Chapter 1

Introduction

Our main focus is randomness, and specifically, the way randomness may be used to expose hidden structures in sets. Using randomness in this way has been one of the central themes of *Asymptotic Geometric Analysis*, an area devoted to the study of convex sets in $\mathbb{R}^n$. What is less widely known is that connections between randomness and structure are at the heart of *Statistical Learning Theory*.

Statistical Learning Theory, and more generally, Nonparametric Statistics, study *prediction* and *estimation* problems. Roughly put, a random sample is used to approximate an unknown random variable by a function selected wisely from a given class of functions. Because of the nature of the given data, randomness obviously plays an essential role in learning problems, but connecting this with 'structure' seem a little far-fetched at this point.

Exploring the roles of structure and its preservation through random sampling have in Statistical Learning Theory will be the main theme of these notes.

To give some indication of why problems involving sampling tend to be difficult, let us begin by describing a toy example: selecting randomly a subset of the coordinates of a single vector in $\mathbb{R}^M$. This example displays many of the issues we will have to contend with in what follows, and which will have to be addressed not for a single object – like a single vector in $\mathbb{R}^M$ – but rather uniformly, for an infinite family of objects.

Contrary to what one may intuitively think, the fact that a vector (or a function for that matter) is bounded with respect to some natural norm says very little on the effectiveness of sampling, and despite being bounded, the outcome of a sampling procedure may be totally different than the original

object. To illustrate this, let us consider a very simple sampling model: given a vector $v \in \mathbb{R}^M$ (and M is very large), one selects uniformly at random N coordinates from $\{1, ..., M\}$. If I is the set of the coordinates selected, the hope is that the sampled vector $(v_i)_{i\in I}$ 'inherits' the properties of the vector $(v_i)_{i=1}^M$ one considers significant. Take for example, the ℓ_2 norm

$$\|v\|_2 = \left(\sum_{i=1}^{M} |v_i|^2\right)^{1/2}$$

and let us see the extent in which sampling preserves it. Clearly, the best outcome one can hope for is when v is the constant vector, in which case, for every $I \subset \{1, ..., M\}$, if $P_I v = (v_i)_{i\in I}$ then

$$\|v\|_2 = \left(\frac{M}{|I|}\right)^{1/2} \|P_I v\|_2.$$

This will serve as a basis for comparison: a perfect outcome of a choice of I means that $\|P_I v\|_2 = (|I|/M)^{1/2}\|v\|_2$ – sampling shrinks the Euclidean norm by a factor of $(|I|/M)^{1/2}$. Now, let

$$v_1 = (1, 0, ..., 0) \quad \text{and} \quad v_2 = (1/\sqrt{M}, ..., 1/\sqrt{M});$$

both vectors belong to the Euclidean unit sphere. However, v_1 and v_2 respond in very different ways to a choice of a random subset of their coordinates. Indeed, for any reasonable definition of a random subset $I \subset \{1, ..., M\}$, even of large cardinality, and certainly for the uniform choice, the likely outcome is that the first coordinate will not be selected. Therefore, in the typical case, $\|P_I v_1\|_2 = 0$ – very far from the benchmark value of $\sqrt{|I|/M}$. In contrast, $\|P_I v_2\|_2 = \sqrt{|I|/M}$, as expected.

The way the two vectors respond to sampling happens to be an outcome of their different structures. The Euclidean norm of v_1 is due to a single coordinate, and in that sense, v_1 is a *peaky* vector. In contrast, v_2 is *well-spread*, as all of its coordinates are the same[1]. Clearly, having some information on the Euclidean norm of v says absolutely nothing about the success of sampling, and like-wise, information on the ℓ_p norm of the vector for other values of p is equally useless. We will return to this observation when we explore learning problems involving classes of functions that are all bounded by some fixed constant.

[1] Although we will use the terms rather freely, one should take care when considering the notions of 'peaky' and 'well spread'; for example, is the vector $(1/\sqrt{2}, 1/\sqrt{2M}, ...1/\sqrt{2M})$ 'peaky' or 'well-spread'?

Intuitively and somewhat inaccurately, sampling works reasonably well when the desired property is captured by a 'large set', and in this case, by a set consisting of many coordinates. Indeed, a natural way of ensuring that $\|P_I v\|_2$ is large enough, say of the order of $\sqrt{|I|/M}\|v\|_2$ is that the set of coordinates

$$J_{\alpha,\beta} = \left\{ i : \|v\|_2 \frac{\alpha}{\sqrt{M}} \leq |v_i| \leq \|v\|_2 \frac{\beta}{\sqrt{M}} \right\} \tag{1.1}$$

has cardinality that is proportional to M – say for $\alpha = 1/2$ and $\beta = 2$; in that case, and because $J_{\alpha,\beta}$ is such a large set, a typical random choice of $I \subset \{1, ..., M\}$ satisfies that $|I \cap J_{\alpha,\beta}| \sim N$, leading to the desired outcome. However, the property that $J_{\alpha,\beta}$ is large is not captured by some natural norm of v.

At this point, a word of warning is called for: the first step in exploring sampling problems is to identify the property one wishes sampling to preserve. For example, let

$$v = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2M}}, ..., \frac{1}{\sqrt{2M}} \right),$$

note that $\|v\|_2 = 1$ and that for $p > 2$,

$$\|v\|_p = \left(\sum_{i=1}^{M} |v_i|^p \right)^{1/p} = \frac{1}{\sqrt{2}} \left(1 + \frac{M-1}{M^{p/2}} \right)^{1/p}, \tag{1.2}$$

which is of the order of 1. Set $\alpha = 1/2$ and $\beta = 2$, and thus $|J_{\alpha,\beta}| = M - 1$, implying, at least intuitively, that a random choice of coordinates will not 'collapse' the ℓ_2 norm. But will the same be true for the ℓ_p norm for $p > 2$?

