmers library
- Shalev-Shwartz and Ben-David, Understanding machine learning: from theory to algorithms (2014). Available online through Chalmers library

These notes borrow heavily from all five sources, while trying to keep the notation consistent.

1.3 Prerequisites

- We will often use basic results from probability theory, such as union bound. A good overview of the basic results that will be needed in this course can be found in the first chapters of [1].
- We will also use basic results in linear algebra related to matrix decomposition and vector and matrix norms. A good reference is [2].

We will assume that the students are familiar with this material.

1.4 Some comments on these lecture notes

Proofs are mostly omitted at this stage of the draft. Most of them will be provided during the lectures and some of them are sketched in the course slides. Finally, some of them will be covered in the homework assignments.

2 Surprises in high dimension

Why is the high-dimension regime *special*? Why may standard approaches in statistics and machine learning fail to capture the peculiarities of this regime? To illustrate why this is the case, we will describe in this chapter some phenomena occurring in high dimensions, which are somewhat counterintuitive.

2.1 The curse of dimensionality

Curse of dimensionality: many algorithmic approaches to problems in $\mathbb{R}^d$ become exponentially more difficult as d grows.

Example: if we want to sample uniformly the unit interval so that the distance between adjacent points is at most 0.01, it suffices to have 100 evenly-spaced points. If we now want to achieve the same result when sampling uniformly a five-dimensional unit cube, we need 10^{10} sample points.

Punchline: a modest increase in dimensions results in a dramatic increase in data points to cover the space at the same density.

2.2 Geometric surprises in high dimension

- Source: Chapter 2 of [3].

2.2.1 Geometry of spheres and balls in high dimension

Our intuition about space is based on two and three dimensions, and can often be misleading when we move to high dimensions. Properties of even very basic objects become counterintuitive in high dimensions. It is important to be aware of this to avoid pitfalls when designing machine-learning algorithms and statistical methods for high-dimensional data.

Let's study some of the properties of the ball and the cube, two objects we are very familiar with in 3 dimensions, as the number of dimensions increases. The d-dimensional ball of radius R is defined by

$$B^d(R) = \{x \in \mathbb{R}^d : x_1^2 + \cdots + x_d^2 \leq R^2\}. \tag{1}$$

The d-dimensional sphere of radius R is given by

$$S^{d-1}(R) = \{x \in \mathbb{R}^d : x_1^2 + \ldots x_d^2 = R^2\}. \tag{2}$$

Finally, the d-dimensional cube with side length $2R$ is the subset of $\mathbb{R}^d$ defined as

$$C^d(R) = \underbrace{[-R, R] \times \cdots \times [-R, R]}_{d \text{ times}}. \tag{3}$$

To keep notation compact, we set $B^d(1) = B^d$, $S^{d-1}(1) = S^{d-1}$, and $C^d(1/2) = C^d$. Here are some surprising results about these objects.

The volume of $B^d(1)$ vanishes as d grows

Theorem 1 *The volume of $B^d(R)$ is given by*

$$\mathrm{Vol}(B^d(R)) = \frac{\pi^{d/2} R^d}{\Gamma(d/2+1)}. \tag{4}$$

Here, $\Gamma(z)$ is the so-called Gamma function, defined as

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \mathrm{d}x. \tag{5}$$

When z is a positive integer, one can show that $\Gamma(z) = (z-1)!$. To obtain some insights on the behavior of (4), we use Stirling's formula, a well-known approximation for the factorial function. It states that for every integer n,

$$n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n. \tag{6}$$

Here, we used the convention that $f(n) \sim g(n)$ means that $\lim_{n\to\infty} |f(n)/g(n)| = 1$. Using (6), we conclude that, when $d/2$ is an integer,

$$\Gamma\left(\frac{d}{2}+1\right) \sim \sqrt{\pi d} \left(\frac{d}{2e}\right)^{d/2}. \tag{7}$$

We now substitute (7) into (4) and conclude that we can approximate the volume of the **unit** d-ball for large d as

$$\mathrm{Vol}(B^d) \sim \frac{1}{\sqrt{d\pi}} \left(\frac{2\pi e}{d}\right)^{d/2} = \frac{1}{\sqrt{d\pi}} e^{-(d/2)\log(d/(2\pi e))}. \tag{8}$$

Note that the right-hand side vanishes when $d \to \infty$. This implies the following perhaps surprising result: the unit ball B^d has vanishing volume as d grows large. The next two results provide some intuition on why this is the case.

The volume of $B^d(1)$ is concentrated close to the equator. Let us now study where most of the volume is concentrated. If we cut a 3D ball in horizontal slabs of the same thickness, then we know that the slide in the middle will be the largest. This effect increases dramatically when the number of dimensions increases.

Theorem 2 *Fix $p \in [0,1]$ and let $P = \{x \in \mathbb{R}^d : \|x\| \le 1, x_1 > p\}$ denote the polar cap, i.e., the part of the ball $B^d(1)$ above the slab of width $2p$ at the equator. Then, for sufficiently large d,*

$$\frac{2\mathrm{Vol}(P)}{\mathrm{Vol}(B^d)} \le e^{-\frac{d-1}{2}p^2}. \tag{9}$$

In words, no matter how small p is, the ratio between the volume of both polar caps and the total volume of the ball goes to zero, which implies that all volume concentrates close to the equator. Note that the decay is exponential in d.

The volume of the ball is concentrated on a shell. Consider two concentric balls $B^d(1)$ and $B^d(1-\epsilon)$ for some arbitrary $\epsilon \in (0,1)$. It follows from (4) that

$$\frac{\mathrm{Vol}(B^d(1-\epsilon))}{\mathrm{Vol}(B^d(1))} = (1-\epsilon)^d. \tag{10}$$

Note now that for every $\epsilon \in (0,1)$, we have that $(1-\epsilon)^d \to 0$ as $d \to \infty$. This means that the spherical shell given by the region between $B^d(1)$ and $B^d(1-\epsilon)$ contains most of the volume of $B^d(1)$ for large enough d, no matter how small ϵ is. In other words, most of the volume of the ball $B^d(1)$ is "concentrated on the surface".

Geometry of the cube in $\mathbb{R}^d$. The cube in $\mathbb{R}^d$ exhibits an even more interesting volume-concentration behavior. We will highlight some strange phenomena occurring as d grows, deferring the proof of these phenomena to later in the course. The proofs will rely on concentration of measure techniques in high-dimensional probability, which will one of the main new tools introduced in this course.

Let us start with a somewhat trivial observation: the cube $C^d(1/2)$ has volume 1 and diameter, $\sqrt{d}$ defined as maximum distance between two points. It is easy to verify that, in 2 dimensions, $C^2(1/2)$, which is a square of side 1, is contained in $B^d(1)$, which is a circle of radius 1. Furthermore, $B^d(1/2)$ is the largest ball that is inscribed in the square. Indeed, the diameter is $\sqrt{2}$ and each vertex is $\sqrt{2}/2$ away from the center. But as d increases, each vertex of the cube moves to a distance $\sqrt{d}/2$ from the center. This means that when $d > 5$, the vertexes of the cube are no longer contained in $B^d(1)$. However, the largest ball inscribed in the cube has still radius $1/2$. To summarize, the maximum distance between two points increases with the dimension d, but the largest ball that can be inscribed in the cube has radius that does not grow with d. This implies that cubes in high dimension are somewhat "pointy", despite being convex. Indeed, one can prove that most of the volume of a cube in high dimension is located around its vertexes. We shall prove this result in Section 3.5.2.

3 Tail bounds and concentration of measure

3.1 Basics concepts from probability

- Source: Vershynin, Section 1.1.
- Mean of random variable X: $\mathbb{E}[X]$.
- Variance: $\mathbb{V}\mathrm{ar}[X]$.
- Moment generating function: $M_X(t) = \mathbb{E}[e^{tX}]$, $t \in \mathbb{R}$. If $M_X(t)$ exists in a neighborhood $(-b, b)$ of $t = 0$, then

$$M_X(t) = \sum_{k=0}^{\infty} \frac{\mathbb{E}\left[X^k\right] t^k}{k!}. \tag{11}$$

  This means that we can obtain all moments $\mathbb{E}\left[X^k\right]$ of X simply by evaluating at zero the derivative of the corresponding order of $M_X(t)$. Hence, the name of $M_X(t)$.

- L^p norm of a random variable: $\|X\|_{L^p} = \mathbb{E}[|X|^p]^{1/p}$, $p \in [0, \infty]$, with the usual extension $\|X\|_{L^\infty} = \operatorname{ess\,sup} |X|$. Note: strictly speaking, this quantity is a norm only when $p \geq 1$.
- L^p is sometimes used also to indicate the space of all random variables with finite L^p norm.
- Standard deviation: $\sigma(X) = \sqrt{\mathbb{V}\mathrm{ar}[X]}$
- Covariance of the random variables X and Y: $\mathbb{C}\mathrm{ov}[X, Y] = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$.
- Union bound: let $\mathcal{A}_1, \mathcal{A}_2, \ldots$ be a infinite countable set of events; then

$$\mathbb{P}\left[\bigcup_{i=1}^{\infty} \mathcal{A}_i\right] \leq \sum_{i=1}^{\infty} \mathbb{P}[\mathcal{A}_i]. \tag{12}$$

3.2 Some classical inequalities

- Source: Vershynin, Section 1.3.
- **Hölder's inequality**: given the random variables X and Y and $p, q \geq 1$ with $1/p + 1/q = 1$,

$$|\mathbb{E}[XY]| \leq \|X\|_{L^p} \|Y\|_{L^q}. \tag{13}$$

  The special case $p = q = 2$ is known as the **Chauchy-Schwarz inequality**
- **Jensen's inequality**: For any real-valued random variable X and convex function $\psi : \mathbb{R} \to \mathbb{R}$,

$$\psi(\mathbb{E}[X]) \leq \mathbb{E}[\psi(X)]. \tag{14}$$

Note: since $\psi(x) = x^{q/p}$ is a convex function for $q \geq p \geq 0$, it follows from Jensen's inequality that

$$\|X\|_{L_p} \leq \|X\|_{L_q}. \tag{15}$$

- **Minkovskii's inequality**: for every $p \in [1, \infty]$ and every random variables X, Y we have

$$\|X + Y\|_{L_p} \leq \|X\|_{L_p} + \|Y\|_{L_p}. \tag{16}$$

- The cumulative distribution of X is defined as $F_X(t) = \mathbb{P}[X \leq t]$, $t \in \mathbb{R}$. The tail of X is the function $t \to \mathbb{P}[X \geq t]$.

The next theorem establishes as useful connection between expectation and tails.

- **Integral identity**: Let X be a **non-negative** random variable. Then

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}[X > t] \mathrm{d}t. \tag{17}$$

- **Markov's inequality**: For any nonnegative random varibale X, we have that

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}. \tag{18}$$

  This inequality is often too weak to give good tail bounds. As we shall see, it is however the key tool to obtain much more powerful inequalities.

- **Chebyshev's inequality**: it is a direct consequence of Markov's inequality. Let X be a random variable with mean μ and variance σ^2. Then for all $t > 0$,

$$\mathbb{P}[|X - \mu| \geq t] \leq \frac{\sigma^2}{t^2}. \tag{19}$$

Chebyshev's inequality is a simple form of concentration inequality: it guarantees that X must be close to its mean μ when the variance σ^2 of X is small. Both Markov and Chebyshev are sharp inequalities, in the sense that there exist random variables for which they hold with equality. Note that Markov's inequality requires only the existence of the mean of a random variable, whereas Chebyshev's inequality requires the existence of the second moment. One can of course generalize these inequalities to higher moments, provided that they exist.

- **Chernoff bound**: Assume that the moment generating function of X exists in a neighbor around zero, i.e., there exists some constant $b > 0$ such that $\mathbb{E}[e^{\lambda(X-\mu)}]$ exists for all $\lambda \in [-b, b]$. In this case, by applying Markov's inequality to the random variable $Y = e^{\lambda(X-\mu)}$, for $\lambda \in [0, b]$ we obtain the Chernoff bound

$$\mathbb{P}[(X-\mu) \geq t] \leq \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda t}}. \tag{20}$$

Observe that λ needs to be non-negative for this inequality to hold. The tightest bound is obtained by optimizing over λ:

$$\log \mathbb{P}[(X-\mu) \geq t] \leq \min_{\lambda \in [0,b]} \{\log \mathbb{E}[e^{\lambda(X-\mu)}] - \lambda t\}. \tag{21}$$

As we shall see later in the course, a variety of very important concentration bounds follow directly from Chernoff bound.

3.3 Sub-Gaussian random variables

- Source: Section 2.2.1 of [4].

As we have just shown, Chernoff bound depends on the moment generating function. Since Chernoff bound is an important tool in obtaining tail bounds in high-dimensional statistics, it is natural to classify random variables in terms of their moment generating function. An important class of random variables is that of **sub-Gaussian** random variables. To introduce this class of random variables, it is instructive to first discuss in details the case of Gaussian random variables.

Gaussian tail bounds: Let $X \sim \mathcal{N}(\mu, \sigma^2)$. The moment generating function of $X - \mu$ is given by

$$\mathbb{E}[e^{\lambda(X-\mu)}] = e^{\frac{\sigma^2\lambda^2}{2}} \tag{22}$$

which is valid for all $\lambda \in \mathbb{R}$. Substituting this expression in the Chernoff bound (20), we get

$$\inf_{\lambda \geq 0} \left\{ \log \mathbb{E}[e^{\lambda(X-\mu)}] - \lambda t \right\} = -\frac{t^2}{2\sigma^2} \tag{23}$$

This means that $\mathcal{N}(\mu, \sigma^2)$ random variables satisfy the following upper deviation inequality

$$\mathbb{P}[(X-\mu) \geq t] \leq e^{-t^2/(2\sigma^2)}, \quad \forall t \geq 0. \tag{24}$$

This bound is actually fairly tight, as shown in Fig. 1, although it can be improved (see Exercise 9). It turns out that tail bounds of the same form as (24) can be obtained for non-Gaussian random variables, as long as they have a moment-generating function that can be upper-bounded by that of a Gaussian random variable. This motivates the following definition.

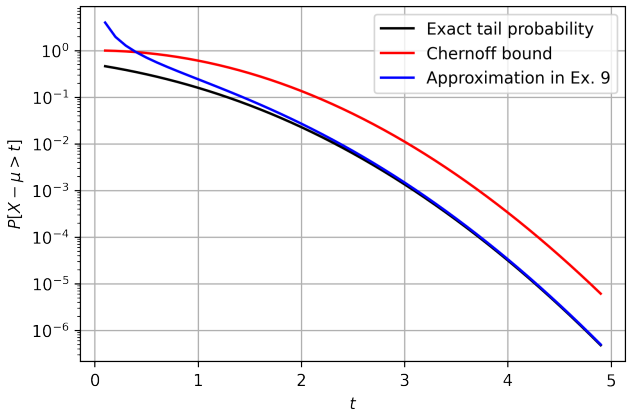

Figure 1: Exact tail bound of a $\mathcal{N}(0,1)$ random variable as well as the Chernoff bound (24) and the approximation suggested in Exercise 9.

Definition 3 (Sub-Gaussian random variables) *A random variable X with mean $\mu = \mathbb{E}[X]$ is sub-Gaussian if there exists a positive constant σ such that*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\sigma^2\lambda^2}{2}}, \quad \forall \lambda \in \mathbb{R}. \tag{25}$$

The constant σ is a proxy of the standard deviation in the Gaussian case. In fact, it is possible to verify that if X is sub-Gaussian according to (25), then $\mathbb{V}\mathrm{ar}[X] \leq \sigma^2$. We will refer to σ as the *sub-Gaussian parameter*. Note that if (25) holds for some σ, it also holds for all $\gamma \geq \sigma$. In this course, we will typically not be interested in determining the smallest σ for which (25) holds.

It is easy to check that the upper-deviation inequality (24) holds for all σ sub-Gaussian random variables. We state this result in the following theorem.

Theorem 4 (sub-Gaussian upper deviation inequality) *Assume that X is sub-Gaussian with parameter σ. Then*

$$\mathbb{P}[(X-\mu) \geq t] \leq e^{-t^2/(2\sigma^2)}, \quad \forall t \geq 0. \tag{26}$$

Note that if X is σ sub-Gaussian, then $-X$ is also σ sub-Gaussian. It then follows from the union bound applied to the probability of the event $\{(X-\mu) \geq t\} \cup \{(X-\mu) \leq -t\}$ that

$$\mathbb{P}[|X-\mu| \geq t] \leq 2e^{-t^2/(2\sigma^2)}, \quad \forall t \geq 0. \tag{27}$$

Note that **bounded random variables** are sub-Gaussian. Specifically, assume that $X \in [a,b]$ with probability 1. Then X is sub-Gaussian with parameter $(b-a)/2$.

3.4 Sub-exponential random variables

- Source: Wainwright, Section 2.1.3.