For a typical subset $I \subset \{1, ..., M\}$, say of cardinality $N \ll M$,

$$\left(\frac{M}{N} \right)^{1/p} \left(\sum_{i \in I} |v_i|^p \right)^{1/p} \sim \left(\frac{1}{M} \right)^{1/2 - 1/p} \ll 1,$$

and the p-norm of $P_I v$ is much smaller than what we would like it to be.

The reason for this is the relative to the p-norm, v is very peaky – the contribution to the norm comes from a single coordinate, and all the other coordinates are negligible. In such a situation, sampling is useless. Upon reflection, the correct p-analog of the set $J_{\alpha,\beta}$ is

$$\left\{ i : \|v\|_p \frac{\alpha}{M^{1/p}} \leq |v_i| \leq \|v\|_p \frac{\beta}{M^{1/p}} \right\}, \tag{1.3}$$

and for our vector v and constant values of α and β, that set contains just a single coordinate.

Having large level sets like 1.1 or 1.3 has strong ties with the so-called *small-ball property* and in turn will help to guarantee that the norm of a sampled object does not 'collapse'. In contrast, ensuring that the sampled object is not too big – for example, in this case, that $\sqrt{M/|I|}(\sum_{i\in I}|v_i|^2)^{1/2}$ is not significantly larger than $\|v\|_2$ – is based on a totally different property: a tail estimate, captured in this case by the cardinality of the sets

$$\left\{ i : |v_i| \geq t\frac{\|v\|_2}{\sqrt{M}} \right\} \tag{1.4}$$

for $t \geq 1$.

It turns out that obtaining good tail estimates, and thus ensuring that the sampled object is not 'too large', requires rather restrictive assumptions. In contrast, the small-ball estimate which guarantees that the sampled object is not 'too small' is almost universally true and requires minimal assumptions.

In the next few chapters we will present the technical machinery needed for the study of sampling of a single object like a vector or a function. We will then show how to tackle uniform estimates that hold for a collection of objects (e.g., a set of vectors or a class of functions). But before we dive into technicalities, let us show why understanding the way sampling preserves certain norms takes centre-stage in learning theory.

1.1 A function class

In a learning problem one is given a class of functions F, defined on a probability space (Ω, μ), but very little is known about the measure μ, and the learner usually does not have access to it. For example, if X is distributed according to μ, the learner does not have access to $L_2(\mu)$ distance between $f_1, f_2 \in F$ that is to

$$\|f_1 - f_2\|^2_{L_2(\mu)} = \int_\Omega |f_1 - f_2|^2(x)d\mu(x) = \mathbb{E}|f_1(X) - f_2(X)|^2.$$

Let $\Omega = \mathbb{R}^n$ and set μ to be some probability measure on $\mathbb{R}^n$. For $T \subset \mathbb{R}^n$ we define the class of linear functionals associated with T

$$F_T = \{\langle t, \cdot \rangle : t \in T\}.$$

Now, every $t \in T$ has a two roles: a vector in $\mathbb{R}^n$, and a linear functional. Clearly, for $u, v \in T$ one may easily compute various distances between u and v, for example, the ℓ_p^n distances,

$$\|u - v\|_p = \|\sum_{i=1}^n (u_i - v_i)e_i\|_p = \left(\sum_{i=1}^n |u_i - v_i|^p\right)^{1/p},$$

where $(e_i)_{i=1}^n$ is the standard basis in $\mathbb{R}^n$. However, the $L_p(\mu)$ distance between u and v, via their identification as a linear functionals is a completely different story. The measure μ is not known, so computing

$$\|u - v\|_{L_p}^p = \int_{\mathbb{R}^n} |\langle u - v, x\rangle|^p d\mu(x) = \mathbb{E}|\langle u - v, X\rangle|^p$$

is impossible, and even worse, there are no guarantees that this distance is finite, because the integral may diverge.

It is important to keep in mind that there is no reason why $\|u - v\|_p$ should have anything to do with $\|u - v\|_{L_p}$ – these are completely different objects. For reasons that will be clarified immediately, the type of results we will obtain involve estimates on L_p distances. If one is interested in bounds on $\|u - v\|_p$ one has to assume more on the measure μ – allowing to relate $\|u-v\|_{L_p}$ with $\|u-v\|_p$. An assumption we will encounter frequently is that the measure μ is *isotropic*.

Definition 1.1.1 *The measure μ on $\mathbb{R}^n$ is isotropic if it is symmetric, in the sense that for every measurable set A, $\mu(A) = \mu(-A)$, and if for every $u \in \mathbb{R}^n$,*

$$\|u\|_{L_2(\mu)}^2 = \mathbb{E}\langle u, X\rangle^2 = \|u\|_2^2,$$

where X is distributed according to μ.

Thus, the L_2 norm endowed on linear functionals in $\mathbb{R}^n$ coincides with $\|\ \|_2$, and although we don't know μ, isotropicity implies a certain covariance structure, namely that $\mathbb{E}\langle X, e_i\rangle\langle X, e_j\rangle = \delta_{ij}$.

Assume that one is given is a random sample $X_1, ..., X_N$, that is, N points in Ω, selected independently according to the (unknown) measure μ. Given a class of functions F, One has access to the values of every function in F on each one of the sample points, and the goal is to identify, or at least approximate, the $L_2(\mu)$ distances between any $f, h \in F$, and to do so with high probability. The obvious candidate is

$$\Phi\left((X_i)_{i=1}^N, f, h\right) = \frac{1}{N}\sum_{i=1}^N (f - h)^2(X_i), \tag{1.5}$$

and Φ is the square of the empirical L_2 distance between f and h. The hope is that with high probability, for most of the pairs in F, if not all,

$$A\|f-h\|_{L_2}^2 \leq \frac{1}{N}\sum_{i=1}^{N}(f-h)^2(X_i) \leq B\|f-h\|_{L_2}^2$$

and in fact, one would like the constants A and B to be as close to 1 as possible.

Although (1.5) is the obvious candidate, one should keep an open mind and consider other alternatives. In fact, it turns out that (1.5) is not optimal (see Section **??** for more details).

With this choice in mind, a natural question that comes to mind is

Question 1.1.2 *What is the best possible choice of a function $\Phi : \Omega^N \times F \times F \to \mathbb{R}_+$, and constants $0 < \delta_N < 1$, $A_N, B_N, r_N, r'_N > 0$ for which the following holds: with probability at least $1-\delta_N$, if $h, f \in F$ and $\|f-h\|_{L_2} \geq r_N$ then*

$$\Phi\left((X_i)_{i=1}^N, f, h\right) \geq A_N\|f-h\|_{L_2}^2, \tag{1.6}$$

and if $h, f \in F$ and $\|f-h\|_{L_2} \geq r'_N$ then

$$\Phi\left((X_i)_{i=1}^N, f, h\right) \leq B_N\|f-h\|_{L_2}^2? \tag{1.7}$$

What is the connection between the structure of F, the measure μ and the choice of Φ, and the constants δ_N, A_N, B_N r_N and r'_N?

Alternatively, one may formulate the same question not for distances between every pair of functions in F, but rather for the norm of each $f \in F$, namely, if $\|f\|_{L_2} \geq r_N$ then

$$\Phi\left((X_i)_{i=1}^N, f, 0\right) \geq A_N\|f\|_{L_2}^2, \tag{1.8}$$

and if $h, f \in F$ and $\|f-h\|_{L_2} \geq r'_N$

$$\Phi\left((X_i)_{i=1}^N, f, 0\right) \leq B_N\|f\|_{L_2(\mu)}^2. \tag{1.9}$$

Remark 1.1.3 *The reason Question 1.1.2 is split into two components – a lower estimate on the $L_2(\mu)$ distances and an upper one is not coincidental. There is a fundamental difference between the two questions that lead to very different answers. We will show that the lower estimate* (1.6) *is in some sense universal, while* (1.7) *is a far more restrictive condition that requires stronger assumptions.*

As it happens, Question 1.1.2 is only a step towards understanding learning problems (which we have not defined yet). Having said that, it is still a formidable question in its own right. To put it is some context let us give three examples of problems that may be resolved using a solution to Question 1.1.2.

Almost isometric Embedding of a finite subset of $\mathbb{R}^n$

Let $T \subset (\mathbb{R}^n, \| \ \|_2)$ be a finite set. We would like 'reduce the dimension' of T, while preserving all of its metric structure: i.e., to map T to $\mathbb{R}^k$ for k that is, hopefully, significantly smaller than n, in a way that (almost) preserves the Euclidean distances between the points in T. Thus, the goal is to find a mapping $\psi : T \to \mathbb{R}^k$ that satisfies that for every $u, v \in T$, and $\varepsilon > 0$ as small as possible,

$$1 - \varepsilon \leq \frac{\|\psi(u) - \psi(v)\|_2}{\|u - v\|_2} \leq 1 + \varepsilon \tag{1.10}$$

where $\| \ \|_2$ denotes the Euclidean norm in both $\mathbb{R}^n$ and $\mathbb{R}^k$.

This problem has been studied extensively since the mid-80's, when Johnson and Lidenstrauss proved their celebrated lemma. They showed that with high probability, a correctly normalized random projection[2] onto a k-dimensional subspace of $\mathbb{R}^n$ for $k = c\varepsilon^{-2} \log |T|$ satisfies (1.10).

Remark 1.1.4 *It should be noted that despite its popularity, the Johnson-Lindenstrauss Lemma was just that, a component in the proof of a different result, on extending a function between a finite subset of a metric space X and ℓ_2, to the entire space X, without distorting the Lipschitz constant by much.*

There has been significant progress in the study of linear operators that satisfy (1.10) over the last 30 years. One class of such operators that is of particular interest in our context consists of random matrices with independent rows.

Let X be an isotropic random vector in $\mathbb{R}^n$ (i.e., for every $u \in \mathbb{R}^n$, $\mathbb{E}\langle X, u\rangle^2 = \|u\|_2^2$). Let $X_1, ..., X_N$ be independent copies of X, and set

$$\Gamma = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \langle X_i, \cdot \rangle e_i$$

[2]The notion of randomness Johnson and Lindenstrauss used was relative to the Haar measure on the appropriate Grassmann manifold

be the matrix whose rows are $X_1, ..., X_N$.

Observe that on average, Γ preserves the Euclidean norm of $u \in \mathbb{R}^n$:

$$\mathbb{E}\|\Gamma u\|_2^2 = \mathbb{E}\frac{1}{N}\sum_{i=1}^{N}\big\langle X_i, u\big\rangle^2 = \|u\|_2^2.$$

Of course, having a well-behaved mean does not imply that $\|\Gamma u\|_2^2$ is close to that mean with high probability, nor that uniform control is possible on a large collection of points.