The class of sub-Gaussian random variables includes many important families of random variables. However, several probability distribution with heavier tails, which occur frequently, are not in this class. A typical example is the so-called *double-sided exponential distribution*, also known as *Laplace distribution*. This distribution has the following probability density function:

$$f_X(x) = \frac{1}{2b} e^{-|x-\mu|/b}. \tag{28}$$

It has mean μ and variance $2b^2$. Note that, for $\mu = 0$, this distribution is obtained by gluing together two exponential distributions. Hence, its name. Assume now that $\mu = 0$ and $b = 1$. Then one can verify that

$$\mathbb{P}[|X| > t] = 2\,\mathbb{P}[X > t] = e^{-t}. \tag{29}$$

If we compare (29) with (26) (for the case $\mu = 0$), we see that the tails of the double-sided exponential distribution are *heavier* than that of sub-Gaussian random variables. It turns out instructive to analyze the moment-generating function of the double-sided exponential distribution. It is given by

$$\mathbb{E}\left[e^{\lambda X}\right] = \frac{1}{1-\lambda^2}, \quad |\lambda| \leq 1 \tag{30}$$

and it is not defined for $\lambda > 1$. One can verify that (see Fig. 2)

$$\frac{1}{1-\lambda^2} \leq e^{2\lambda^2}, \quad \lambda < 1/2 \tag{31}$$

where the value $1/2$ is actually a conservative estimate of the range of values for which (31) holds. If we now compare (31) with (25) (for the case $\mu = 0$, we see that, in the neighborhood of the origin, the moment-generating function of the double-sided exponential distribution behaves like the moment-generating function of a sub-Gaussian random variable with parameter $\sigma^2 = 4$. It turns out that this property is shared by a large class of random variables whose tails are heavier than that of sub-Gaussian random variables but no heavier than that of a double-sided exponential random variable. This motivates the following definition:

Definition 5 (Sub-exponential random variable) *A random variable X with mean $\mu = \mathbb{E}[X]$ is sub-exponential if there exist nonnegative numbers ν and b such that*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\nu^2\lambda^2/2}, \quad \forall |\lambda| \leq 1/b. \tag{32}$$

In this case, we say that this random variable is (ν, b)-sub-exponential.

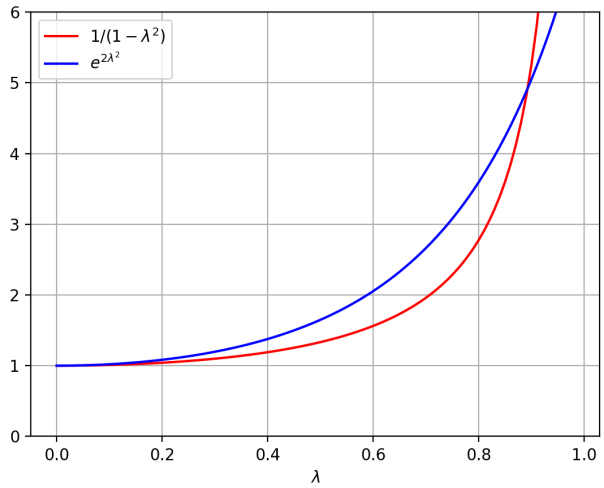

Figure 2: Comparison between the left-hand side and the right-hand side of (31).

As in the sub-Gaussian case, we will often choose the parameters (ν, b) in a convenient way, without trying to find the best (i.e., smallest) ones.

Note that the double-sided exponential distribution is sub-exponential with parameters $\nu = 2$ and $b = 2$. Furthermore, any sub-Gaussian random variable with parameter σ is also sub-exponential with parameters $(\sigma, 0)$. Here is another nontrivial example of sub-exponential random variable. Let $X \sim \mathcal{N}(0,1)$. Let $Z = X^2$. Note that $\mathbb{E}[Z] = 1$. One can show that

$$\mathbb{E}[e^{\lambda(Z-1)}] \leq e^{4\lambda^2/2}, \quad \forall |\lambda| \leq 1/4. \tag{33}$$

This means that Z is sub-exponential with parameters $\nu = 2$, $b = 4$. We now generalize the upper-deviation inequality (26) to the case of sub-exponential random variables.

Theorem 6 (Sub-exponential upper-deviation inequality) *Assume that X is sub-exponential with parameters (ν, b). Then*

$$\mathbb{P}[(X-\mu) \geq t] \leq \begin{cases} e^{-t^2/(2\nu^2)}, & \text{if } 0 \leq t \leq \nu^2/b \\ e^{-t/(2b)}, & \text{if } t \geq \nu^2/b. \end{cases} \tag{34}$$

Example: Consider the previous example of sub-exponential random variable. Specifically, let $X \sim \mathcal{N}(0,1)$ and let $Z = X^2$. Then Z is sub-exponential with parameters $(2, 4)$. It then follows from (34) that

$$\mathbb{P}[Z - 1 \geq t] \leq \begin{cases} e^{-t^2/8}, & \text{if } 0 \leq t \leq 1 \\ e^{-t/8}, & \text{if } t \geq 1. \end{cases} \tag{35}$$

More compactly,

$$\mathbb{P}[Z - 1 \geq t] \leq \exp\left(-\frac{1}{8}\min\{t, t^2\}\right). \tag{36}$$

3.5 The blessing of high dimensionality

3.5.1 Concentration of sums of independent random variables

Say that we compute the empirical average of n independent and identically distributed (iid) random variables. We know from the law of large numbers that the empirical average converges to the expectation as $n \to \infty$. The concentration of measure results we investigate in this section help us quantify how fast is this convergence.

The results we review next are an example of the so-called **blessing of dimensionality**. This expression refers to the fact that certain random fluctuations, which are complicated to model in the low-dimension regime, can be controlled accurately in high dimensions. The concentration inequalities we will establish in this section take the following form: let $X_1, \ldots, X_n$ be iid random variables with mean μ. Then

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{k=1}^{n} X_n - \mu\right| \geq t\right] \leq \text{something small}. \tag{37}$$

Three fundamental results

- Source: Vershynin, Chapter 2.1.

We start by reviewing three fundamental results in probability theory that, however, do not provide estimates as tight as the ones we shall establish later.

The first result is the well-known law of large numbers.

Theorem 7 (Strong law of large numbers) *Let $X_1, \ldots, X_n$ be iid random variables with mean μ. Let $S_n = X_1 + \cdots + X_n$. Then as $n \to \infty$, we have that S_n/n converges to μ almost surely.*

The next result, called the central limit theorem, identifies the limiting distribution of a properly scaled version of S_n.

Theorem 8 (Lindeberg-Lèvi central limit theorem) *Let $X_1, \ldots, X_n$ be iid random variables with mean μ and variance σ^2. Consider the normalized random variable $Z_n = \sum_{k=1}^{n}(X_k - \mu)/\sqrt{n\sigma^2}$. Then, as $n \to \infty$, the probability distribution of Z_n converges to that of a $\mathcal{N}(0,1)$ random variables.*

The final result we review here quantifies the rate at which convergence in the central-limit theorem occurs.

Theorem 9 (Berry-Esseen central limit theorem) *Let Z_n be as in Theorem 8. Let $Q(\cdot)$ denote the Gaussian Q function and let $c = \mathbb{E}[|X_1 - \mu|^3]/\sigma^3$. Then*

$$|\mathbb{P}[Z_n \geq t] - Q(t)| \leq \frac{6c}{\sqrt{n}}. \tag{38}$$

This last result implies that the rate of convergence is of order $1/\sqrt{n}$. Note that the constant that multiplies the $1/\sqrt{n}$ term can be improved.

Hoeffding's bound

- Source: Wainwright, Chapter 2.1.2 & Chapter 2.1.3.

As we shall see, we can obtain stronger concentration results than the ones just reviewed, if we are told that the $\{X_k\}$ are sub-Gaussian or sub-exponential.

We start by noting the following result. Let $X_1, \ldots, X_n$ be independent sub-Gaussian random variables with parameters $\sigma_1, \ldots, \sigma_n$, respectively. Then it is easy to verify that $S_n = X_1 + \cdots + X_n$ is sub-Gaussian with parameter $\sqrt{\sum_{k=1}^n \sigma_k^2}$. The following large-deviation inequality, which is known as Hoeffding's bound, then follows from (26).

Theorem 10 (Hoeffding's bound) *Let $X_1, \ldots, X_n$ be independent and assume that each X_k has mean μ_k and is sub-Gaussian with parameter σ_k. Then for all $t > 0$,*

$$\mathbb{P}\left[\sum_{k=1}^n (X_k - \mu_k) \geq t\right] \leq \exp\left(-\frac{t^2}{2\sum_{k=1}^n \sigma_k^2}\right). \tag{39}$$

Note that if the X_k are supported on the interval $[a, b]$, then each X_k is sub-Gaussian with parameter $(b-a)/2$ and (39) reduces to

$$\mathbb{P}\left[\sum_{k=1}^n (X_k - \mu_k) \geq t\right] \leq \exp\left(-\frac{2t^2}{n(b-a)^2}\right). \tag{40}$$

In particular, if we set $t = n\gamma$, and assume that $\mu_1 = \cdots = \mu_n = \mu$, the probability that the empirical average of these random variable deviates from the mean decays as $\exp(-2n\gamma^2/(b-a)^2)$.

We can establish an inequality similar to Hoeffding's inequality also for sub-exponential random variables. To do so, we need the following observation. Let $\{X_k - \mu_k\}$ be sub-exponential with parameters (ν_k, b_k), $k = 1, \ldots, n$. Then $\sum_{k=1}^n (X_k - \mu_k)$ is sub-exponential with parameters $\nu_\star = \sqrt{\sum_{k=1}^n \nu_k^2}$ and $b_\star = \max_k b_k$. This result, combined with (34) implies the following concentration of measure result.

Theorem 11 (Large-deviation inequality for sum of sub-exp. RVs) *Let the RVs X_k, $k = 1, \ldots, n$ be defined as above. Then*

$$\mathbb{P}\left[\sum_{k=1}^n (X_k - \mu_k) \geq t\right] \leq \begin{cases} \exp(-t^2/(2\nu_\star^2)) & \text{if } 0 \leq t \leq \frac{\nu_\star^2}{b_\star} \\ \exp(-t/(2b_\star)) & \text{if } t > \frac{\nu_\star^2}{b_\star}. \end{cases} \tag{41}$$

Improving Hoeffding's inequality: Bernstein's inequality

- Source: Wainwright, Chapter 2.1.3.

Let us start with a motivating example. Let X_k, $k = 1, \dots, n$ be iid and assume that they are drawn according to the following distribution:

$$X_k = \begin{cases} -1, & \text{with prob. } p/2 \\ 0, & \text{with prob. } 1-p \\ 1, & \text{with prob. } p/2 \end{cases} \tag{42}$$

Since the $\{X_k\}$ are supported on $\{-1, 0, 1\}$, they are sub-Gaussian with parameter 1. Also, they have zero mean. It then follows from Hoeffding's inequality that

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{k=1}^{n} X_k\right| \geq t\right] \leq 2\exp\left(-\frac{nt^2}{2}\right). \tag{43}$$

However, this bound does not depend on p. Intuitively, the smaller p is the more unlikely large deviations from the mean 0 should be. We next seek a bound that hold for sequences $X_1, \dots, X_n$ with more general distribution than the one just considered. We want that the resulting bound when applied to the setup described above, improves on Hoeffding by capturing the dependence on p. We shall assume for simplicity that all the $\{X_k\}$ have the same mean μ and the same variance σ^2. We will also assume that these random variables satisfy the following condition, which we will refer to as Bernstein's condition.

Definition 12 (Bernstein condition) *Let X be a RV with mean μ and variance σ^2. We say that X satisfies the Bernstein condition with parameter b if*

$$|\mathbb{E}[(X-\mu)^k]| \leq \frac{1}{2}k!\sigma^2 b^{k-2}, \quad k = 2, 3, \dots \tag{44}$$

Bernstein condition is satisfied for bounded random variables (indeed, if $|X| \leq c$ with probability one, then $b = c$), but it also hold for some unbounded random variables. Bernstein condition can be used to tighten (in some cases) Hoeffding's inequality. Specifically, the following result holds.

Theorem 13 (Bernstein-type bound) *For any random variable X satisfying the Bernstein condition (44), we have that*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq \exp\left(\frac{\lambda^2\sigma^2/2}{1-b|\lambda|}\right), \quad \forall |\lambda| \leq 1/b. \tag{45}$$

Furthermore, the following tail bound holds:

$$\mathbb{P}[|X-\mu| \geq t] \leq 2\exp\left(-\frac{t^2}{2(\sigma^2+bt)}\right), \quad \forall t \geq 0. \tag{46}$$

Note that the Bernstein-type bound in (46) depends both on the variance σ^2 of the random variable, and on the Berstein parameter, which for a bounded random variable is related to the size of its support.

These two inequalities imply the following concentration of measure result for sum of of i.i.d. random variables satisfying (44)

$$\mathbb{P}\left[\left|\sum_{k=1}^{n} X_k - \mu\right| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2(n\sigma^2 + bt)}\right) \quad \forall t \geq 0. \tag{47}$$

Let us now go back to the example we started with. The random variables in (42) are bounded on $[-1, 1]$. Hence, they satisfy Bernstein condition with $b = 1$. Furthermore, $\sigma^2 = p$. It then follows from (47) that

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{k=1}^{n} X_k\right| \geq t\right] \leq 2\exp\left(-\frac{nt^2}{2(p+t)}\right). \tag{48}$$

Note that when $t \ll p$ this bound is much tighter than (43) and it essentially reveals that the random variables behave as if they were sub-Gaussian with parameter p instead of sub-Gaussian with parameter 1.

3.5.2 The geometry of the cube (revisited)

- Source: Bandeira, Chapter 2.4.2

We can use the concentration of measure results just established to prove that almost all the volume of the cube in high dimensions is located in its corners. The proof will be based on a probabilistic argument: it illustrates the connection between geometry and probability in high dimensions.

Let $C^d(1/2) = [-1/2, 1/2]^d$ the cube of length 1 in d dimensions. Let $B^d(1/2)$ be the ball with radius $1/2$ in d dimensions. We will show that $\mathrm{Vol}(C^d(1/2)/B^d(1/2))$ is large. Let $X = (X_1, \ldots, X_n)$ be drawn uniformly at random within the cube. This means that each component X_k, $k = 1, \ldots, d$. is uniformly distributed on $[-1/2, 1/2]$. We will show that the probability that $X \in B^d(1/2)$, i.e., the probability that $\sum_{k=1}^{d} X_k^2 \leq 1/4$, is small. Let $Z_d = 4X_d^2$. Note that $\mathbb{E}[Z_k] = 1/3$. Furthermore, the $\{Z_k\}$ are sub-Gaussian with parameter $1/2$, since they are supported on $[0, 1]$.

An application of Hoeffding's inequality reveals that, when $d \geq 3$,

$$\mathbb{P}\left[\sum_{k=1}^{d} X_k^2 \leq \frac{1}{4}\right] = \mathbb{P}\left[\sum_{k=1}^{d} Z_k - \frac{d}{3} \leq 1 - \frac{d}{3}\right] \leq \exp\left(-\frac{2(1-d/3)^2}{d}\right) \tag{49}$$

Note that for $d \gg 1$, the right-hand side of this inequality is approximately equal to $\exp(-2d/9)$. This means that the probability that a point drawn from a uniform distribution on the cube is also in the sphere decays exponentially with d. So most points must be outside the sphere, i.e., in the corners.

Most points on the surface of a d-dimensional sphere are close to the equator

- Source: Bandeira, Chapter 2.4.3

We next establish this result using a similar method. Let us consider a random point uniformly drawn from the surface of the sphere. To generate this point, we normalize a Gaussian-distributed d-dimensional random vector. Specifically, let $\tilde{X} \sim \mathcal{N}(0, I_d)$. Then it follows by circular symmetry of the Gaussian distribution that $X = \tilde{X}/\|\tilde{X}\|$ is uniformly distributed on $S^{d-1}(1)$.

Note that, by construction $\|X\|^2 = 1$. Furthermore, $\mathbb{E}[\|X\|^2] = \sum_{k=1}^{d} \mathbb{E}[X_k^2] = 1$. By symmetry, we must have $\mathbb{E}[X_k^2] = 1/d$ for all $k = 1, \dots, d$.

To demonstrate that most of the surface is close to the equator, we just need to show that, for example, $\mathbb{P}[|X_1| > \epsilon]$. It follows from Chebyshev's inequality that

$$\mathbb{P}[|X_1| > \epsilon] \leq \frac{\mathbb{E}[|X_1|^2]}{\epsilon^2} = \frac{1}{d\epsilon^2}. \tag{50}$$

3.5.3 Random vectors in high dimensions

- Source: Bandeira, Chapter 2.4.4.

In this section, we will investigate two questions:

- What length do we expect a random vector in $\mathbb{R}^n$ to have?
- What is the angle between two random vectors?

Length of a random vector: Assume that the n-dimensional random vector $X = [X_1, \dots, X_n]$ has iid entries with zero mean and unit variance. Since $\mathbb{E}[\|X\|^2] = \sum_{k=1}^{n} \mathbb{E}[X_k^2] = n$, we expect the typical length of X to be $\sqrt{n}$. We will use the concentration of measure results just developed to make this statement more precise.

For simplicity we shall focus on the case in which $X \sim \mathcal{N}(0, I_d)$. Then $\|X\|^2 \sim \chi^2(n)$. As discussed in Section 3.4, each X_k^2 is sub-exponential with parameters $(2, 4)$. It then follows from (41) that

$$\mathbb{P}\left[\left|\frac{1}{n}\|X\|^2 - 1\right| \geq t\right] \leq 2\exp\left(-\frac{n}{8}\min\{t, t^2\}\right). \tag{51}$$

This gives a concentration result on $\|X\|^2$. But what can we say about $\|X\|$? To establish a concentration result on $\|X\|$, we use that for all $z \geq 0$, if $|z - 1| \geq \delta$ then $|z^2 - 1| \geq \max\{\delta, \delta^2\}$. Using this result, we conclude that

$$\mathbb{P}\left[\left|\frac{1}{\sqrt{n}}\|X\| - 1\right| \geq \delta\right] \leq \mathbb{P}\left[\left|\frac{1}{n}\|X\|^2 - 1\right| \geq \max\{\delta, \delta^2\}\right] \leq 2\exp\left(-\frac{n}{8}\delta^2\right). \tag{52}$$

So we see that, in both cases, the probability that the vector has a length that is δ away from the expected length $\sqrt{n}$ vanishes exponentially in n.