Now, consider the class of linear functionals $F_T = \{\langle t, \cdot\rangle : t \in T\}$. Note that by selecting Φ as in (1.5), Question 1.1.2 implies that for every $u, v \in T$,

$$\Phi\left((X_i)_{i=1}^N, f_v, f_v\right) = \frac{1}{N}\sum_{i=1}^{N}(f_u - f_v)^2(X_i) = \frac{1}{N}\sum_{i=1}^{N}\big\langle u-v, X_i\big\rangle^2 = \|\Gamma(u-v)\|_2^2,$$

and $\|f_u - f_v\|_{L_2}^2 = \mathbb{E}\big\langle u - v, X\big\rangle^2 = \|u - v\|_2^2$. Thus, (1.10) follows from a positive answer to Question 1.1.2 for the right value of N that suffices to ensure that $1 - \delta_N > 0$, and with the choices of $A_N = 1 - \varepsilon$ and $B_N = 1 + \varepsilon$ and $r_N = r'_N = 0$.

The reason why $N = c\varepsilon^{-2}\log|T|$ is the right choice may still appear mysterious, but after we develop the necessary machinery, we will show the the logarithm of the cardinality of a set is a actually a rather crude measure of the set's complexity. We will then suggest much sharper alternatives. We will also explain where the factor of ε^{-2} comes from.

Extremal singular values of a random matrix

The spectral theory of random matrices has attracted considerable attention in recent years. One well studied question has to do with the largest and smallest singular values of a random matrix Γ, and those have a very simple geometric description according to the way Γ acts on the Euclidean unit sphere S^{n-1}:

$$\lambda_{\max} = \sup_{x \in S^{n-1}} \|\Gamma x\|_2 \quad \text{and} \quad \lambda_{\min} = \inf_{x \in S^{n-1}} \|\Gamma x\|_2.$$

In other words, the largest and smallest singular values of Γ are the outer radius and the inner radius, respectively, of the ellipsoid ΓB_2^n – the image of the Euclidean unit ball B_2^n.

Let us consider once again the Γ mentioned earlier: a matrix whose rows are independent copies of an isotropic random vector in $\mathbb{R}^n$. If follows that

$$\lambda_{\max}^2 = \sup_{x \in S^{n-1}} \frac{1}{N} \sum_{i=1}^N \langle X_i, x \rangle^2, \quad \text{and}$$
$$\lambda_{\min}^2 = \inf_{x \in S^{n-1}} \frac{1}{N} \sum_{i=1}^N \langle X_i, x \rangle^2,$$

corresponding to (1.9) and (1.8), respectively.

The fact that we chose to separate the upper estimate from the lower one will prove to be significant (see Section **??**). For example, we will show that under minimal assumptions on X, and with high probability,

$$\lambda_{\min} \geq 1 - c\sqrt{\frac{n}{N}}.$$

On the other hand,

$$\lambda_{\max} \leq 1 + c\sqrt{\frac{n}{N}}$$

is true, but only under more restrictive conditions. Moreover, these estimate hide what is the natural complexity parameter associated with the Euclidean unit sphere: $\sqrt{n}$. The 'error term' $\sqrt{n/N}$ happens to be the ratio between the 'complexity' of the indexing set – in this case, the unit sphere – and the square root of the cardinality of the given sample. This will prove to be a general phenomenon. Why $\sqrt{n}$ captures the complexity of sphere has to be explained, and, obviously, we will have to develop the right tools that will help us identify the right complexity parameters of more general classes of functions and their roles in 'error terms'.

Simple exact recovery

Let $T \subset \mathbb{R}^n$ and assume that some $t_0 \in T$ is selected but not revealed to us. Our goal is to identify t_0, or if that is impossible, to approximate it with respect to the $\| \ \|_2$ norm. The information we are provided with is a set of linear measurements, $(\langle X_i, t_0 \rangle)_{i=1}^N$, with $X_1, ..., X_N$ selected independently, according to an underlying measure μ.

Given that information, an obvious guess is to select any $t \in T$ that agrees with the given measurements; that is, take any $t \in T$ for which $\langle t, X_i \rangle = \langle t_0, X_i \rangle$ for every $1 \leq i \leq N$.

Now, assume that Question 1.1.2 may be answered for this problem; specifically, that with probability at least $1-\delta$, if $\|u-v\|_{L_2}^2 = \mathbb{E}\langle u-v, X\rangle^2 \geq r_N$ then

$$\frac{1}{N}\sum_{i=1}^{N}\langle u-v, X_i\rangle^2 \geq A_N\|u-v\|_{L_2}^2$$

for some $A_N > 0$. On this event, if $\langle t, X_i\rangle = \langle t_0, X_i\rangle$ for $1 \leq i \leq$, then $\|t-t_0\|_{L_2}^2 \leq r_N$, and if μ happens to be an isotropic measure, then

$$\|t-t_0\|_2^2 = \|t-t_0\|_{L_2}^2 \leq r_N.$$

Moreover, if $r_N = 0$ then no other point in T agrees with the t_0 on $X_1, ..., X_N$; the system of equations

$$\langle t, X_i\rangle = \langle t_0, X_i\rangle \quad \text{for every } 1 \leq i \leq N$$

has a unique solution in T, and that solution is t_0.

This observation has different names in different fields. In a naive version of *sparse recovery*, the set T consists of all the vectors in $\mathbb{R}^n$ that are s-sparse; that is, supported on at most s coordinates relative to the standard basis in $\mathbb{R}^n$. In learning theory, this is an example of a *realizable learning problem*, or a noise-free problem when the class consists of linear functionals in $\mathbb{R}^n$. And, in *Asymptotic Geometric Analysis*, for T that is convex and centrally symmetric, the argument leads to a bound on the so called random Gelfand widths of T, i.e., on the Euclidean diameter of $\ker(\Gamma) \cap T$: if $\langle X_i, t\rangle = 0$ for $i = 1, ..., N$, then $\|t-0\|_2^2 = \|t-0\|_{L_2}^2 \leq r_N$.

All these examples will be explored in great detail later, but for now, their role is convince the reader that Question 1.1.2 is a nontrivial theoretical question that has far-reaching implications in modern areas of mathematics, statistics, computer science and engineering – even if we only consider the restricted setup of classes of linear functionals in $\mathbb{R}^n$ and sampling according to an isotropic measure. Because if is such a fundamental question, it should not be surprising that developing the machinery necessary for addressing it will require some effort.

1.2 A Learning problem

Let us now turn to the "main event" – the definition of a learning problem. The starting point is the same as in the previous section: a class of functions defined on a probability space (Ω, μ), where the measure μ is not known.

Let $\mathcal{Y}$ be a collection of *admissible targets*, consisting of the random variables from which the (unknown) target will be selected; obviously, we would like to keep that set as large as possible. We will return to assumptions on $\mathcal{Y}$ a little later.

Consider some (unknown) $Y \in \mathcal{Y}$, and the goal is to find some $f \in F$ that is as close to Y as possible. The notion of similarity is up to the learner, and is calibrated using a *loss function*. While there are more general notions of a loss function, here we will only consider the following:

Definition 1.2.1 *A loss is a function $\ell : \mathbb{R} \to \mathbb{R}_+$ that is even, convex, increasing in $\mathbb{R}_+$ and satisfies $\ell(0) = 0$.*

What is arguably the most important example of a loss function and the one we will focus on in what follows is the squared loss $\ell(t) = t^2$.

Given $y \in \mathbb{R}$ and $x \in \Omega$ the cost of predicting $f(x)$ instead of y is $\ell(f(x) - y)$, and for the squared loss, $(f(x) - y)^2$. Hence, the best function one can find in the class F is the minimizer in F of the *risk functional*,

$$f \to \mathbb{E}\ell\left(f(X) - Y\right) \equiv R(f),$$

with the expectation taken with respect to the joint distribution of X and Y. We will assume that the minimizer exists and is unique, which happens to be a rather minimal assumption, and we denote that empirical minimizer by f^*.

The obvious difficulty in identifying f^* is that both X and Y are not known, and therefore it is impossible to solve the risk minimization problem

$$\text{argmin}_{f \in F} \mathbb{E}\left(f(X) - Y\right)^2.$$

The difference between standard problems in Approximation Theory and the ones we are interested in is the information one has at one's disposal: rather than knowing Y and X, one has access only to a random sample $\mathcal{D} = (X_i, Y_i)_{i=1}^N$, selected according to the joint distribution of (X, Y). Using the random sample one is expected to produce some $\hat{f}$ and ensure that for most samples, $\hat{f}$ approximates f^*.

There are several notions of approximation we will consider:

(1) $\hat{f}$ is selected in F, and one would like $\hat{f}$ to be close to f^* in the $L_2(\mu)$ sense, i.e., with high probability,

$$\|\hat{f} - f^*\|_{L_2} = \mathbb{E}\left((\hat{f} - f^*)^2(X)|\mathcal{D}\right) \leq \mathcal{E}_e,$$

and the *estimation error* $\mathcal{E}_e$ should be as small as possible.

(2) $\hat{f}$ is selected in F and one would like the risk of $\hat{f}$ to be almost optimal, that is, with high probability,

$$R(\hat{f}) \leq \inf_{f \in F} R(f) + \mathcal{E}_p,$$

and the *prediction error* $\mathcal{E}_p$ should be as small as possible.

(3) One is allowed more freedom, and the restriction that $\hat{f} \in F$ is removed. Still, the goal is to select $\hat{f}$ whose risk is not much larger than the best in F; that is, with high probability,

$$R(\hat{f}) \leq \inf_{f \in F} R(f) + \mathcal{E}_{agg},$$

and the *aggregation error*[3] $\mathcal{E}_{agg}$ is as small as possible.

(4) It is possible to show that if the class F is 'too rich' in a sense that will be clarified later, the estimation error and prediction error will be too big to be of any use, regardless of the way $\hat{f}$ is selected. Instead of restricting the problems we consider to only small classes, it is possible to use *regularization methods*. The idea behind regularization is that some functions within F are preferred to others: each class member has a 'price-tag' attached to it and if two functions have a similar fit to the random data, preference is given to the function with the smaller price-tag. The hope is that with a well-chosen penalty one may find a procedure $\hat{f}$ that has a small estimation/prediction error, despite the fact that F is seemingly too large.

There are many obvious questions one can ask at this point, but the most fundamental one is this:

Question 1.2.2 *What determines the prediction error and the estimation error? Specifically, how do the estimation error and prediction error scale with sample size N, the probability estimate one is aiming for, the structure of F, the underlying measure μ and the class of admissible targets $\mathcal{Y}$? And, finally, what is the right choice of $\hat{f}$?*

The main goal of these notes is to address Question 1.2.2 and other questions of its kind.

[3]The name "aggregation error" is not standard, and we will explain its origins in Section **??**.

Before we present a formal definition of a learning problem, let us give an example of a prediction problem and of an estimation problem, both in $\mathbb{R}^n$ and with respect to the squared loss.

Let μ be the standard gaussian measure μ on $\mathbb{R}^n$: i.e., the measure whose density is proportional to $\exp(-\|t\|_2^2/2)$ (in what follows we will not use any of the special properties of the gaussian measure – other than it is isotropic).