Angle between two random vectors

Theorem 14 (Angle between Rademacher vectors) *Assume that X and Y are d-dimensional random vectors[1] that are independent and have iid ± 1 entries with equal probability (Rademacher entries). Define the cosine of the angle $\theta(X,Y)$ between the two vectors as follows:*

$$\cos\theta(X,Y) = \frac{Y^{\mathrm{T}}X}{\|X\|\|Y\|} = \frac{Y^{\mathrm{T}}X}{d}. \tag{53}$$

Then

$$\mathbb{P}\left[|\cos\theta(X,Y)| \geq \sqrt{\frac{2\log d}{d}}\right] \leq \frac{2}{d}. \tag{54}$$

This result follows directly from Hoeffding inequality (40). Indeed, note that $Y^{\mathrm{T}}X = \sum_{k=1}^{d} X_k Y_k$ is the sum of i.i.d. Rademacher RVs. Hence,

$$\mathbb{P}[|Y^{\mathrm{T}}X| \geq t] = \mathbb{P}\left[\left|\sum_{k=1}^{d} X_k Y_k\right| \geq t\right] = \mathbb{P}\left[\frac{\left|\sum_{k=1}^{d} X_k Y_k\right|}{\|X\|\|Y\|} \geq \frac{t}{d}\right] \leq 2\exp\left(-\frac{t^2}{2d}\right). \tag{55}$$

Then set $t = \sqrt{2d\log d}$.

Theorem 14 implies that, as d grows large, it is more and more likely that two randomly generated vectors with Rademacher entries are orthogonal. Indeed as $d \to \infty$, we have that $\sqrt{(2\log d)/d} \to 0$.

A similar result holds also for Gaussian random vectors and for random vectors chosen uniformly from the sphere S^{d-1}. Note also that while we can have only d vectors that are orthogonal in $\mathbb{R}^d$, for large d we can have exponentially many vectors that are *almost* orthogonal. This is explored in exercise 13.

3.6 Exercises

Exercise 1 (Volume of $B^d(R)$) *Prove the formula for the volume of $B^d(R)$ provided in (4). Plot $B^d(1)$ as a function of d. For which value of d is $B^d(1)$ maximized?*

Exercise 2 (p-moments via tails) *Let X be a random variable and $p \in (0,\infty)$. Show that*

$$\mathbb{E}[|X|^p] = \int_0^\infty pt^{p-1}\mathbb{P}[|X| > t]\mathrm{d}t. \tag{56}$$

Exercise 3 (Chebyshev from Markov) *Deduce Chebyshev's inequality (19) from Markov's inequality (18).*

[1] Throughout these notes, we will use the convention that vectors are always column vectors.

Exercise 4 (Symmetry and sub-Gaussian random variables) *Prove that if X is sub-Gaussian with parameter σ, then $-X$ is also sub-Gaussian with same parameter.*

Exercise 5 (Double-sided bound for sub-Gaussian random variables) *Let X be sub-Gaussian with parameter σ. Use the union bound and (26) to show that*

$$\mathbb{P}[|X-\mu| \geq t] \leq 2e^{-t^2/(2\sigma^2)}, \quad \forall t \geq 0 \tag{57}$$

Exercise 6 (Bounded RVs are sub-Gaussian) *Prove that bounded RVs are sub-Gaussian. Specifically prove that if $X \in [-a,a]$ with probability 1, and $\mathbb{E}[X]=0$, then X is sub-Gaussian with parameter a. Here is a hint on how to proceed:*

- *Use convexity to show that*

$$e^{\lambda x} \leq \frac{a+x}{2a}e^{\lambda a} + \frac{a-x}{2a}e^{-\lambda a}, \quad \forall x \in [-a,a] \tag{58}$$

- *Apply this inequality to prove that $\mathbb{E}[e^{\lambda X}] \leq \cosh(\lambda a)$*
- *Conclude the proof by arguing that $\cosh(\lambda a) \leq e^{\lambda^2 a^2/2}$.*

Note that a more general result holds: if X is supported on $[a,b]$, then X is sub-Gaussian with parameter $\sigma = (b-a)/2$ (independently of its mean). However, the proof is more involved.

Exercise 7 (Moments of sub-Gaussian random variables) *Assume that X has zero mean and it is sub-Gaussian with parameter σ. Then $\mathbb{E}[|X|^p] \leq 2^{p/2}p\sigma^p\Gamma(p/2)$. Hint: use (56).*

Exercise 8 (Sub-exponential from sub-Gaussian) *Let X be a zero-mean sub-Gaussian random variable. Prove that $X^2 - \mathbb{E}[X^2]$ is sub-exponential. Hint: Take the Taylor expansion of $\mathbb{E}\left[e^{\lambda\left(X^2-\mathbb{E}[X^2]\right)}\right]$. Use also that, for every non-negative random variable Z, we have that $\mathbb{E}[(Z-\mathbb{E}[Z])^p] \leq \mathbb{E}[Z^p]$, $p \in \mathbb{N}$ as well as the result from the previous exercise.*

Exercise 9 (Tail probability of a Gaussian random variable) *Let $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ and $\Phi(u) = \int_{-\infty}^{u}\phi(x)dx$ be the density and distribution function, respectively, of the standard normal distribution. Show, e.g. using partial integration, that*

$$1-\Phi(u) \sim \frac{\phi(u)}{u}$$

i.e. that the probability that that a standard normal variable is larger than u asymptotically is $\frac{\phi(u)}{u}$. Compare with the subgaussian bound.

Exercise 10 (Fair coins) *Toss a fair coin n times: what is the probability that we get at least 75% heads? Obtain the exact expression (which needs to be evaluated numerically) and compare it with the estimates obtained using Chebyshev's inequality, the central-limit theorem, and Hoeffding's inequality.*

Exercise 11 (Improving the performance of a randomized algorithm) *Assume you are given a randomized algorithm for solving some decision problems. The algorithm returns the correct answer with probability $1/2+\delta$ for some $\delta > 0$. To improve the performance, you decide to run the algorithm n times and take a majority vote. How large should n be so that the answer is correct with probability at least $1-\epsilon$?*

Exercise 12 (Concentration of measure for chi-squared random variables) *Let Y be χ^2 distributed with n degrees of freedom, i.e., $Y = \sum_{k=1}^{n} Z_k^2$ where the Z_k are independent and $\mathcal{N}(0,1)$-distributed. Provide an estimate for $\mathbb{P}[\frac{1}{n}\sum_{k=1}^{n}(Z_k^2 - 1) \geq t]$*

Exercise 13 (Angle between binary vectors) *Let $X_1, \ldots, X_n$ be independent d-dimensional vectors with Rademacher entries. How large can we choose n if we want that with probability at least $7/8$ the cosine of the angle between any of two vectors is at most $1/100$?*

4 Large random matrices

Random matrices occur naturally in many problems in high-dimensional statistics, such as spectral clustering, principal component analysis, and covariance-matrix estimation. In this chapter, we start an investigation of the non-asymptotic behavior of random matrices.

4.1 Preliminaries on matrices

- Source: Vershynin, Chapter 4.1

4.1.1 Singular-value decomposition

We shall focus on a $m \times n$ dimensional matrix A with real entries. Such a matrix can be represented using the **singular value decomposition**. Specifically, let r be the rank of A. Then

$$A = \sum_{i=1}^{r} s_i u_i v_i^{\mathrm{T}}. \tag{59}$$

Here, $s_1 \geq s_2 \geq \cdots \geq s_r \geq 0$ are non-negative number called **singular values**; $u_1, \ldots, u_r$ are m-dimensional vectors referred to as the **left singular vectors**. These are the orthonormal eigienvectors of the matrix AA^{T}; finally, $v_1, \ldots, v_r$ are n-dimensional vectors referred to as the **right singular vectors**. These are the orthonormal eigienvectors of the matrix $A^{\mathrm{T}}A$. Note that the singular values are related to the eigenvalues of AA^{T} and of $A^{\mathrm{T}}A$, which we denote by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$ (note that the matrices AA^{T} and $A^{\mathrm{T}}A$ are positive semidefinite, so all eigenvalues are nonnegative). Specifically,

$$s_k = \sqrt{\lambda_k}, \quad k = 1, \ldots, r. \tag{60}$$

In general, if A is a symmetric matrix (i.e., $A = A^{\mathrm{T}}$), then the singular values are equal to the absolute value of the eigenvalues of the matrix (note that the eigenvalues may be negative, whereas the singular values are always non-negative).

4.1.2 Norm of matrices

The space of $m \times n$ matrices can be equipped with several different norms. We will mainly consider the **operator norm**, which is defined as follows:

$$\|A\| = \max_{x \in S^{n-1}} \|Ax\|. \tag{61}$$

Equivalently, the operator norm can be computed by maximizing the quadratic form $y^{\mathrm{T}}Ax$. Specifically,

$$\|A\| = \max_{x \in S^{n-1}, y \in S^{m-1}} y^{\mathrm{T}}Ax. \tag{62}$$

Recall now that if B is a $n\times n$ symmetric matrix, the maximum and minimum eigenvalues admit the following variational characterization (this result is known as the **Rayleigh-Ritz Theorem**)

$$\lambda_1 = \max_{x\in S^{n-1}} x^T Bx \tag{63}$$

$$\lambda_n = \min_{x\in S^{n-1}} x^T Bx \tag{64}$$

This variational characterization implies that

$$\|A\| = \sqrt{\lambda_1(A^T A)} = s_1(A). \tag{65}$$

Furthermore, it also implies that, if A is a symmetric $n \times n$ matrix, then

$$\|A\| = \max_{x\in S^{n-1}} |x^{\mathrm{T}} Ax|. \tag{66}$$

These results will turn out useful in various parts of the course.

4.2 The operator norm of sub-Gaussian random matrices

We are now ready to obtain our first result on random matrices. Specifically, we shall show that if a $m \times n$ matrix A has independent sub-Gaussian entries, then $\|A\|$ is no larger than $\mathrm{c}(\sqrt{m} + \sqrt{n})$, where c denotes a constant that does not depend on m or n, with high probability.

It follows from (62) that to establish such a result, we need to show that the probability that $y^{\mathrm{T}} Ax$ is larger than $\mathrm{c}(\sqrt{m}+\sqrt{n})$ for some $x \in S^{n-1}$, $y \in S^{m-1}$ is small. The problem is that to do so, we will need to check uncountably many x and y. Instead of doing so, we employ the so-called ϵ-covering argument. The idea is to discretize the spheres S^{n-1} and S^{m-1} using only a finite number of suitably chosen points (the ϵ-covering), control the error incurred by this discretization step, and then apply concentration of measure results combined with the union bound *only* over the chosen points.

4.2.1 Covering and packing

- Wainwright, Chapter 5.1, Varshynin, Chapter 4.2.1

We start by formally introducing the concept of ϵ-cover. Then we will discuss how to bound the cardinality of the minimal ϵ-covers for the unit-radius hyperspheres S^{n-1} and S^{m-1}.

Definition 15 (Covering number) *An ϵ-cover of a set $\mathcal{T} \subset \mathbb{R}^n$ is a set $\{t_1, \dots, t_N\} \subset \mathcal{T}$ for which, for all $t \in \mathcal{T}$, there exists a $i \in \{1, \dots, N\}$ such that $\|t_i - t\| \leq \epsilon$. The ϵ-covering number $N(\epsilon, \mathcal{T})$ is the cardinality of the smallest ϵ-cover.*

A related concept is the one of ϵ-packing.

Definition 16 (Packing number) *An ϵ-packing of a set $\mathcal{T} \subset \mathbb{R}^n$ is a set $\{t_1, \ldots, t_P\} \subset \mathcal{T}$ for which $\|t_i - t_j\| > \epsilon$ for all $i, j \in \{1, \ldots, P\}$. The ϵ-packing number $P(\epsilon, \mathcal{T})$ is the cardinality of the largest ϵ-packing.*

An ϵ-packing for $\mathcal{T}$ can be visualized as collection of balls of of radius $\epsilon/2$ and center in $\mathcal{T}$ such that no two balls intersect. The following relation holds:

$$P(2\epsilon, \mathcal{T}) \leq N(\epsilon, \mathcal{T}) \leq P(\epsilon, \mathcal{T}). \tag{67}$$

In the next theorem, we bound the ϵ-covering number of the unit ball B^n, from which a useful upper bound on the ϵ-covering number of the unit sphere S^{n-1} follows.

Theorem 17 (Covering number of the Euclidean ball) *The ϵ-covering number $N(\epsilon, B^n)$ of the Euclidean ball B^n satisfies*

$$(1/\epsilon)^n \leq N(\epsilon, B^n) \leq (2/\epsilon + 1)^n. \tag{68}$$

The upper bound is true also for the Euclidean sphere S^{n-1}.

4.2.2 Computing the operator norm on an ϵ-cover

As already mentioned, to evaluate $\|A\|$ we need to maximize $y^{\mathrm{T}} A x$ over $x \in S^{n-1}$ and $y \in S^{m-1}$. In the next theorem, we control the error resulting when the maximization is performed on ϵ covers of S^{n-1} and S^{m-1}.

Theorem 18 (Operator norm on a cover) *Let A be an $m \times n$ matrix and let $\epsilon \in [0, 1/2)$. For every ϵ-cover $\mathcal{N}$ of the sphere S^{n-1} and for every ϵ-cover $\mathcal{M}$ of the sphere S^{m-1}, we have*

$$\sup_{x \in \mathcal{N}, y \in \mathcal{M}} y^{\mathrm{T}} A x \leq \|A\| \leq \frac{1}{1 - 2\epsilon} \sup_{x \in \mathcal{N}, y \in \mathcal{M}} y^{\mathrm{T}} A x. \tag{69}$$

So this theorem tells us that if we replace the maximization over all $x \in S^{n-1}$ and $y \in S^{m-1}$, with a maximization over two ϵ nets of the two spheres, we pay a multiplicative penalty equal to $1/(1 - 2\epsilon)$.

4.2.3 The norm of a sub-Gaussian random matrix

- Source: Vershynin, Chapter 4.4.2 (with some modifications to make the constant explicit)

We are now ready to state our first result on random matrices.

Theorem 19 (Norm of matrices with sub-Gaussian entries) *Let A be an $m \times n$ random matrix whose entries A_{ij} are independent, zero mean sub-Gaussian random variables with parameter σ_{ij}. Let $\sigma = \max_{ij} \sigma_{ij}$. Then for all $t \geq 0$ and for all $\epsilon \in (0, 1/2)$ we have that*

$$\mathbb{P}\left[\|A\| \geq \frac{\sqrt{2\sigma^2}}{1 - 2\epsilon}\left(t + (\sqrt{m} + \sqrt{n})\sqrt{\log\left(\frac{2}{\epsilon} + 1\right)}\right)\right] \leq e^{-t^2}. \tag{70}$$

To be concrete, let us assume $\sigma = 1$, $m = n$, and $t = \sqrt{n}$. Then by optimizing the bound over ϵ, we conclude that

$$\mathbb{P}\big[\|A\| \geq 7.63\sqrt{n}\big] \leq e^{-n}. \tag{71}$$

Note that $e^{-10} \approx 4.5 \cdot 10^{-5}$. So this bound provides a very sharp concentration of measure result already for matrices as small as 10×10. This theorem can be extended to symmetric matrices, which will play an important role in the reminder of the course.

Theorem 20 (Norm of symmetric matrices with sub-Gaussian entries) *Let A be an $n \times n$ symmetric random matrix whose entries A_{ij} on and above the main diagonal are independent, zero-mean, sub-Gaussian random variables with parameter σ_{ij}. Let $\sigma = \max_{ij} \sigma_{ij}$. Then for all $t \geq 0$ and for all $\epsilon \in (0, 1/2)$ we have that*

$$\mathbb{P}\left[\|A\| \geq 2\frac{\sqrt{2\sigma^2}}{1-2\epsilon}\left(t + 2\sqrt{n}\sqrt{\log\left(\frac{2}{\epsilon}+1\right)}\right)\right] \leq 2e^{-t^2}. \tag{72}$$

The idea of the proof is to split A in an upper-triangular part and a lower-triangular part (the diagonal should be part of only one of the two matrices), use Theorem 19 on each part (show that this theorem generalizes to this setting), and then apply triangular inequality and the union bound.

Consider again the case $\sigma = 1$, $m = n$ and $t = \sqrt{n}$. Then, after optimizing over ϵ the bound in (72) becomes

$$\mathbb{P}\big[\|A\| \geq 15.3\sqrt{n}\big] \leq 2e^{-n} \tag{73}$$

In general, for an arbitrary value of σ, we can conclude that, if A is sufficiently large, then with high probability, $\|A\|$ is no larger than $\mathrm{c}\sqrt{n}$, where c is a constant that depends on σ.

4.3 Application: community detection in networks

- Source: Vershyinin, Chapter 4.5

Results on random matrices are useful in many applications. Here is one such example. Real-world networks are often organized in "communities", i.e., clusters of vertexes that are tightly connected. The problem of community detection deals with finding communities within a network.

Consider the problem depicted in Fig. 3. This is a real data set of political blogs ($n = 1222$) from the 2004 US political elections. Specifically, we have access to the following information: which blog refers to which other blog through hyperlinks. Our task is to decide the political inclination of the blogs (democratic vs republican).