Let $T = B_1^n = \{x : \sum_{i=1}^n |x_i| \leq 1\}$ be the unit ball of the normed space $\ell_1^n = (\mathbb{R}^n, \| \ \|_1)$ and set $F_T = \{\langle t, \cdot \rangle : t \in T\}$. Let $t_0 \in \mathbb{R}^n$ (not necessarily in B_1^n) and set $Y = \langle t_0, X \rangle + W$, for W is a centred random variable that has variance σ^2 and is independent of X. Because W and X are independent, W is mean-zero, and X is isotropic, it is evident that for every $t \in T$,

$$\mathbb{E}(Y - \langle t, X \rangle)^2 = \mathbb{E}\langle t_0 - t, X \rangle^2 + \sigma^2 = \|t_0 - t\|_2^2 + \sigma^2.$$

Hence, the minimizer in F_T of the risk is attained by $f^* = \langle t^*, \cdot \rangle$ for t^* that is closest to t_0 with respect to the Euclidean distance.

The data one is given is a random sample $(X_i, \langle X_i, t_0 \rangle + W_i)_{i=1}^N$ for $X_1, ..., X_N$ that are independent and distributed according to μ, and W_i that are independent copies of W and are also independent of $(X_i)_{i=1}^N$. The goal is to use that given sample to produce some $\hat{t} \in B_1^n$. With this choice, $\hat{f} = \langle \hat{t}, \cdot \rangle$

(1) the estimation error of $\hat{f} = \langle \hat{t}, \cdot \rangle$ is

$$\mathcal{E}_\varepsilon = \|\hat{f} - f^*\|_{L_2}^2 = \|\hat{t} - t^*\|_2^2, \quad \text{and}$$

(2) The prediction error of $\hat{f} = \langle \hat{t}, \cdot \rangle$ is

$$\mathcal{E}_p = R(\hat{f}) - R(f^*) = \|\hat{t} - t_0\|_2^2 - \|t^* - t_0\|_2^2.$$

With Question 1.2.2 in mind, how should $\hat{t}$ be selected? How is the fact that $T = B_1^n$ is reflected in the estimation error and prediction error? And how would the estimation and prediction error change for different X or Y – for example, if X happen to be more 'heavy tailed' than gaussian?

1.3 Some definitions

Let (Ω, μ) be a probability space; the probability measure μ is fixed, but not known, and let X be distributed according to μ. Let F be a class of real-valued functions defined on Ω, and set $\mathcal{Y}$ to be a class of admissible targets.

Definition 1.3.1 *A set of admissible targets is minimal if it contains all targets of the form* $\{f(X) + W : f \in F\}$*, where* W *is a centred gaussian random variable with variance at most* σ^2 *that is independent of* X*. With a slight abuse of notation we will denote a minimal set of targets by* $\mathcal{Y}_{\min}$*.*

The idea is that a reasonable class of admissible targets must at least contain what are arguably the most natural and simplest of all targets: realizable targets, i.e., of the form $Y = f_0(X)$ for some $f_0 \in F$, and additive shifts of realizable targets by independent gaussian noise, $Y = f_0(X) + W$.

Definition 1.3.2 *Given a sample size* N*, a learning procedure is a mapping* $\Phi_N : (\Omega \times \mathbb{R})^N \to F$*. In other words, it assigns each* $(x_i, y_i)_{i=1}^N$ *to some* $f \in F$*.*

Remark 1.3.3 *For the time being, we shall focus our attention to* learning *procedures – maps that take values within the given class. We will relax this restriction later, allowing the procedure to take values outside* F*.*

Next, let us define the estimation error and prediction error rate of a learning problem both with respect to the squared loss. The modifications needed for the definition of estimation error and prediction error relative to more general loss functions are obvious and at this point are omitted.

Definition 1.3.4 *Given a class* F*, a set of admissible targets* $\mathcal{Y}$ *and an integer* N*, a learning procedure* Φ_N *performs with estimation accuracy* $\mathcal{E}_e$ *and confidence parameter* δ *if for any* $Y \in \mathcal{Y}$*, with probability at least* $1 - \delta$*,*

$$\|\Phi_N\left((X_i, Y_i)_{i=1}^N\right) - f^*\|_{L_2} \leq \mathcal{E}_e \tag{1.11}$$

where f^* *denotes the minimizer in* F *of the true risk functional* $f \in \mathbb{E}(f(X) - Y)^2$ *and the probability is with respect to the* N *product of the joint distribution of* X *and* Y*.*

The procedure performs with prediction accuracy $\mathcal{E}_p$ *and confidence parameter* δ *if for any* $Y \in \mathcal{Y}$*, with probability at least* $1 - \delta$*,*

$$R\left(\Phi_N\left((X_i, Y_i)_{i=1}^N\right)\right) \leq \inf_{f \in F} R(f) + \mathcal{E}_p \tag{1.12}$$

Remark 1.3.5 *Note that both* $\mathcal{E}_e$ *and* $\mathcal{E}_p$ *depend on the cardinality of the given sample* N*.*

In other words, the performance of a procedure is measured by its success with regard to any admissible target $Y \in \mathcal{Y}$ and that success is measured via the accuracy/confidence tradeoff: the error a procedure may guarantee and the probability with which it can guarantee it.

The obvious way of deciding if a procedure is useful is by comparing the accuracy/confidence tradeoff it achieves with the benchmark performance of a hypothetical procedure: the theoretical limitations on the accuracy/confidence tradeoff. And to make this comparison more interesting, the hypothetical procedure only has to contend with a minimal set of admissible targets. This type of an error rate is often called the *minimax error rate*, though in some places its meaning is slightly different than the way we use it here.