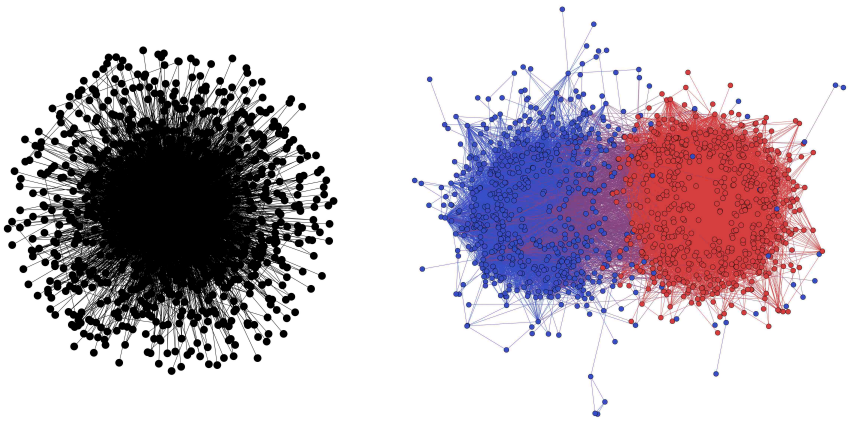

Figure 3: Real data set of political blogs: each vertex represents a blog and each edge represents a hyperlink to another blog. The left graph depics the available data, the right graph the output of a modern community detection algorithm. This figure is taken from [5].

4.3.1 The stochastic block model

We will consider a simple community detection problem: we have a network with n vertexes (n even), which are divided in two communities of $n/2$ vertexes each. We construct a random graph consisting of these vertexes as follows. Two vertexes are connected with probability p if they belong to the same community; otherwise they are connected with probability q. Note that we allow self loops. This model is an example of the so-called stochastic block model. We assume that $p > q$ so that edges are more likely to occur within communities than across communities.

For a given random graph, the community detection problem involves deciding which vertexes belong to the first community and which vertexes belong to the second community. We describe one such random graph using the **adjacency matrix** A, a binary matrix, whose entry in position (i, j) is equal to 1 if there exists an edge between vertex i and vertex j, and it is equal to zero otherwise. It follows from our construction that A is symmetric and that the entries of A are Bernoulli distributed. Specifically, they are either Bern(p) or Bern(q), depending whether the corresponding vertexes belong to the same community or not.

Let $D = \mathbb{E}[A]$, where the expectation is taken entry-wise and let $E = A - D$, so that $A = D + E$. It turns out that if D was known to us (it is not, unfortunately), then we could solve the community detection problem optimally with a simple algorithm. To see why, let us perform a spectral decomposition of D. Let us assume for simplicity that the vertexes $1, \dots, n/2$ belong to

community 1 and the remaining vertexes belong to community 2. Then the matrix D has the following block structure

$$\left(\begin{array}{ccc|ccc} p & \dots & p & q & \dots & q \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ p & \dots & p & q & \dots & q \\ \hline q & \dots & q & p & \dots & p \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ q & \dots & q & p & \dots & p \end{array}\right) \tag{74}$$

Note that the matrix D has rank 2. Furthermore the nonzero eigenvalues and the corresponding eigenvectors are

$$\lambda_1 = n\frac{p+q}{2}, \quad u_1 = [\underbrace{1,\dots,1}_{n/2},\underbrace{1,\dots,1}_{n/2}]^{\mathrm{T}}, \tag{75}$$

and

$$\lambda_2 = n\frac{p-q}{2}, \quad u_2 = [\underbrace{1,\dots,1}_{n/2},\underbrace{-1,\dots,-1}_{n/2}]^{\mathrm{T}}. \tag{76}$$

What is important to note here is the form of the second eigenvector u_2. We can tell which community each vertex belongs to by inspecting the sign of the corresponding entry in the eigenvector. If the entry is positive, then the corresponding vertex belongs to community 1. If it is negative, it corresponds to community 2.

Unfortunately, we do not know $D = \mathbb{E}[A]$: we are just given a single realization of the random matrix A. The question we shall investigate then is the following. Is A sufficiently close to D for large n so that we can perform community detection by using the second eigenvector of A (which we know) rather than that of D (which we do not know)?

4.3.2 Perturbation theory for matrices

To answer this question, we need to understand how the eigenvectors of a matrix change under matrix perturbations. Indeed, $A = D + E$, where we can think of E as a sort of perturbation matrix. Before doing that, let us start with the eigenvalues.

Theorem 21 (Weyl's theorem) *For every two symmetric matrices S and T with the same dimension, we have*

$$\max_k |\lambda_k(S) - \lambda_k(T)| \leq \|S - T\|. \tag{77}$$

So the operator norm controls the stability of the eigenvalues.

A similar result holds for the eigenvectors, although for the result to hold, we need to assume that the eigenvalues are well separated.

Theorem 22 (Davis-Kahan theorem) *Let S and R be symmetric matrices with the same dimensions. Fix k and assume that the kth largest eigenvalue of S is well separated from the rest of the eigenvalues:*

$$\min_{j\neq k} |\lambda_k - \lambda_j| = \delta > 0. \tag{78}$$

Assume that $v_k(S)$ and $v_k(R)$ are the unit eigenvectors associated to the kth eigenvalue of S and R, respectively. Let ϕ be the angle between these two eigenvectors, measured so that this angle is nonnegative. Then

$$\sin\phi \leq \frac{2\|S-R\|}{\delta}.$$

Since we may reverse the sign of $v_k(R)$ to ensure that $\big(v_k(R)\big)^{\mathrm{T}} v_k(S) \geq 0$, we conclude that there exists a $\theta \in \{-1,1\}$ such that

$$\|v_k(S) - \theta v_k(R)\| \leq \frac{2^{3/2}\|S-R\|}{\delta}. \tag{79}$$

4.3.3 Spectral clustering

Let us now come back to the community detection problem. Let us set $S = D$ and $T = A = D+E$. We next compute the δ parameter for the second eigenvalue of D:

$$\delta = \min_{j\in\{1,3\}} |\lambda_2(D) - \lambda_j(D)| = n \underbrace{\min\left\{\frac{p-q}{2}, q\right\}}_{\mu} = n\mu. \tag{80}$$

Let $u_2(D)$ be the eigenvector of D corresponding to the second eigenvalue, normalized so that its norm is $\sqrt{n}$ (and its entries belong to $\{+1,-1\}$ as shown above). Let $u_2(A)$ be the eigenvector of A corresponding to the second eigenvalue, and normalized so that its norm is $\sqrt{n}$. It then follows from Theorem 22 that there exists a $\theta \in \{-1,1\}$ such that

$$\frac{1}{\sqrt{n}}\|u_2(D) - \theta u_2(A)\| \leq \frac{2^{3/2}\|E\|}{n\mu}. \tag{81}$$

Note now that E is sub-Gaussian, because its entries are bounded. Furthermore, since E is symmetric, it follows from (73) that

$$\|E\| \leq \mathrm{c}\sqrt{n} \tag{82}$$

with probability at least $1-2e^{-n}$. Hence, with the same probability,

$$\|u_2(D) - \theta u_2(A)\| \leq \frac{\mathrm{c}}{\mu}. \tag{83}$$

Here, we absorbed the factor $2^{3/2}$ into the constant c.[2] Let us now analyze the term

$$\|u_2(D) - \theta u_2(A)\|^2 = \sum_{k=1}^{n} |[u_2(D)]_k - [\theta u_2(A)]_k|^2. \tag{84}$$

Since the entries of the vector $u_2(D)$ are in $\{+1, -1\}$ it follows that each index in which the two vectors $u_2(D)$ and $\theta u_2(A)$ have entries with opposite signs, contributes to the sum by at least 1. This implies that the number of disagreeing signs must be bounded by c/μ^2. Since this constant does not depend on n, this result implies that as n grows large, only a vanishing fraction of the vertexes in the graph will be mis-classified. In other words, we can use the vector $u_2(A)$ to accurately estimate the vector $u_2(D)$, whose signs identify the two communities.

To summarize consider the following **spectral clustering algorithm**.

- **Input**: graph G with n vertexes and edges drawn according to the stochastic block model
- **Output**: a partition of the vertexes of G into two communities

1. Compute the adjacency matrix A of the graph G
2. Compute the eigenvector $u_2(A)$ corresponding to the second largest eigenvalue of A
3. Partition the vertexes into two communities based on the signs of the entries of $u_2(A)$: If $[u_2(A)]_k > 0$, $k = 1, \dots, n$, put vertex k into the first community, otherwise in the second.

Then we have just shown that with probability at least $1 - 2e^{-n}$, the spectral clustering algorithm identifies the communities of G correctly up to c/μ^2 misclassified vertexes.

Recall now that $\mu = \min\left\{\frac{p-q}{2}, q\right\}$. To have a small number of misclassified vertexes we need μ to be large. This means that the algorithm works well when the graph is sufficiently dense, i.e., q is sufficiently large, and the probability of edges within communities and across communities are sufficiently different, i.e., $p - q$ is sufficiently large.

4.4 Two-sided bounds on the operator norm

- Source: Verhsynin, Chapter 4.6

We shall now generalize the result obtained in Theorem 19 in two ways:

[2] Throughout the remainder of these lecture notes, we will use c to denote a constant. Its value may change at each occurrence.

- We will provide a sharper, two-sided bound.
- Rather than assuming that all entries of the matrix are independent, we will just assume that the rows are independent, but there may be dependence between the entries of the matrix on the same row. This generalization is important in data-science applications, since the rows are often obtained by sampling from some high-dimensional distribution.

To get started, we first extend the notion of sub-Gaussianity to random vectors.

Definition 23 (Sub-Gaussian random vectors) *A zero-mean n-dimensional vector X is sub-Gaussian with parameter σ if for all $v \in S^{n-1}$, the random variable $v^{\mathrm{T}}X$ is sub-Gaussian with parameter σ.*

We say that a random matrix A is **row-wise sub-Gaussian with parameter** σ if its rows are independent sub-Gaussian vectors with parameter σ. We shall also need this auxiliary result.

Theorem 24 (Approximate isometry) *Let A be an $m\times n$ matrix and $\delta > 0$. Suppose that*

$$\|A^T A - I_n\| \leq \max\{\delta, \delta^2\}. \tag{85}$$

Then for all $x \in \mathbb{R}^n$,

$$(1-\delta)\|x\| \leq \|Ax\| \leq (1+\delta)\|x\|. \tag{86}$$

As a consequence, all singular values of A are between $(1-\delta)$ and $(1+\delta)$:

$$1 - \delta \leq s_n(A) \leq s_1(A) \leq 1 + \delta. \tag{87}$$

Here, "approximate isometry" means that the norm of x does not change much when x is multiplied by A. Our main result is in the next theorem.

Theorem 25 (Two-sided bound on sub-Gaussian matrices) *Let A be an $m \times n$ row-wise sub-Gaussian matrix with parameter σ. Assume also that each row A_k of A satisfies $\mathbb{E}[A_k A_k^T] = I_n$ (i.e., the entries in each row are uncorrelated, but not necessarily independent). Let $\epsilon \in (0, 1/2)$, set $r = 16\sigma^2$, and $\delta = \sqrt{\frac{2}{m}}t + \sqrt{\frac{2n}{m}\log\left(\frac{2}{\epsilon}+1\right)}$ Then*

$$\mathbb{P}\left[\left\|\frac{1}{m}A^T A - I_n\right\| \geq \frac{r}{1-2\epsilon}\max\{\delta, \delta^2\}\right] \leq 2e^{-t^2}. \tag{88}$$

If we now combine this result with Theorem 24 and use our absolute constant notation, we conclude that, under the assumption that $m \geq n$,

$$\mathbb{P}\left[1 - \mathrm{c}\left(\frac{t}{\sqrt{m}} + \sqrt{\frac{n}{m}}\right) \leq \frac{s_n(A)}{\sqrt{m}} \leq \frac{s_1(A)}{\sqrt{m}} \leq 1 + \mathrm{c}\left(\frac{t}{\sqrt{m}} + \sqrt{\frac{n}{m}}\right)\right] \leq 2e^{-t^2}. \tag{89}$$

In particular, if we set $t = \sqrt{m}\gamma$ for some $\gamma > 0$,

$$\mathbb{P}\left[1 - \mathrm{c}\left(\gamma + \sqrt{\frac{n}{m}}\right) \leq \frac{s_n(A)}{\sqrt{m}} \leq \frac{s_1(A)}{\sqrt{m}} \leq 1 + \mathrm{c}\left(\gamma + \sqrt{\frac{n}{m}}\right)\right] \leq 2e^{-m\gamma^2}. \quad (90)$$

This result means that when $m \gg n$, then the matrix $A/\sqrt{m}$ is an approximate isometry in the sense of Theorem 24.

What is surprising perhaps about this theorem is the correction factor $\sqrt{m/n}$. Indeed the theorem implies that when both m and n are large, so that their ratio $\zeta = m/n$ is approximately constant, then the singular values belong to the interval $[1 - \mathrm{c}\sqrt{\zeta}, 1 + \mathrm{c}\sqrt{\zeta}]$. In particular, the value of the ratio ζ determines the size $2\mathrm{c}\sqrt{\zeta}$ of the interval. Our intuition, instead, may let us believe that the singular values converges all to 1 since the covariance matrix of each row vector is the identity matrix.

Note that Theorem 25 was proven under the simplifying assumption that the covariance matrix of each row was an identity. As shown in the next section, this assumption can be easily relaxed.

4.5 Application: Covariance matrix estimation

- Source: Vershyinin, Chapter 4.7

Covariance matrices play a fundamental role in statistics. We will be concerned in this section with the estimation of covariance matrices based on data. The problem of covariance estimation is well understood in the low-dimensional regime where the dimensions of the matrix is much smaller than the sample size. In this section, we will be interested in the high-dimensional setting where the matrix dimensions are comparable or possibly much larger than the sample size. This setting arises in many modern relevant applications that deal with community detection, web searches, and recommended systems.

Say that we have m vectors $X_1, \ldots, X_m$ sampled independently from an unknown distribution in $\mathbb{R}^n$. Let us assume for simplicity that this distribution results in zero-mean vector. This means that if X is drawn from this unknown distribution, the covariance matrix we are interested in evaluating is given by $\Sigma = \mathbb{E}[XX^T]$.

Recall

- n: dimensions of all vectors
- m: number of independent samples

Now, to estimate Σ, we can use the sample covariance matrix Σ_m, which is computed from the samples $X_1, \ldots, X_m$ as follows:

$$\Sigma_m = \frac{1}{m}\sum_{k=1}^{m} X_k X_k^T. \quad (91)$$

This estimator is unbiased, i.e., $\mathbb{E}[\Sigma_m] = \Sigma$. It then follows from the law of large number that Σ_m converges to Σ almost surely as $m \to \infty$. This leads

to the following question: how large should m be for the error incurred when approximating Σ with Σ_m be small?

We can use Theorem 25 to answer this question. We first normalize the vectors $X_1, \ldots, X_m$ so that their covariance matrix is the identity. Specifically, we set $X_k = \Sigma^{1/2} A_k$, $k = 1, \ldots, m$, where $\mathbb{E}[A_k A_k^T] = I_n$.

Let now A be the $m \times n$ matrix whose rows are the vectors $\{A_k^T\}$. Note that

$$\Sigma_m = \frac{1}{m} \sum_{k=1}^{m} X_k X_k^T = \frac{1}{m} \sum_{k=1}^{m} \Sigma^{1/2} A_k A_k^T \Sigma^{1/2} = \Sigma^{1/2} (A^T A/m) \Sigma^{1/2}. \tag{92}$$

Then

$$\begin{aligned} \|\Sigma_m - \Sigma\| &= \|\Sigma^{1/2}(A^T A/m)\Sigma^{1/2} - \Sigma\| && (93)\\ &= \|\Sigma^{1/2}(A^T A/m - I_n)\Sigma^{1/2}\| && (94)\\ &\leq \|\Sigma\| \|(A^T A/m - I_n)\|. && (95) \end{aligned}$$

In the last step, we used that for two arbitrary matrices A, B with compatible dimensions, $\|AB\| \leq \|A\|\|B\|$. Assume now that the vectors $X_1, \ldots, X_m$ are sub-Gaussian with parameter σ. Then the vectors $A_1, \ldots, A_m$ are sub-Gaussian with parameter σ at most $\sigma/\sqrt{\|\Sigma\|}$. Then we can apply Theorem 25 and obtain a tail bound on $\|\Sigma_m - \Sigma\|$. Specifically, let, as in Theorem 25, $\delta = c\left[\gamma + \sqrt{\frac{n}{m}}\right]$ for some $\gamma > 0$ (we will not worry about the exact constants.)

Then

$$\mathbb{P}\left[\frac{\|\Sigma_m - \Sigma\|}{\|\Sigma\|} \geq \mathrm{c}(\delta + \delta^2)\right] \leq \mathbb{P}\left[\|(A^T A/m - I_n)\| \geq \mathrm{c}(\max\{\delta, \delta^2\})\right] \leq 2e^{-m\gamma^2}. \tag{96}$$

Substituting the value of δ, we conclude that

$$\mathbb{P}\left[\frac{\|\Sigma_m - \Sigma\|}{\|\Sigma\|} \geq \mathrm{c}\left(\gamma + \sqrt{\frac{n}{m}} + \left(\gamma + \sqrt{\frac{n}{m}}\right)^2\right)\right] \leq 2e^{-m\gamma^2}. \tag{97}$$

It then follows that this method for estimating the covariance matrix yields accurate results when the number of samples is large m, as long as the ratio between the vector dimension n and the number of samples m is sufficiently small. More precisely, the sample covariance matrix Σ_m is a consistent estimate (in the sense of the operator norm) of the population covariance matrix Σ as long as $n/m \to 0$ as $m \to \infty$.

Another way to interpret the result is as follows: roughly speaking, when $m > n$,

$$\frac{\|\Sigma_m - \Sigma\|}{\|\Sigma\|} \lessapprox \sqrt{\frac{n}{m}} + \frac{n}{m}$$

with high probability. So to have an error less than $\epsilon \ll 1$ we need about $m \approx n/\epsilon^2$ points. This is an extremely useful rule of thumb, for practical applications.

4.6 Application: clustering of point sets

- Source: Vershynin, Chapter 4.7.1

We consider now a second application of Theorem 25: the problem of clustering. Specifically, we want to partitions points in $\mathbb{R}^n$, so that the distance between points within the same cluster is small. To keep things simple, we will consider a very simple model for the random generation of points in $\mathbb{R}^n$, introduce an algorithm to perform clustering on points generated according to this model, and then use Theorem 25 to obtain theoretical guarantees on the performance of the algorithm.

As model we will consider a simple **Gaussian mixture model**. Fix a vector $u \in \mathbb{R}^n$ and consider the following random vector:

$$X = Bu + W.$$

Here, B is a Rademacher random variable, which takes value $\{\pm 1\}$ with the same probability, and $W \sim \mathcal{N}(0, I_n)$ and independent of B. The distribution of X is called Gaussian mixture model with means u and $-u$. Then we draw m independent random vectors $X_1, \dots, X_m$ that have the same distribution as X.

The distribution induces two clusters, one centered at u and one centered at $-u$. Our task is to identify which point belongs to which cluster, on the basis of the observations of the points $X_1, \dots, X_m$ alone. In particular the parameter u is unknown to us.

To do so, we use a variation of the spectral clustering algorithm we used for community detection. The idea is as follows. Since all vectors are aligned with u apart from a Gaussian perturbation, we expect that the eigenvector of the normalized matrix $\Sigma_m = \frac{1}{m}\sum_{k=1}^m X_k X_k^T$ corresponding to the largest eigenvalue be close to u. Indeed, consider the matrix $\Sigma = \mathbb{E}[XX^T] = uu^T + I_n$. It is easy to check that $\lambda_1(\Sigma) = \|u\|^2 + 1$ and $\lambda_2 = \cdots = \lambda_m = 1$. Furthermore, the eigenvector $v_1(\Sigma)$ associated to the largest eigenvalue of Σ is $v_1(\Sigma) = u$. Projecting each point on u we can assess to which cluster it belongs. Specifically, a negative projection suggests that the point belongs to the first cluster, whereas a positive projection suggests that the point belongs to the second cluster.

Now, the Davis-Kahan theorem guarantees that if $\|u\|$ is sufficiently large, we can estimate $v_1(\Sigma)$ accurately by computing the eigenvector associated to the largest eigenvalue of Σ_m. This suggests the following spectral-clustering algorithm

---

Input: $X_1, \dots, X_m \in \mathbb{R}^n$ generated according to the Gaussian mixture model
Output: a partition of the point into two clusters

1. Compute the sample covariance matrix Σ_m
2. Compute the eigenvector $v_1(\Sigma_m)$ corresponding to the largest eigenvalue of Σ_m.

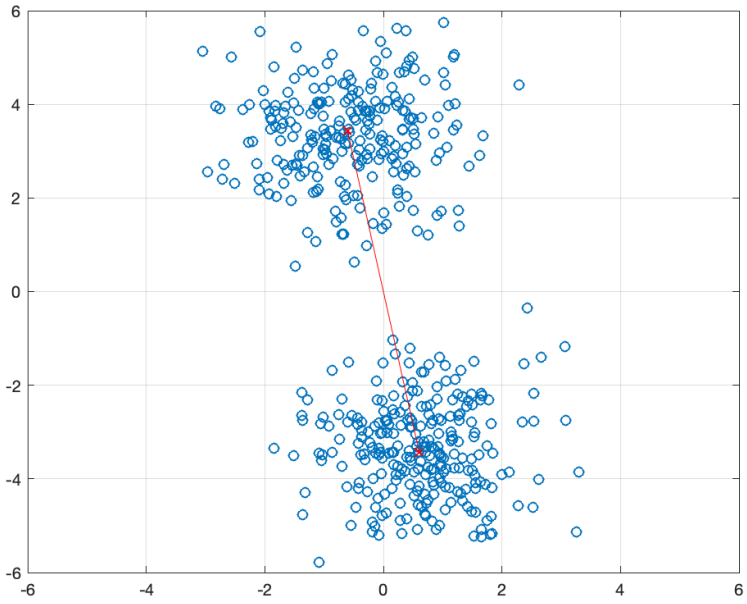

Figure 4: Example of points generated according to the Gaussian mixture model. Here $n = 2$ and $m = 400$

3. For each $k = 1, \dots, m$, put X_k in first community if $v_1^T X_k \geq 0$. Otherwise, put it into the second community.

The theory developed so far allows us to provide theoretical guarantees on the performance of this algorithm. Specifically, we have the following result.

Theorem 26 (Clustering for Gaussian mixture model) *Consider the Gaussian mixture model just described with the additional (simplifying) assumption that $\|u\| = 1$. Let v_1 denote the unit-norm eigenvector of the matrix Σ_m corresponding to the largest eigenvalue. Then there exist a $\phi \in \{\pm 1\}$ such that*

$$\mathbb{P}\left[\|u - \phi v_1\| \geq \mathrm{c}\left(\gamma + \sqrt{\frac{n}{m}} + \left(\gamma + \sqrt{\frac{n}{m}}\right)^2\right)\right] \leq 2e^{-m\gamma^2}$$

One way to interpret this result is that, roughly speaking,

$$\|u - \phi v_1\| \lessapprox \sqrt{\frac{n}{m}} + \frac{n}{m}$$

with high probability. So to have $\|u - \phi v_1\|$ no larger than ϵ one needs $m \approx n/\epsilon^2$ points. The smaller ϵ, the more accurate is our estimate of u and the more the performance of our algorithm approaches the performance that can be achieved in the case in which $\mathbf{u}$ was known to us.

4.7 Exercises

Exercise 14 (Alternative expression for the operator norm) *Prove that the operator norm of a symmetric $n \times n$ square matrix can be expressed as in (66).*

Exercise 15 (Norm of product of matrices) *Show that for two matrices A, B of compatible dimensions, $\|AB\| \leq \|A\|\|B\|$.*

Exercise 16 (Sub-Gaussian vectors) *Let A be a sub-Gaussian vector with parameter 1 (see Definition 23). Let Σ be a symmetric matrix. Show that ΣA is sub-Gaussian with parameter at most $\|\Sigma\|$.*

Exercise 17 (Computing the operator norm on a cover) *Let A be an $m \times n$ matrix and $\epsilon \in [0, 1)$. Prove that for every ϵ-cover $\mathcal{N}$ of the sphere S^{n-1}, we have*

$$\sup_{x \in \mathcal{N}} \|Ax\| \leq \|A\| \leq \frac{1}{1 - \epsilon} \sup_{x \in \mathcal{N}} \|Ax\|. \tag{98}$$

Prove also that if A is $n \times n$ and symmetric, and $\epsilon \in [0, 1/2)$, then for every ϵ-cover $\mathcal{N}$ of the sphere S^{n-1}, we have

$$\sup_{x \in \mathcal{N}} |x^T A x| \leq \|A\| \leq \frac{1}{1 - 2\epsilon} \sup_{x \in \mathcal{N}} |x^T A x|. \tag{99}$$

Exercise 18 (Norm of symmetric matrices with sub-Gaussian entries) *Prove Theorem 20.*

5 Sparse linear models in high dimensions

5.1 Problem formulation and applications

- source: Wainwright, Chapter 7.1

We consider the following linear model: let $x \in \mathbb{R}^n$ be an unknown vector, sometimes called **regression vector**. Suppose we observe a vector $y \in \mathbb{R}^m$ through a (known) measurement matrix $A \in \mathbb{R}^{m\times n}$ via the linear model

$$y = Ax + e$$

where $e \in \mathbb{R}^m$ is a noise vector.

This model is one of the most widely used in statistics and has a long history. In the "low-dimensional" regime where $m \geq n$ the theory on how to recover x from y is well known and includes methods such as least square.

We are interested in this chapter in the "high-dimensional" regime where $m \ll n$. As our intuition suggests, if $m/n < 1$ there is no hope to obtain an accurate estimate of x, even when there is no noise, unless we impose additional constraints on the class of vectors x we intend to recover. One practically relevant assumption on the class of vectors x is that they are **sparse**, i.e., only few of the entries of x are nonzero. Under this sparse assumption, even focusing on the noiseless linear model $y = Ax$ yields nontrivial results. So we will analyze this model, to start with. In particular, we shall be interested in the structure of its "sparse solutions".

Let us start by formalizing the notion of sparsity.

Definition 27 (Support of a vector) *The support of a vector $x \in \mathbb{R}^n$ is the index set of its nonzero entries:*

$$\operatorname{supp}(x) = \{j \in \{1, \dots, n\} : x_j \neq 0\} \tag{100}$$

We say that a vector x is s-sparse if at most s of its entries are nonzero, i.e., if

$$\|x\|_0 = |\operatorname{supp}(x)| \leq s. \tag{101}$$

The quantity $\|x\|_0$ is usually referred to as ℓ_0 norm, although this terminology is somewhat misleading since this quantity is not a norm. Indeed, it does not satisfy the absolute homogeneity property: $\|\alpha x\|_0 \neq |\alpha|\|x\|_0$.

Throughout this chapter, we shall be interested in finding efficient algorithms for solving the system of equations $y = Ax$ for the scenario in which $m \ll n$ and x is s-sparse, with $s \ll n$. Clearly, if we know the position of the s nonzero entries, it is possible to recover x perfectly from $m = s$ measurements. What happens though in the practically relevant case in which the positions of the s nonzero entries are not known? Our main finding is that there exist numerically efficient algorithms to solve this problem, provided that $m \gtrsim s \ln(en/s)$. So if $s \ll n$, a number of measurements m much smaller than the dimension n of the vector x is sufficient to recover x.

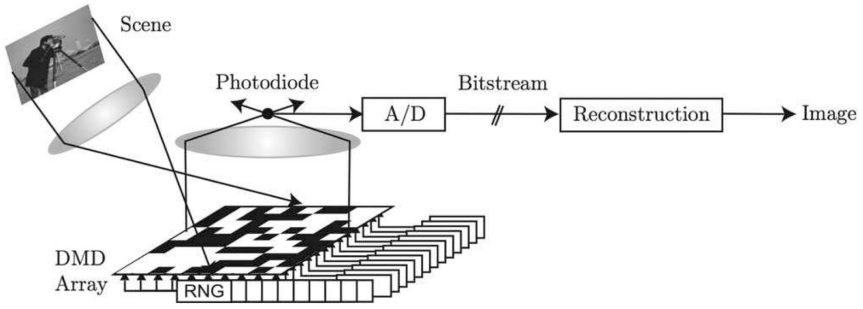

Figure 5: A schematic diagram of the single-pixel camera: a grid of micro-mirrors reflect some parts of the incoming light beam towards the sensor.

5.1.1 Compressive sensing

Compressive sensing is motivated by the wastefulness of the classical approaches to acquire sparse signal.

- Source: chapter 1, [6] (available online through Chalmers library)

Digital photography Consider for example the problem of image acquisition. A raw digital image acquired with a modern digital camera can be as large as tens of Mbytes. However, we can compressed it down to a .jpeg file of size of the order of tens of Kbytes (with a compression factor of order 1000), with only a marginal loss of resolution.

This is possible because digital images are typically sparse when expressed in a suitable basis (e.g., wavelet transform or discrete-cosine transform).

In the current paradigm, we first acquire the whole image (which can we think of a vector in some n dimensional space, with $n \gg 1$), and then we use sparsity to store only its s largest nonzero coefficient, when the vector is expressed in a suitable coordinate system. Compressive sensing is motivated by the following question. Can we avoid taking so many measurements in the first place? Can we exploit sparsity already when measuring the signal, i.e., in the signal acquisition phase? As we shall see, this is indeed possible. And this is not only a theoretical possibility. A proof of concept, called the **single-pixel camera** operating according to this principle was built in 2006 at Rice University in the USA. To learn more about the single-pixel camera, see the following article.

Here is the idea: the camera consists of a micro-array involving a large number of small mirrors that can be turned on and off individually. The light from the image is reflected on this micro-array and all reflected beams from the mirror are focused on a single sensor. The number of active mirrors impact the intensity of the reflected signal. This hardware essentially computes inner products between the image and one row of the measurement matrix. By

flipping the mirrors randomly, one obtains additional measurements. Since the measurement are performed serially, it is desirable to recover the image using as few measurements as possible, which leads precisely to the compressive sensing framework.

Digital cameras built according to this principle are of particular interest for operation at certain wavelengths outside the visible spectrum, where digital sensors are hard to build.

Magnetic resonance imaging Magnetic resonance imaging (MRI) is an important technology in medical imaging, which is used for tasks such as brain imaging, examination of blood vessels and dynamic heart imaging. In traditional approaches, the time required to produce high-resolution images can be several minutes or hours depending on the tasks. This is challenging for patients, and it also prevents the use of this technique in some situations. For example, it is difficult to use it on children, who may have difficulties in staying still for the required long amount of time. In this situation using compressive sensing techniques, which promise accurate images with much fewer measurements, is particularly appealing. The signal measured by MRI turns out to be the spatial Fourier transform of some physical quantity (magnitude of magnetization) from which an image can be recovered. This spatial Fourier transform is actually sparse. So we can use compressive sensing techniques to reduce the number of measurements.

It is worth mentioning that the sparse linear model $y = Ax + e$ occurs naturally in many other applications beyond compressive sensing, including Gaussian sequence models, signal denoising, signal compression, Gaussian graphical models, etc...

5.2 Efficient signal recovery in the noiseless setting

5.2.1 Minimal number of measurements ad the P0 problem

- source: chapter 2.2 and chapter 2.3, [6] (available online through Chalmers library)

Consider the linear model $y = Ax$ for the case in which $m \ll n$, i.e., the number of measurements is much smaller than the dimension of the vector x. We will next investigate under which conditions an s-sparse vector x can be recovered from $y = Ax$.

Note that, for a given sparsity level s, the s-sparse vector x can be recovered from Ax if the vector x is the **unique** s-sparse solution of $Az = y$ with $y = Ax$. In other words, the set $\{z \in \mathbb{R}^n : Az = Ax, \|z\|_0 \leq s\}$ contains only the vector x. Equivalently, The vector x is the **unique** solution of the following minimization problem, which we will refer to as the **(P0)** problem:

$$\min_{z \in \mathbb{R}^n} \|z\|_0 \quad \text{subject to } Az = y. \tag{102}$$

We will first discuss under which conditions on the measurement matrix, the problem (102) admits a unique solution. Then we will discuss how easy it is to solve (102). **Spoiler**: not easy at all!

The theorem below provides an answer to the first question. Not surprisingly, uniqueness of the solution is related to properties of the null-space of A, which is the set of vectors $x \in \mathbb{R}^n$ such that $Ax = 0$. We will denote the null-space by $\ker(A)$.

Theorem 28 (Solution of the (P0) problem) *Let A be a $m \times n$ matrix with real-valued entries. The following statements are equivalent:*

1. *Every s-sparse vector $x \in \mathbb{R}^n$ is the unique s-sparse solution of $Az = Ax$. This means that if $Ax = Az$ and both x and z are s-sparse, then $x = z$.*
2. *The null space $\ker(A)$ of A does not contain $2s$-sparse vectors other than the zero vector.*
3. *Every set of $2s$ columns of A is linearly independent.*

Note that the last condition implies that the number m of measurements for recovering of all s-sparse vectors must satisfy $m \geq 2s$. Indeed, if this is not the case, it is not possible to have $2s$ linearly independent columns. It turns out that one con construct deterministic matrices A with exactly $m = 2s$ rows for which Theorem 28 holds.

For example, the following $2s \times n$ matrix can be shown to satisfy Theorem 28. Let $t_n > \cdots > t_2 > t_1 > 0$. Then set

$$A = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_n \\ \vdots & \vdots & & \vdots \\ t_1^{2s-1} & t_2^{2s-1} & \cdots & t_n^{2s-1} \end{bmatrix}. \tag{103}$$

Unfortunately, the problem (P0) is hard to solve, especially when n is large. Indeed, although the constraint describes a simple subspace, the cost function is non-differentiable and non-convex. Suppose that we know that the vector x to be reconstructed has precisely s non-zero entries and suppose that we have designed A so that the problem (P0) has a unique solution for all s-sparse vectors. Here is how we could recover x from $y = Ax$. Let $A_{\mathcal{S}}$ be the matrix obtained by keeping only the columns of A that are in the subset $\mathcal{S} \subset \{1, \ldots, n\}$ of cardinality s. Then two things can happen.

Either y can be written as a linear combinations of the columns of $A_{\mathcal{S}}$, i.e., y is in the span of $A_{\mathcal{S}}$ or it cannot. If it can be written as a linear combination of columns of $A_{\mathcal{S}}$, then, since these columns are linearly independent, there exists a unique vector $u \in \mathbb{R}^s$ such that $y = A_{\mathcal{S}} u$, and this vector can be found by as follows

$$u = (A_{\mathcal{S}}^{\mathrm{T}} A_{\mathcal{S}})^{-1} A_{\mathcal{S}}^{\mathrm{T}} y. \tag{104}$$

We can then recover x by setting its entries with index in $\mathcal{S}$ equal to the corresponding entries of u and setting to zero all other entries. If y cannot be

written as a linear combination of columns of $A_{\mathcal{S}}$, we need to choose another set $\mathcal{S}$ and start again. Note that, since the solution of the (P0) problem is unique, there is only a single set $\mathcal{S}$ for which this procedure will return a vector u.