Definition 1.3.6 *Given a class of functions F on a probability space (Ω, μ) and a set of admissible targets $\mathcal{Y}$, a procedure Φ_N performs with the γ-minimax accuracy for a confidence parameter δ if*

- *For every $Y \in \mathcal{Y}$, with probability at least $1-\delta$, (1.11) (resp. (1.12)) holds, with an estimation error $\mathcal{E}_e$ (resp. prediction error $\mathcal{E}_p$).*

- *If Ψ is another learning procedure, then there is some $Y \in \mathcal{Y}_{\min}$ for which the event $\|\Psi_N\left((X_i,Y_i)_{i=1}^N\right) - f^*\|_{L_2} \leq \gamma\mathcal{E}_e$ (resp. $R\left(\Phi_N\left((X_i,Y_i)_{i=1}^N\right)\right) \leq \inf_{f\in F} R(f) + \gamma\mathcal{E}_p$) holds probability smaller than $1-\delta$.*

Because it is rather optimistic to identify the best possible procedure, the parameter γ gives one some freedom, and we will consider a procedure to be optimal in the minimax sense if for a given degree of confidence the accuracy is performs with for any $Y \in \mathcal{Y}$ is proportional to the theoretical limitations relative to a minimal set of targets, and for γ that is an absolute constant – independent of N and the class F.

We will discuss the question of the optimality and minimax rates at length in Section **??**.

1.4 Empirical risk minimization

Given a class of functions F on (Ω, μ) and an unknown target $Y \in \mathcal{Y}$, there is a natural candidate for a learning procedure: given a sample $(X_i, Y_i)_{i=1}^N$, the choice is the function in F that best fits the sample, taking into account the loss function. In the case of the squared loss, that candidate is a minimizer

in F of the empirical risk

$$f \to \frac{1}{N}\sum_{i=1}^{N}\left(f(X_i)-Y_i\right)^2. \tag{1.13}$$

The mapping $\Phi : (\Omega \times \mathbb{R})^N \to F$ that selects

$$\hat{f} \in \mathrm{argmin}_{f\in F}\frac{1}{N}\sum_{i=1}^{N}\left(f(X_i)-Y_i\right)^2 \tag{1.14}$$

is called *Empirical Risk Minimization* (ERM).

ERM is one of the important learning procedures and has been studied extensively over the last 50 years in numerous manuscripts.

The analysis of ERM is based on the notion of the excess loss functional and the resulting excess risk.

Definition 1.4.1 *Let ℓ be a loss function and set f^* to be the minimizer in F of the risk $\mathbb{E}\ell(f(X)-Y)$. The excess loss functional assigns to each $f \in F$ the function*

$$\mathcal{L}_f(X,Y) = \ell(f(X)-Y) - \ell(f^*(X)-Y),$$

and the excess risk is

$$\mathbb{E}\mathcal{L}_f(X,Y) = \mathbb{E}\ell(f(X)-Y) - \mathbb{E}\ell(f^*(X)-Y) = R(f) - R(f^*).$$

Note that while the loss $\ell(f(X)-Y)$ can be computed on a given sample point (X,Y), the same is no longer true for the excess loss $\mathcal{L}_f(X,Y) = \ell(f(X)-Y)-\ell(f^*(X)-Y)$ – because the value $f^*(X)$ is not known. Thus, the excess loss functional may be used only as a theoretical object, and one cannot suggest learning procedures that are based on the loss. Having said that, the excess risk has two important features:

(1) For every $f \in F$, $\mathbb{E}\mathcal{L}_f \geq 0$, and if f^* is uniquely determined then $\mathbb{E}\mathcal{L}_f = 0$ only for $f = f^*$.

(2) The empirical minimizer of the loss coincides with the empirical minimizer of the excess loss. In other words,

$$\hat{f} \in \mathrm{argmin}\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_f(X_i,Y_i).$$

It follows that

$$\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{\hat{f}}(X_i,Y_i)\leq 0 \tag{1.15}$$

because f^* is a 'competitor' and $\mathcal{L}_{f^*}=0$.

Equation (1.15) presents an opportunity to analyze ERM: potentially one could 'map' the values of the empirical excess risk in F on a high-probability set of samples. For every fixed sample in that set, the fact that $\sum_{i=1}^{N}\mathcal{L}_f(X_i,Y_i)>0$ excludes f as a potential empirical minimizer. Hence, if there is some $r>0$ for which

$$\{f\in F:\|f-f^*\|_{L_2}\geq r\}\subset\left\{f\in F:\sum_{i=1}^{N}\mathcal{L}_f(X_i,Y_i)>0\right\}, \tag{1.16}$$

it implies that $\|\hat{f}-f^*\|_{L_2}\leq r$, establishing an estimate of $\mathcal{E}_e$. In a similar fashion, if

$$\{f\in F:R(f)-R(f^*)\geq r\}\subset\left\{f\in F:\sum_{i=1}^{N}\mathcal{L}_f(X_i,Y_i)>0\right\} \tag{1.17}$$

then $R(\hat{f})-R(f^*)\leq r$, leading to an estimate on $\mathcal{E}_p$.

Let us examine (1.16) for the squared loss. To obtain a useful estimate we have to show that with high probability, for any function that is sufficiently far away from (the unknown) f^*, the excess empirical risk is positive. We will try to show a little more: that with high probability, if $\|f-f^*\|_{L_2}\geq r$ then

$$\frac{1}{N}\sum_{i=1}^{N}(f(X_i)-Y_i)^2-\frac{1}{N}\sum_{i=1}^{N}(f^*(X_i)-Y_i)^2\geq\theta\|f-f^*\|_{L_2}^2, \tag{1.18}$$

which would imply (1.16).