Let $A_{\mathcal{S}}$ be the matrix obtained by keeping only the columns of A that are in the subset $\mathcal{S} \in \{1, \dots, n\}$ of cardinality s. Then we can recover x by solving all square systems of equations $A_{\mathcal{S}}^T A_{\mathcal{S}} u = A_{\mathcal{S}}^T y$ for $u \in \mathbb{R}^s$, with $\mathcal{S}$ running through all possible s dimensional subsets of $\{1, \dots, n\}$. By assumption, one and only one of these systems of equations will admit a (unique) solution, whereas all the other systems will have no solutions. Unfortunately, the number of systems of equations to solve is $\binom{n}{s}$, which is very large in the high-dimension regime of interest in this course. For example, if $n = 1000$ and $s = 10$, $\binom{n}{s} \geq 10^{20}$. This means that to find x we need to solve up to 10^{20} systems of equations. Even if we could solve each one of them very rapidly, say in 0.1 ns, we would still need more than 300 years!

So this method is completely impractical. It turns out that any method we may come up with to solve (P0) suffers from the same problem. More specifically, in the language of computational complexity, one can prove that the problem (P0) is NP hard.

5.2.2 A convex relaxation of the P0 problem

One way to avoid the computational complexity associated with the (P0) problem, is to replace the ℓ_0 norm in (102) which is noncovex, with the nearest convex member of in the family of ℓ_q norms, which is the ℓ_1 norm.

Specifically, we define the ℓ_q "norm" of the vector $x \in \mathbb{R}^n$ as

$$\|x\|_q = \left(\sum_{k=1}^{n} |x_k|^q \right)^{1/q}. \tag{105}$$

Note that $\|x\|_q$ is a norm in the strict sense only if $q \geq 1$.

In the figure below we plotted the shape of the unit balls induced by the ℓ_q norm for different values of q. As one can see from the figure, the ℓ_1 norm is convex, whereas the ℓ_q norms, $q < 1$, are not.

Replacing ℓ_0 with ℓ_1 norm is an instance of **convex relaxation**, where a noncovex optimization problem is replaced by a convex one. Doing so we obtain the following **(P1)** optimization problem

$$\min_{z \in \mathbb{R}^n} \|z\|_1 \quad \text{subject to } Az = y. \tag{106}$$

The optimization problem (P1), which is commonly referred to as **basis pursuit**, is a convex program because the cost function is now convex. The question we shall answer next is the following: under which conditions is solving (P1) equivalent to solving (P0)?

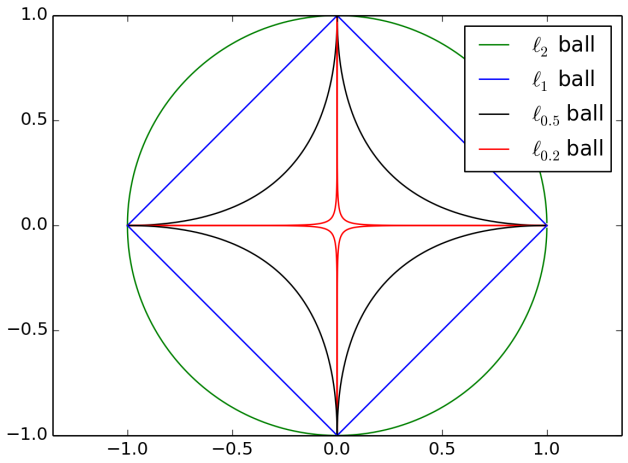

Figure 6: Unit balls induced by different ℓ_q norms in $\mathbb{R}^2$: as $q \to 0$, the ball becomes increasingly spiky

5.2.3 Restricted null-space property and restricted isometry property

- source: Chapter 7.2.2, Wainwright

To answer the question we have just posed, let us first introduce some notation. Let $\mathcal{S}$ be an arbitrary subset of $\{1, \dots, n\}$ of cardinality s. For a vector $x \in \mathbb{R}^n$, we let $x_{\mathcal{S}}$ be the vector in $\mathbb{R}^n$ whose entries with index in $\mathcal{S}$ coincide with that of x and whose remaining entries are zero. Similarly, we denote by $x_{\mathcal{S}^c}$ be the vector in $\mathbb{R}^n$ whose entries with index in $\mathcal{S}^c$ coincide with that of x and whose remaining entries are zero. Here $\mathcal{S}^c = \{1, \dots, n\} \setminus \mathcal{S}$. The following definition will turn out useful.

Definition 29 (Restricted null-space property) *A matrix $A \in \mathbb{R}^{m \times n}$ is said to satisfy the restricted null-space property with respect to a set $\mathcal{S} \subset \{1, \dots, n\}$ if*

$$\|v_{\mathcal{S}}\|_1 < \|v_{\mathcal{S}^c}\|_1, \quad \text{for all } v \in \ker(A) \setminus \{0\}. \tag{107}$$

The following theorem gives a necessary and sufficient condition for the problem (P1) to admit a unique solution.

Theorem 30 ((P1) problem and restricted null-space property) *Given a matrix $A \in \mathbb{R}^{m \times n}$, every vector $x \in \mathbb{R}^n$ supported on the set $\mathcal{S}$ is the unique solution of (P1) with $y = Ax$, if and only if A satisfies the restricted null-space property (RNP) with respect to $\mathcal{S}$.*

We are still left with the problem of how to verify whether a matrix A satisfies the restricted null-space property. We will next introduce a definition that

will help us check whether a matrix satisfies the restricted null-space property. Specifically, the idea is introduce a definition that can be shown to hold with high probability (using our concentration of measure tools) for randomly constructed matrices of appropriate size.

Definition 31 (Restricted isometry property) *For a given integer $s \in \{1, \dots, n\}$, we say that the matrix $A \in \mathbb{R}^{m\times n}$ satisfies the restricted isometry property (RIP) of order s with constant δ_s, if*

$$\left\| \frac{1}{m} A_{\mathcal{S}}^T A_{\mathcal{S}} - I_s \right\| \leq \delta_s \tag{108}$$

for all subsets $\mathcal{S}$ of size s.

The following result shows that satisfying the restricted isometry property is a sufficient condition for the restricted null-space property to hold

Theorem 32 (RIP implies RNP) *If, for a given matrix A, the RIP constant of order $2s$ is bounded as $\delta_{2s} < 1/3$, then the RNP holds for any subset $\mathcal{S}$ of cardinality s.*

To establish this result, we will rely on the following three additional results, which we state separately in the following three lemmas. Note that we will denote now the ℓ_2 norm of a vector x by $\|x\|_2$ instead of $\|x\|$, to avoid confusions with other norms.

Lemma 33 *Let the vector x have s nonzero entries. Then*

$$\|x\|_1 \leq \sqrt{s}\|x\|_2 \tag{109}$$

Lemma 34 *Assume that the matrix $A \in \mathbb{R}^{m\times n}$ satisfies the RIP of order $2s$ with constant δ_{2s}. Let u and v be s-sparse vectors. If $\operatorname{supp}(u) \cap \operatorname{supp}(v) = \emptyset$, then*

$$\frac{1}{m}|u^T A^T A v| \leq \delta_{2s}\|u\|_2\|v\|_2. \tag{110}$$

Lemma 35 *Let a and b be two s-dimensional vectors such that the largest entry in absolute value of b is no larger than the smallest entry in absolute value of a. Then*

$$\|a\|_1 \geq \sqrt{s}\|b\|_2. \tag{111}$$

5.2.4 Random measurement matrices and restricted isometry property

We are now going to show that a large class of random matrices satisfy the RIP with high probability provided that the number of measurements satisfy $m \gtrapprox s\ln(en/s)$. We will consider the class of row-wise sub-Gaussian random matrices we have already encountered in Theorem 25. And indeed, the proof of our main results will turn out to follow along the same lines as the proof of Theorem 25. We will start with a preliminary intermediate result that will be useful to establish the desired result.

Theorem 36 (Concentration bound for row-wise sub-Gaussian matrices) *Let A be an $m \times n$ row-wise sub-Gaussian matrix with parameter σ. Assume also that each row A_k of A satisfies $\mathbb{E}[A_k A_k^T] = I_n$ (i.e., the entries in each row are uncorrelated, but not necessarily independent). Let $r = 16\sigma^2$. Then for all $t \in (0, r]$ and for all $x \in \mathbb{R}^n$*

$$\mathbb{P}\left[\left|x^T\left(\frac{1}{m}A^TA - I_n\right)x\right| \geq t\|x\|_2^2\right] \leq 2e^{-\mathrm{c}t^2 m}. \tag{112}$$

Here, as usual c is an absolute constant that depends only on r. In the reminder of this section, we shall assume that $r \geq 1$. This assumption is not fundamental and can be relaxed with a more refined analysis of the error probability in (112). Equipped with this preliminary result, we can now state the main result of this section.

Theorem 37 (RIP for row-sub-Gaussian matrices) *Let A be an $m \times n$ random matrix satisfying (112). Then,*

$$\mathbb{P}\left[\left\|\frac{1}{m}A_{\mathcal{S}}^TA_{\mathcal{S}} - I_s\right\| > \delta \text{ for some subset } \mathcal{S} \subset \{1,\dots,n\} \text{ of cardinality } s\right] \leq 2e^{-\mathrm{c}\delta^2 m} \tag{113}$$

provided that $m \geq \mathrm{c}\delta^{-2}s\log(en/s)$. This means that, provided that $m \geq \mathrm{c}\delta^{-2}s\log(en/s)$, the matrix A satisfies the RIP of order s with parameter $\delta_s \leq \delta$ with probability at least $1 - 2e^{-\mathrm{c}\delta^2 m}$.

Recall now that the (P1) problem admits a unique s-sparse solution provided that the matrix A satisfies the RIP of order $2s$ with constant $\delta_{2s} < 1/3$. The theorem we have just proven (with s replaced by $2s$ and δ set so that it does not exceed $1/3$) shows that this happens with probability no smaller than $1 - 2e^{-\mathrm{c}\delta^2 m}$, i.e., with high probability!

5.2.5 Generalizations: robustness and stability

In practical applications, the vector x may be sparse only in an approximate sense, i.e, it may have s components whose absolute value is much larger than that of the other components. Furthermore, our measurement y may be affected by noise. It turns out that the theory we have just developed can be extended to these more practical scenarios. If we have sparsity only in an approximate sense, we would like our recovery algorithm to return the best s-sparse approximation of the measured vector. This property is usually referred to as **stability**. Specifically, for a given n-dimensional vector x, we shall denote by $\sigma_s(x)$ the error incurred when approximating x by its best s-term representation in the ℓ_1 sense:

$$\sigma_s(x) = \inf\{\|x - z\|_1,\, z \in \mathbb{R}^n,\, z \text{ is } s\text{-sparse}\}. \tag{114}$$

If we have noise, i.e., $y = Ax + e$, we need to seek the reconstructed sparse vector that is closest to y in some norm (usually, the ℓ_2). A reconstruction algorithm that is able to return such a vector is called **robust**.

To account for the noise, we can modify the (P1) problem as follows:

$$\min_{z\in\mathbb{R}^n} \|z\|_1 \text{ subject to } \|Az-y\|_2 \leq \eta \tag{115}$$

where η expresses the noise tolerance.

This modified (P1) problem is sometimes referred to as **relaxed basis pursuit**. It turns out to be closely related to the so-called **Lasso program** in statistics.

In the following theorem, we show that the relaxed basis pursuit problem is stable and robust whenever the measurement matrix A satisfies the RIP of order $2s$ with sufficiently small constant. This implies that $m \gtrapprox s\ln(en/s)$ measurements are **also** sufficient for the robust and stable reconstruction of approximately sparse vectors from noisy measurements.

Theorem 38 (Robust and stable RIP) *Assume that the $m\times n$ matrix A satisfies the RIP of order $2s$ with constant $\delta_{2s}\leq 0.63$. Then for all $x\in\mathbb{R}^n$ and $y\in\mathbb{R}^m$ with $\|Ax-y\|_2\leq\eta$, a solution $x_\star$ of the relaxed basis pursuit optimization problem (115) approximates the vector x with errors*

$$\|x-x_\star\|_1 \leq c_1\sigma_s(x)+c_2\sqrt{s}\eta \tag{116}$$

$$\|x-x_\star\|_2 \leq \frac{c_1}{\sqrt{s}}\sigma_s(x)+c_2\eta. \tag{117}$$

Here, $c_1>0$ and $c_2>0$ depend only on δ_{2s}.

5.3 Exercises

Exercise 19 (ℓ_q balls and convexity) *Let $q\geq 1$. Prove that the function*

$$\|x\|_q = \left(\sum_{k=1}^n |x_k|^q\right)^{1/q} \tag{118}$$

is convex. Hint: use Minkovski's inequality.

Exercise 20 (Need for ℓ_1 norm) *Let $A\in\mathbb{R}^{m\times n}$, with $m<n$. Prove that there exists a 1-sparse vector x that is not a minimizer of the optimization problem*

$$\min_{z\in\mathbb{R}^n} \|z\|_2 \quad \text{subject to } Az=Ax. \tag{119}$$

Argue that the same is true if we replace the ℓ_2 norm with any ℓ_q norm with $q>1$.

Exercise 21 (Useful lemmas) *Prove Lemma 33, Lemma 34 and Lemma 35*

Exercise 22 ((P1) algorithm) *Implement the (P1) algorithm described in this chapter. Argue first that this algorithm can be recast as a linear program if one introduces the (slack) variables z^+ and z^- where the entries of z^+ coincide*

with that of z whenever the entries of z are positive, and are zero otherwise. Similarly, the entries of z^- coincide with that of $-z$ if the entries of z are negatives and are zero otherwise. This implies that $z = z^+ - z^-$. Choose $A \in \mathbb{R}^{n\times m}$ with independent random entries equal to $1/\sqrt{m}$ or $-1/\sqrt{m}$, each with probability $1/2$. Test the algorithms on randomly generated s-sparse signals, where first the support is chosen at random and then the nonzero coefficients. By varying n, m, s, evaluate the empirical success probability of recovery. Present your findings in a couple of slides, discussing the agreement with the theory. Optional*: test the modified version of the P1 algorithm (relaxed basis pursuit) on noisy measurements and approximately sparse vectors.* Additional resources*: Consult https://web.stanford.edu/~boyd/papers/admm_distr_stats.html for numerically efficient methods to solve the (P1) problem and similar problems.*

6 Low-rank matrix recovery

The problem of low rank matrix recovery (a.k.a., low-rank matrix completion) is the problem of estimating an unknown matrix $X \in \mathbb{R}^{n_1 \times n_2}$ based on (possibly noisy) observations of a subset of its entries. As for the compressive sensing problem studied in the previous section, this problem is ill-posed unless we impose further structure on the class of matrices we want to reconstruct. One practically relevant assumption is that the underlying matrix has low rank, or can be well-approximated by a low-rank matrix.

Why low rank?: an arbitrary $n_1 \times n_2$ matrix ($n_1 < n_2$) is given by specifying $n_1 n_2$ parameters. However, if the matrix is of rank r, the number of parameters reduces to $n_1 r + r(n_2 - r) = r(n_1 + n_2) - r^2$. To get this bound, observe that r columns of the matrix should be linearly independent. So I need $n_1 r$ parameters to describe them. The remaining $n_2 - r$ columns can be expressed as linear combinations of the first r columns. So I need r parameters to describe each one of these columns. This means that it should be possible to reconstruct the matrix by observing roughly $r(n_1 + n_2)$ entries rather than $n_1 n_2$.

6.1 Motivating example: the Netflix problem

- Source: chapter 10 [4] (see Example 10.2).

In 2006, Netflix launched a competition for the best algorithm to predict user ratings for movies, based on previous ratings, without any additional information about the users. Netflix released partial information about how some users rated some movies, and asked the competitors to guess the missing entries. We can organize the available data in a large matrix, where the rows correspond to the users, and the columns correspond to the movies, and the matrix entry in position i, j represent the rating assigned by user i to movie j. The matrix is incomplete and the task given to the competitors was to fill it. The goal is to provide recommendation to the users based on their ratings, i.e., to suggest movies that they have not seen and that they will most likely enjoy watching. It turns out that the singular values of this recommender matrix decay fairly rapidly, which means that the matrix can be well-approximated by a low-rank matrix.

Hence, there is hope to solve this problem. Indeed, the competition was run for several years and many algorithms with satisfactory performance were proposed. Eventually, the competition was stopped in 2010 because of privacy constraints. It turned out that by taking into account additional information available on movie-rating websites, some researchers managed to reconstruct the identity of some of the (supposedly anonymous) users in the portion of database released by Netflix.

In the next section, we will discuss a matrix completion algorithm that is directly inspired by the compressive sensing problem we studied in the previous chapter.

6.2 Efficient matrix recovery

- Source: Chapter 4.6 [6] (see also exercises 6.25 and 9.12 in the same reference)

We will consider the following setup. A matrix $X \in \mathbb{R}^{n_1 \times n_2}$ of rank at most r is observed via the measurement vector $y = \mathcal{A}(X) \in \mathbb{R}^m$. Here $\mathcal{A}(\cdot)$ is some linear mapping from $\mathbb{R}^{n_1 \times n_2}$ to $\mathbb{R}^m$. This mapping could for example describe a mask matrix that allows one to observe only m entries of the matrix X.

As in the compressive sensing case, a natural first approach is to solve this matrix completion problem by seeking the matrix of lowest rank that is compatible with the observation:

$$\min_{Z \in \mathbb{R}^{n_1 \times n_2}} \operatorname{rank}(Z) \quad \text{subject to } \mathcal{A}(Z) = y. \tag{120}$$

Unfortunately, similar to the (P0) problem in compressive sensing, this problem turns out to be NP hard. Note indeed that the rank of Z is the ℓ_0 norm of the vector containing its singular values $\{s_1(Z), \dots, s_n(Z)\}$ where $n = \min\{n_1, n_2\}$.

Motivated by the compressive sensing (P1) problem, we replace the rank minimization problem (120) with a minimization of the ℓ_1 norm of the vector of the singular values. Specifically, let $\|X\|_\star$ be the so called **nuclear norm** of X:

$$\|X\|_\star = \sum_{k=1}^{n} s_k(X). \tag{121}$$

Then we replace the constrained rank minimization problem with the following constrained nuclear norm minimization problem:

$$\min_{Z \in \mathbb{R}^{n_1 \times n_2}} \|Z\|_\star \quad \text{subject to } \mathcal{A}(Z) = y. \tag{122}$$

This problem is a convex optimization problem and it is equivalent to a semidefinite program, which can be solved efficiently.

Similar to the compressive sensing problem, we can establish conditions under which every matrix X with rank at most r is the unique solution of the nuclear norm minimization problem (122).

To state one such condition, we need the following definition.

Definition 39 (Rank-restricted isometry property) *A linear map $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$, satifies the rank restricted isometry property of order r with constant δ_r if*

$$(1 - \delta_r)\|X\|_F^2 \le \|\mathcal{A}(X)\|^2 \le (1 + \delta_r)\|X\|_F^2 \tag{123}$$

for all matrices $X \in \mathbb{R}^{n_1 \times n_2}$ of rank at most r.

Here $\|X\|_F$ is the so-called Frobenius norm of X, which is given by the square root of the sum of the absolute squares of its elements:

$$\|X\|_F = \sqrt{\sum_{i,j} |X_{ij}|^2}. \tag{124}$$

Similar to the compressive sensing setup, one can show that if the mapping $\mathcal{A}$ satisfies the rank restricted isometry property of order $2r$ with parameter $\delta_{2r} < 1/3$, then every matrix $X \in \mathbb{R}^{n_1 \times n_2}$ of rank at most r is the unique solution to the nuclear norm minimization problem (122). Furthermore, a similar result holds also for the case in which the vector of singular values is only approximately sparse, or the observation are contaminated by additive noise with a bounded norm.

We next show that randomly designed measurement maps satisfy the rank restricted isometry property with high probability. Specifically, let us call a linear measurement map $\mathcal{A}$ sub-Gaussian if for every matrix X the m-dimensional vector $\mathcal{A}(X)$ has independent zero-mean sub-Gaussian entries with the same sub-Gaussian parameter σ. Then one can prove that $\mathcal{A}$ satisfies the rank restricted isometry property of order r with constant $\delta_r \leq \delta$ with probability $1 - \epsilon$ provided that

$$m \geq \mathrm{c}(r(n_1 + n_2) + \log(2\epsilon^{-1})) \tag{125}$$

where the constant c depends on δ.

So this result shows that to reconstruct the $n_1 \times n_2$ matrix X of (low) rank r, it is sufficient to take around $r(n_1 + n_2) \ll n_1 \times n_2$ random measurements.

7 Sample complexity in statistical learning theory

In this chapter, we will show how the concentration of measure tools developed so far can be used to understand the performance of general machine learning algorithms used to perform classification or regression in the supervised setting. Throughout this chapter, we will be interested in the following learning problem. We want to develop a learning/prediction rule on the basis of some available training data, and be able to assess the performance of this prediction rule on unseen data.

For example, we may train an machine learning algorithm (say a deep neural network) to perform image classification for, say traffic safety applications, on the basis of the images available in some data set. Then we would like to predict how such algorithm would work when shown unseen (maybe real-world) data.

The field that deals with establishing such results is known as **statistical learning theory**. This is a very active area of research. We will review in this chapter some classical results in statistical learning theory. Many researchers are currently trying to refine these results, so that they become useful in predicting the performance of modern machine learning algorithms, such as deep neural networks.

7.1 The statistical learning framework

- Source: [7], Chapter 2 and 3

In the following, we will use the classification problem as a reference problem to introduce the statistical learning framework. A statistical learning model consists of the following elements:

- **Domain set**: This is an arbitrary set $\mathcal{X}$ that contains the objects we may wish to classify or perform regression over. An example of such a set–the Fashion-MNIST dataset–is provided in Fig. 7. Each object in $\mathcal{X}$ is sometimes referred to as an **instance**, and $\mathcal{X}$ is sometimes referred to as **instance space**.

- **Label set**: This is the set $\mathcal{Y}$ of all possible labels assigned to the objects in $\mathcal{X}$. In the Fashion-MNIST data set considered in Fig 7, the set of labels has cardinality 10, which is the number of classes in the data set.

- **Training data**: $\mathcal{S} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ is a finite sequence of pairs in $\mathcal{X} \times \mathcal{Y}$. Each pair corresponds to a labeled object. For example, the pair (image 4, t-shirt) in the Fashion-MNIST data set is an example of a labeled object. These labeled objects are often called **training examples**, and $\mathcal{S}$ is often referred to as **training set**. In the following, we shall denote each pair (x_k, y_k) by z_k. So the training set $\mathcal{S}$ contains the training examples $z_1, \ldots, z_m$.

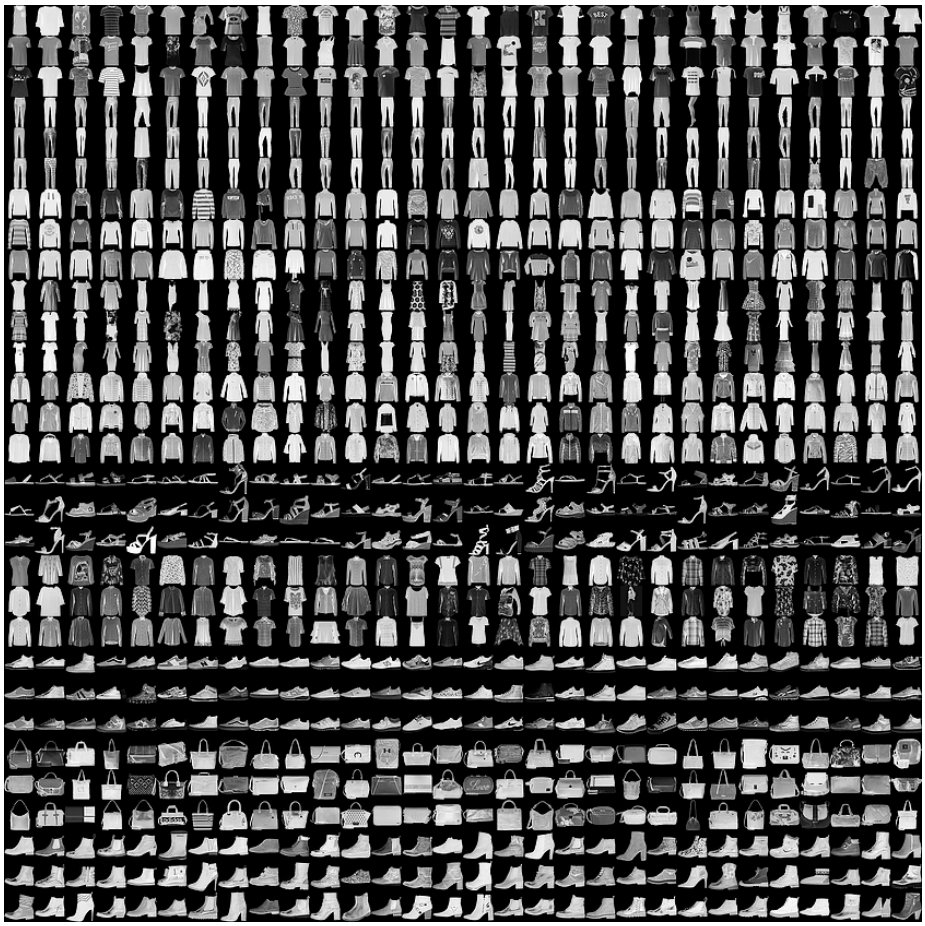

Figure 7: The fashion MNIST data set–a data-set of 70 000 Zalando's article images. Each image is a 28×28 grayscale image, which belongs to one out of 10 classes

- **The learner output**: the goal of the learner is to produce a *prediction rule* $h : \mathcal{X} \to \mathcal{Y}$. This function is also referred to as *predictor*, *hypothesis*, or *classifier*. For a given algorithm A, we denote by $A(\mathcal{S})$ the hypothesis that the learning algorithm A returns on the basis of the training sequence $\mathcal{S}$. The prediction rule is typically assumed to belong to a set $\mathcal{H}$, referred to as *hypothesis class*. Choosing this set appropriately turns out to be very important to enable learning, as we are going to discuss soon.

- **A data generation model**: We assume that the pairs $(x_k, y_k) = z_k$ are generated according to some probability distribution P_Z on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and that the training examples are extracted independently from this distribution. The distribution P_Z may for example consist of a distribution P_X on $\mathcal{X}$ and a deterministic rule that assigns a label y to each x . But we will consider a more general and practically relevant setup where the label assigned to each x is described by a probabilistic map. A key assumption in statistical learning theory is that the distribution P_Z is unknown to the learner.

- **Measures of success**: We measure the performance of a classifier h by assessing the quality of its prediction on a random example generated according to P_Z. Specifically, the quality of the prediction provided by a hypothesis h is measured in terms of a nonnegative loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$. In a classification problem, it is natural to consider as loss function the following 0–1 loss. Let $z = (x, y)$ and fix a classifier h. Then, for the case of 0–1 loss, we have that $\ell(h, z) = 0$ if $h(x) = y$ and $\ell(h, z) = 1$ if $h(x) \neq y$. In other words, the 0–1 loss is the indicator function of the error event. In a regression problem, we may instead be interested in finding some patterns in the data. For example, we may want to learn a predictor for the relation between wage and educational level of people living in Gothenburg. In such a case, a measure of success is better expressed as the square difference between the true label and their predicted values. So it is natural to set $\ell(h, z) = (h(x) - y)^2$. We define the **risk function** $L_{P_Z}(h)$ to be the expected loss of the classifier, namely

$$L_{P_Z}(h) = \mathbb{E}_{P_Z}[\ell(h, Z)]. \tag{126}$$

  This quantity is also referred to in the literature as **population error**, **true error** or **generalization error** of h.

Ideally, we would like to choose a classifier that minimizes $L_{P_Z}(h)$. Unfortunately, this is not possible because the learner does not know P_Z and, hence, it cannot compute the risk function $L_{P_Z}(h)$. We will discuss next a simple strategy to choose a classifier h.

7.2 Empirical risk minimization

- Source: [7], Chapter 2 and 3

The learner receives as input the training set $\mathcal{S}$. The goal of the learner is to choose an $h \in \mathcal{H}$ on the basis of $\mathcal{S}$ that leads to low risk $L_{P_Z}(h)$. To do so, the learner can compute the performance of h on the training data. Specifically, it evaluates the **training error** $L_{\mathcal{S}}(h)$ defined as the empirical average of the loss function over the training set:

$$L_{\mathcal{S}}(h) = \frac{1}{n} \sum_{k=1}^{n} \ell(h, z_k). \tag{127}$$

This quantity is sometimes referred to also as **empirical error** or **empirical risk**. It is then natural for the learner to choose the hypothesis h that minimizes the training error. We refer to this learning paradigm as **empirical risk minimization** (ERM) and to the corresponding hypothesis as **ERM classifier** (or **ERM hypothesis**):

$$h_{\mathcal{S}} \in \arg\min_{h \in \mathcal{H}} L_{\mathcal{S}}(h). \tag{128}$$

Note that the notation stresses that the hypothesis may not be unique.

The hope is that an h that minimizes the empirical risk with respect to the training sample $\mathcal{S}$ is also a risk minimizer, or has risk close to the minimum with respect to the true data distribution as well. In the remainder of this chapter, we will be interested in determining when this is true.

7.3 ERM and overfitting

- Source: [7], Chapter 2 and 3

Consider the following regression problem. We want to learn the polynomial function $y = \sum_{d=0}^{D'} a_d x^d$ from noisy observations of points on this polynomial. It is natural then to choose as $\mathcal{H}$ the set of all polynomials up to a given degree D. The question we investigate next is the following: how does the ERM classifier perform as a function of D? We consider as loss function the quadratic loss, so determining the ERM classifier on the basis of some training samples is equivalent to solving a simple least-square problem, which can be solved efficiently. Some numerical results are shown in Figure 8.

For the example shown in the figure, one achieves a good fit (i.e., a low risk) between the predicted polynomial function and the ground truth when $D = 3$. If we increase D further, the fit becomes worst. However, the training error keeps on becoming smaller and smaller, until it reaches 0 when $D = 9$ and the blue curve passes through the 10 training data. This phenomenon is called **overfitting** and occurs when the selected hypothesis fits the training data too well.

As shown in the figure, one way to reduce overfitting is to restrict the set of hypotheses $\mathcal{H}$. Specifically, the learner should choose in advance (before seeing the data) a set of predictors. By limiting the set of predictors, we introduce a bias, which is referred to as **inductive bias**. The inductive bias should ideally reflect some prior knowledge on the problem to be learned. For example, in

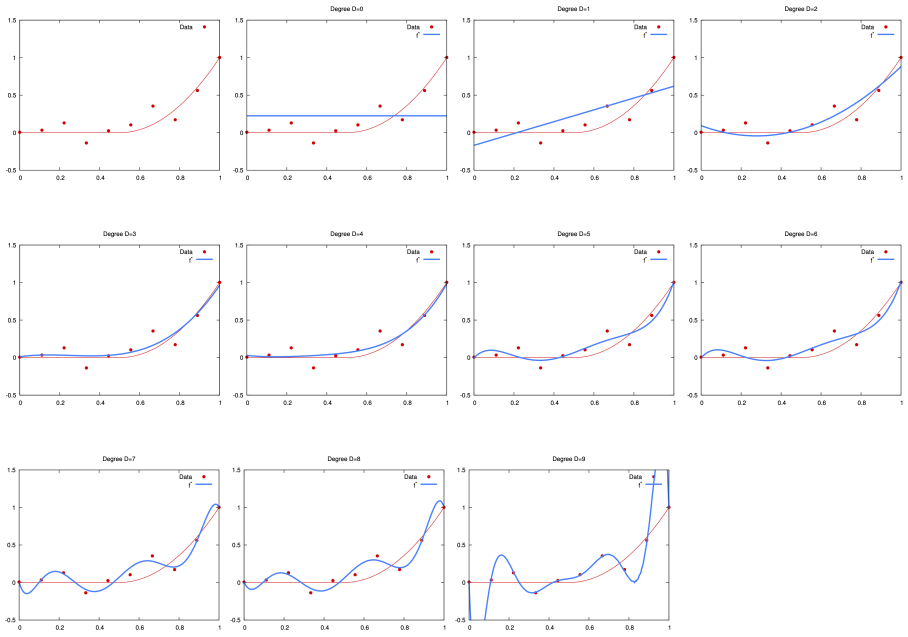

Figure 8: Estimated polynomial function (in blue) versus the ground truth (in blue). The training examples are marked as red dots. This figures is taken from the lecture notes of Prof. Francois Fleuret's EE-559 Deep Learning course at EPFL

the example shown on Fig. 8, the inductive bias may reflect some *a priori* information about the rate at which the function to be learned is supposed to change.

As discussed in the last group project, one possible way to determine parameters such as D in our example is through **cross-validation**: a sub-set of the available training data is held out and used to evaluate empirically the generalization performance of the chosen classifier.

7.4 PAC learning

- Source: [7], Chapter 3

Let $h_{\mathcal{S}}$ be a classifier that is chosen on the basis of $\mathcal{S}$, for example via the ERM principle, i.e., by minimizing the training error. Since the set $\mathcal{S}$ contains randomly generated training examples, the classifier $h_{\mathcal{S}}$ is random, and, hence, also the risk function $L_{P_Z}(h_{\mathcal{S}})$ is random. We will be interested in characterizing the probability that the difference between the risk $L_{P_Z}(h_{\mathcal{S}})$ and the risk achieved by the best algorithm $h \in \mathcal{H}$, which is given by $\min_{h' \in \mathcal{H}} L_{P_Z}(h')$ does not exceed a certain *accuracy parameter* $\epsilon \in (0,1)$. In particular, we are interested in determining the minimum number of training examples that guarantees that this probability is no smaller than some *confidence parameter* $1-\delta$ with $\delta \in (0,1)$.

For a given classifier $h_{\mathcal{S}}$, chosen as a function of $\mathcal{S}$, we interpret the event $L_{P_Z}(h_{\mathcal{S}}) > \min_{h' \in \mathcal{H}} L_{P_Z}(h') + \epsilon$ as a failure of the learner, whereas if $L_{P_Z}(h_{\mathcal{S}}) \leq \min_{h' \in \mathcal{H}} L_{P_Z}(h') + \epsilon$, we view the output of the algorithm as **approximately correct**.

To summarize, we want to study under which conditions the following probability

$$\mathbb{P}\left[L_{P_Z}(h_{\mathcal{S}}) \leq \min_{h' \in \mathcal{H}} L_{P_Z}(h') + \epsilon\right] \tag{129}$$

is larger or equal to $1-\delta$. This probability is evaluated with respect to the randomness in the generation of the training set $\mathcal{S}$. In other words, we are interested in determining when the output hypothesis is **probably approximately correct (PAC)**. We are now ready to provide a formal definition of **PAC learnability**.

Definition 40 (PAC learnability) *A hypothesis class $\mathcal{H}$ is PAC learnable with respect to a set $\mathcal{Z}$ and a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$ if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: for every $\epsilon, \delta \in (0,1)$ and for every distribution P_Z over $\mathcal{Z}$, when running the algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated using P_Z, the algorithm returns a $h \in \mathcal{H}$ such that with probability at least $1-\delta$ over the choice of the m training examples*

$$L_{P_Z}(h) \leq \min_{h' \in \mathcal{H}} L_{P_Z}(h') + \epsilon. \tag{130}$$

The function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ determines the **sample complexity** required for learning $\mathcal{H}$, i.e., how many examples are required to guarantee a probably approximately correct solution. The sample complexity is a function of the accuracy ϵ and confidence δ parameters. It also depends on the property of the hypothesis class. For example, as we will see soon, for a hypothesis class $\mathcal{H}$ with finite cardinality, the sample complexity depends on the logarithm of the size of $\mathcal{H}$.

7.5 No-free-lunch theorem

In this section, we show that without introducing some restriction of the hypothesis class $\mathcal{H}$, PAC learnability is not possible. Specifically, for every algorithm A, one can construct a distribution P_Z for which learning is hard. To prove that this is indeed the case, it is sufficient to focus on the simple problem of binary classification, where $\mathcal{Y} = \{0,1\}$ and consider the 0–1 loss.

Theorem 41 (No-free-lunch theorem) *Let A be a learning algorithm for the task of binary classification with respect to the 0–1 loss over a domain $\mathcal{X}$. Let $m \leq |\mathcal{X}|/2$ be the training size. Then, there exists a distribution P_Z over $\mathcal{X} \times \{0,1\}$ such that i) There exists a function $f : \mathcal{X} \to \{0,1\}$ with $L_{P_Z}(f) = 0$. ii) With probability at least $1/7$ over the choice of the training set $\mathcal{S}$ of dimension m, we have that $L_{P_Z}(A(\mathcal{S})) \geq 1/8$.*

Roughly speaking, this theorem says the following: pick an arbitrary algorithm for binary classification over $\mathcal{X}$ with respect to the 0–1 loss function. Then one can construct a data distribution P_Z, such that there exists a binary classification function with zero population error, but such that the algorithm we selected fails to find a binary classification function that has population error smaller than $1/8$ with probability $1/7$ even when the algorithm is given as input a training set containing half of the elements in $\mathcal{X}$.

The following result then follows directly from the no-free-lunch theorem.

Theorem 42 (A class that is not PAC learnable) *Let $\mathcal{X}$ an infinite domain set and let $\mathcal{H}$ be the set of all functions from $\mathcal{X}$ to $\{0,1\}$. Then $\mathcal{H}$ is not PAC learnable.*

Indeed, consider for example the ERM predictor over the hypothesis class $\mathcal{H}$ of all the functions from $\mathcal{X}$ to $\{0,1\}$. Note that no prior knowledge (or inductive bias) is embedded in the choice of this class: every possible function is considered as a potential candidate. But according to the no-free-lunch theorem, any algorithm that chooses a hypothesis from $\mathcal{H}$ on the basis of $m \leq |\mathcal{X}|/2$ examples, including the EMR algorithm, will fail on some learning task specified by a distribution P_Z. In particular, if $|\mathcal{X}| = \infty$, no PAC-learnability can be guaranteed, because m can be chosen arbitrarily large.

7.6 Uniform convergence is sufficient for PAC learnability

- Source: [7], Chapter 4.1

As already discussed, a practical way to select an algorithm h on the basis of the set $\mathcal{S}$ is to use empirical risk minimization, i.e., to select the h that minimizes the empirical risk, which can be computed by the learner. Intuitively, if we ask that all $h \in \mathcal{H}$ have an empirical risk $L_{\mathcal{S}}(h)$ that is close to their true risk $L_{P_Z}(h)$, we should be able to ensure PAC learnability. In other words, we are asking for the empirical risk to be close to the true risk uniformly over all hypotheses in the hypothesis class. This is formalized in the following definition

Definition 43 (ϵ-representative sample) *A training set $\mathcal{S}$ is called ϵ-representative if for all $h \in \mathcal{H}$,*

$$|L_{P_Z}(h) - L_{\mathcal{S}}(h)| \leq \epsilon. \tag{131}$$

In the next lemma, we show that if a training set if $\epsilon/2$ representative, then the ERM learning rule returns a "good" hypothesis. By good we mean that the true risk achieved by the EMR learning rule is at most ϵ away from the true risk achieved by the best $h \in \mathcal{H}$.

Lemma 44 *Assume that the training set $\mathcal{S}$ is $\epsilon/2$-representative. Then every ERM hypothesis $h_{\mathcal{S}} \in \arg\min_{h\in\mathcal{H}} L_{\mathcal{S}}(h)$ satisfies*

$$L_{P_Z}(h_{\mathcal{S}}) \leq \min_{h\in\mathcal{H}} L_{P_Z}(h) + \epsilon. \tag{132}$$

Now to conclude that the ERM rule is a PAC learner, it suffices to show that the training set $\mathcal{S}$ is $\epsilon/2$-representative with probability at least $1-\delta$. We will show next that this is indeed the case for bounded loss functions and when the hypothesis class $\mathcal{H}$ has finite cardinality.

7.7 Finite hypothesis classes are PAC learnable

- Source: [7], Chapter 4.2

Let us assume that $\mathcal{H}$ is a finite hypothesis class, i.e., $|\mathcal{H}| < \infty$. Let us also assume that $\ell(\cdot,\cdot)$ is supported on the bounded set $[0,1]$. The following result holds.

Theorem 45 (Sample complexity of finite classes) *Under the assumptions just stated, $\mathcal{H}$ is PAC learnable using the ERM algorithm with sample complexity*

$$m_{\mathcal{H}}(\epsilon,\delta) \leq \left\lceil \frac{2}{\epsilon^2} \log\left(\frac{2|\mathcal{H}|}{\delta}\right) \right\rceil. \tag{133}$$

7.8 Infinite-size classes can also be learnable

- Source: [7], Chapter 6.1

The requirement that the size of the hypothesis class is finite turns out to be too stringent. Indeed, hypothesis classes with infinite size are also learnable. Consider the following example. Let $\mathcal{H}$ be the set of threshold functions on the real line. Namely $\mathcal{H} = \{h_a, a \in \mathbb{R}\}$ where $h_a(x) = 1$ if $x < a$ and $h_a(x) = 0$ otherwise. Clearly $\mathcal{H}$ has infinite size. It turns out that $\mathcal{H}$ is PAC learnable using the ERM algorithm.

Theorem 46 (Learning threshold functions) *Let $\mathcal{H}$ be the class of threshold functions just defined. Then $\mathcal{H}$ is PAC learnable with respect to the* 0–1 *loss, using the ERM rule with sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{1}{\epsilon} \log \frac{2}{\delta} \right\rceil$.*

Roughly speaking, this hypothesis class is PAC learnable, because it is characterized by a single parameter (i.e., a). Examples similar to this one have spurred interest in seeking properties of the hypothesis classes that give a correct characterization of their learnability. The underlying theory has been developed by Vapnik and Chervonenkis for the case of binary classification problems and 0–1 loss and resulted in a fundamental quantity known as VC dimension. We will not have the time to formally introduce this quantity in this course. It suffices to say that the VC dimension of the set of threshold functions over the real line is 1 and the VC dimension of a finite hypothesis class is no larger than the logarithm of its cardinality. This leads us to the following fundamental theorem of PAC learning.

Theorem 47 (Fundamental theorem of statistical learning) *Let $\mathcal{H}$ be a hypothesis class of functions from $\mathcal{X}$ to the set $\{0, 1\}$. Let the loss function be the* 0–1 *loss. Assume that the VC dimension of $\mathcal{H}$ is finite and equal to d. Then there exist absolute constant c_1 and c_2 such that $\mathcal{H}$ is PAC learnable with sample complexity*

$$c_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq c_2 \frac{d + \log(1/\delta)}{\epsilon^2} \tag{134}$$

Conversely, if $d = \infty$, the hypothesis class is not PAC learnable.

7.9 PAC learning and deep neural networks

The reason why modern machine learning algorithms perform so well is still a puzzle to theoreticians. Answering this question is an important research topic in theoretical machine learning and an active area of research at Chalmers. In this section, we will outline through an example why the classical PAC learning framework that we have just introduced cannot be used to explain the performance of neural networks.

The VC dimension of a neural network is roughly speaking proportional to the number of parameters (i.e., weights) of the neural network. But this number is typically very large: for example, state of the art convolutional deep neural networks used for image classifications contain around 10^6 parameters. It then follows from Theorem 47 that to get an error not exceeding 10^{-1} one would need a number of examples on the order of 10^8. Yet, numerical experiments suggest that training the network on 10^4 examples yields the desired accuracy on the CIFAR-10 image dataset. This shows the existence of a profound discrepancy between theory and practice. Many approaches are currently under investigation to address this discrepancy. One such approach is to obtain algorithmic-dependent bounds that are based on a Bayesian generalization of the PAC framework (PAC-Bayes) and that rely on information-theoretic tools.

7.10 Exercises

Note: for Exercises 25, 26, 27, we assume realizability, i.e., that there exists a $h^* \in \mathcal{H}$ with zero population error.

Exercise 23 (True error equals expected empirical error) *Let $\mathcal{H}$ be a hypothesis class of classifiers over a domain $\mathcal{X}$. Let $\mathcal{D}$ be an unknown distribution over $\mathcal{X}$, and let f be the target hypothesis in $\mathcal{H}$. The data generation rule P_Z is determined by $(\mathcal{D}, f)$: each pair (x_i, y_i) in the training data set $\mathcal{S}$ of size m is generated by first sampling x_i according to $\mathcal{D}$ and then labeling it by $y_i = f(x_i)$. Fix some $h \in \mathcal{H}$ and consider the 0-1 loss, i.e., for $z = (x, y)$, $\ell(z, h) = \mathbb{1}\{h(x) \neq y\}$. Show that the expected value of the empirical error $L_{\mathcal{S}}(h)$ over the choice of the set $\{x_i\}_{i=1}^m$ equals the true error $L_{P_Z}(h)$, i.e.,*

$$\mathbb{E}_{\{x_i\}_{i=1}^m \sim \mathcal{D}^m}[L_{\mathcal{S}}(h)] = L_{P_Z}(h). \tag{135}$$

Exercise 24 (Monotonicity of sample complexity) *Let $\mathcal{H}$ be a hypothesis class for a binary classification task. Suppose that $\mathcal{H}$ is PAC learnable with sample complexity $m_{\mathcal{H}}(\cdot, \cdot)$. Show that $m_{\mathcal{H}}$ is monotonically nonincreasing in each of its parameters. That is, show that given $\delta \in (0, 1)$ and $0 < \epsilon_1 \leq \epsilon_2 < 1$, it holds that $m_{\mathcal{H}}(\epsilon_1, \delta) \geq m_{\mathcal{H}}(\epsilon_2, \delta)$. Similarly, show that given $\epsilon \in (0, 1)$ and $0 < \delta_1 \leq \delta_2 < 1$, it holds that $m_{\mathcal{H}}(\epsilon, \delta_1) \geq m_{\mathcal{H}}(\epsilon, \delta_2)$.*

Exercise 25 (Axis-aligned rectangles) *An axis aligned rectangle classifier in the plane is a classifier that assigns the value 1 to a point if and only if it is inside a certain rectangle. Formally, given real numbers $a_1 \leq b_1$, $a_2 \leq b_2$, define the classifier $h_{(a_1, b_1, a_2, b_2)}$ by*

$$h_{(a_1, b_1, a_2, b_2)}(x_1, x_2) = \mathbb{1}\{a_1 \leq x_1 \leq b_1, a_2 \leq x_2 \leq b_2\}. \tag{136}$$

With a slight abuse of notation, we call the instances with label 1 the positive instances, and the instances with label 0 the negative instances. The class of all axis aligned rectangles in the plane is defined as

$$\mathcal{H}_{rec}^2 = \{h_{(a_1, b_1, a_2, b_2)}(x_1, x_2) \colon a_1 \leq b_1, a_2 \leq b_2\}.$$

Note that this is an infinite-size hypothesis class.

1. *Let A be the algorithm that returns the smallest rectangle enclosing all positive instances in the training set. Show that A is an empirical risk minimizer (ERM).*

2. *Show that if A receives a training set of size $m \geq \left\lceil \frac{4 \log(4/\delta)}{\epsilon} \right\rceil$ then, with probability of at least $1 - \delta$ it returns a hypothesis with error of at most ϵ with respect to the 0-1 loss. That is, show that $\mathcal{H}_{rec}^2$ is PAC learnable with respect to the 0-1 loss with sample complexity $m_{\mathcal{H}_{rec}^2}(\epsilon, \delta) \leq \left\lceil \frac{4 \log(4/\delta)}{\epsilon} \right\rceil$.*

   Hint*: For some distribution $\mathcal{D}$ over $\mathcal{X}$, let $R^* = R(a_1^*, b_1^*, a_2^*, b_2^*)$ be the rectangle that generates the labels, and let f be the corresponding hypothesis. Let $a_1 \geq a_1^*$ be a number such that the probability mass (with respect*

to $\mathcal{D}$) of the rectangle $R_1 = R(a_1^*, a_1, a_2^*, b_2^*)$ is exactly $\epsilon/4$. Similarly, let b_1, a_2, b_2 be the numbers such that the probability masses of the rectangles $R_2 = R(b_1, b_1^*, a_2^*, b_2^*)$, $R_3 = R(a_1^*, b_1^*, a_2^*, a_2)$, and $R_4 = R(a_1^*, b_1^*, b_2, b_2^*)$ are all exactly $\epsilon/4$. Let $R(\mathcal{S})$ be the rectangle returned by A for a training set $\mathcal{S}$. See the illustration in Fig. 9.

- *Show that $R(\mathcal{S}) \subseteq R^*$.*
- *Show that if $\mathcal{S}$ contains (positive) instances in all of the rectangles R_1, R_2, R_3, R_4, then the hypothesis returned by A has error of at most ϵ.*
- *For each $i \in \{1, 2, 3, 4\}$, upper bound the probability that $\mathcal{S}$ does not contain an instance from R_i.*
- *Use the union bound to conclude the argument.*

Remark: *More generally, the class $\mathcal{H}_{rec}^d$ of axis aligned rectangles in $\mathbb{R}^d$ is PAC learnable with respect to the* 0-1 *loss with sample complexity $m_{\mathcal{H}_{rec}^2}(\epsilon, \delta) \leq \left\lceil \frac{2d \log(2d/\delta)}{\epsilon} \right\rceil$.*

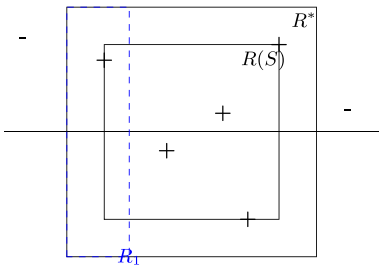

Figure 9: Axis-aligned rectangles

Exercise 26 (Singleton) *Let $\mathcal{X}$ be a discrete domain, and let $\mathcal{H}_{\text{singleton}} \triangleq \{h_z \colon z \in \mathcal{X}\} \cup \{h^-\}$ where $h_z(x) \triangleq \mathbb{1}\{x = z\}, x \in \mathcal{X}$, and h^- is the all-negative hypothesis, i.e., $h^-(x) = 0, \forall x \in \mathcal{X}$. That is, a hypothesis in $\mathcal{H}_{\text{singleton}}$ labels negatively all instances in the domain, perhaps except one.*

1. *Describe an algorithm that implements the ERM rule for learning $\mathcal{H}_{\text{singleton}}$.*
2. *Show that $\mathcal{H}_{\text{singleton}}$ is PAC learnable. Provide an upper bound on the sample complexity.*

Exercise 27 (Concentric circles) *Let $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{0, 1\}$, and let $\mathcal{H}_{\text{circ}}$ be the class of concentric circles in the plane, that is, $\mathcal{H}_{\text{circ}} \triangleq \{h_r \colon r \in \mathbb{R}_+\}$, where $h_r(x) \triangleq \mathbb{1}\{\|x\| \leq r\}$.*

1. *Describe an algorithm that implements the ERM rule for learning* $\mathcal{H}_{\text{circ}}$.
2. *Show that* $\mathcal{H}_{\text{circ}}$ *is PAC learnable with sample complexity bounded by* $m_{\mathcal{H}_{\text{circ}}}(\epsilon, \delta) \leq \left\lceil \frac{\log(1/\delta)}{\epsilon} \right\rceil$.

Exercise 28 (Bounded loss function) *In Theorem 45 (sample complexity of finite classes) in the lecture notes, we assumed that the range of the loss function in* $[0, 1]$. *Prove that if the range of the loss function is* $[a, b]$ *then the sample complexity satisfies*

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{2(b-a)^2}{\epsilon^2} \log\left(\frac{2|\mathcal{H}|}{\delta}\right) \right\rceil. \tag{137}$$

Hint: *Apply Hoeffding's inequality to obtain*

$$\mathbb{P}_{\mathcal{S}}\left[|L_{P_Z}(h) - L_{\mathcal{S}}(h)| \geq \epsilon/2\right] \leq 2\exp\left(-\frac{2m\epsilon^2}{(b-a)^2}\right)$$

where m *is the training set size.*

References

[1] R. D. Yates and D. J. Goldman, *Probability and Stochastic Processes*. Singapore: Wiley, 2015.

[2] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

[3] A. S. Bandeira, Singer, and Strohmer, *Mathematics of Data Science*, Jun. 2020, draft 0.1. [Online]. Available: https://people.math.ethz.ch/~abandeira/BandeiraSingerStrohmer-MDS-draft.pdf

[4] M. J. Wainwright, *High-dimensional statistics: a nonasymptotic viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[5] E. Abbe, *Community detection and stochastic block model*, ser. Foundations and Trends in Communications and Information Theory. Now Publisher, 2018, vol. 14, no. 1-2.

[6] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*, ser. Applied and numerical harmonic analysis. New York, NY: Birkhäuser, 2013.

[7] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: from theory to algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2014.