Observe that

$$\begin{aligned}(f(X)-Y)^2-(f^*(X)-Y)^2&=(f(X)-f^*(X)+f^*(X)-Y)^2-(f^*(X)-Y)^2\\&=(f(X)-f^*(X))^2+2\,(f(X)-f^*(X))\,(f^*(X)-Y)\,.\end{aligned}$$

Set $\xi(X,Y)=f^*(X)-Y$ and for a sample $(X_i,Y_i)_{i=1}^N$, let

$$\mathcal{Q}_{f-f^*}=\frac{1}{N}\sum_{i=1}^{N}(f-f^*)^2(X_i), \tag{1.19}$$

and

$$\mathcal{M}_{f-f^*} = \frac{1}{N}\sum_{i=1}^{N} \xi_i (f-f^*)(X_i) - \mathbb{E}\xi(f-f^*)(X), \tag{1.20}$$

where $\xi_i = f^*(X_i) - Y_i$. Therefore,

$$\frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_f(X_i, Y_i) = \mathcal{Q}_{f-f^*} + \mathcal{M}_{f-f^*} + \mathbb{E}\xi(f-f^*)(X), \tag{1.21}$$

and we will call $\mathcal{Q}_{f-f^*}$ the *quadratic component* of the empirical excess risk and $\mathcal{M}_{f-f^*}$ the multiplier component. To prove (1.18) it suffices to show that with high probability, if $\|f-f^*\|_{L_2} \geq r$ then

$$\frac{1}{N}\sum_{i=1}^{N} (f-f^*)^2(X_i) \geq 2\theta \|f-f^*\|_{L_2}^2, \tag{1.22}$$

$$\left| \frac{1}{N}\sum_{i=1}^{N} \xi_i (f-f^*)(X_i) - \mathbb{E}\xi(f-f^*)(X) \right| \leq \theta \|f-f^*\|_{L_2}^2. \tag{1.23}$$

and for every $f \in F$, $\mathbb{E}\xi(f-f^*)(X) \geq 0$.

The lower bound on the quadratic component $\mathcal{Q}_{f-f^*}$ is of a similar nature to (1.6), and is indeed a question about sampling not shrinking the L_2 norm by more that a constant factor – uniformly in a class. In contrast, (1.23) has to do with an upper bound on the supremum of a certain empirical process. The estimates we require has to do with the oscillation of empirical means involving the *multipliers* $(\xi_i)_{i=1}^N$ around the true mean. Moreover, the estimate has to be uniform: on a large event the bound has to hold for every $f \in F$ that is sufficiently far from f^*.

Obtaining the required bounds for a single function is not trivial – let alone the uniform bounds we actually need. It will take some effort to develop the necessary technical machinery, and we will do that in the next few chapters. The connection these questions have with the geometric structure of F will also become apparent.

Finally, the issue of the sign of $\mathbb{E}\xi(f-f^*)(X)$ happens to be relatively simple, as long as F is closed and convex. In that case one may show that for every $f \in F$, $\mathbb{E}\xi(f-f^*)(X) \geq 0$. For example, for the squared loss $\mathbb{E}(f^*(X)-Y)(f-f^*)(X) \geq 0$ for every $f \in F$ is the characterization of f^* as the unique minimizer of the functional $f \to \mathbb{E}(f(X)-Y)^2$; i.e., that f^* is the nearest point to Y in the convex class F with respect to the L_2 norm. A similar fact is true for any (convex) loss function. Whether $\mathbb{E}\xi(f-f^*)(X) \geq 0$ for every $f \in F$ becomes significantly more difficult when F is not convex, and we will address that question in Chapter **??**.

Chapter 2

Basic function spaces

We will be interested in various properties of normed spaces consisting of functions defined on a probability space (Ω, μ).

The simplest norms we will explore are the L_p norms, for $1 \leq p \leq \infty$.

Definition 2.0.1 *Let (Ω, μ) be a probability space and let X be distributed according to μ. For $1 \leq p < \infty$ and a measurable function $f : \Omega \to \mathbb{R}$ set*

$$\|f\|_{L_p(\mu)} = (\mathbb{E}|f(X)|^p)^{1/p} = \left(\int_\Omega |f(x)|^p d\mu(x)\right)^{1/p},$$

and

$$\|f\|_{L_\infty(\mu)} = \inf\{a : \mu(|f| > a) = 0\}$$

is the essential supremum of f. Thus, up to a set of measure 0, $|f| \leq \|f\|_{L_\infty}$.

In what follows we will often omit the measure μ and denote the $L_p(\mu)$ norm by $\|f\|_{L_p}$.

It is standard to verify that for $1 \leq p \leq \infty$, the $\| \ \|_{L_p}$ are indeed norms, and that because (Ω, μ) is a probability space, the L_p norms for a hierarchy

$$\|f\|_{L_p} \leq \|f\|_{L_q} \quad \text{if} \quad 1 \leq p \leq q \leq \infty. \tag{2.1}$$

$L_p(\mu)$ is the space of measurable functions with a finite L_p norm, and again, we will omit the dependence on the measure and write L_p instead of $L_p(\mu)$. It follows from (2.1) that if $1 \leq p \leq q \leq \infty$, $L_q \subset L_p$.

It is straightforward to verify that for $1 \leq p < \infty$,

$$\|f\|_{L_p}^p = p \int_0^\infty t^{p-1} Pr(|f| > t) dt. \tag{2.2}$$

and we will use (2.2) extensively in what follows.

One of the questions we would like to explore has to do with the preservation of L_p norms. If $f \in L_p$, X is distributed according to μ and $X_1, ..., X_N$ are independent copies of X, then sampling preserves the L_p norm if

$$\left(\frac{1}{N} \sum_{i=1}^{N} |f(X_i)|^p \right)^{1/p} \sim \|f\|_{L_p(\mu)}, \tag{2.3}$$

with the obvious one-sided versions.

In other words, random sampling yields points in Ω, selected independently according to μ. This defines an alternative (random) measure on Ω, a measure that assigns equal weight of $1/N$ to each one of the $X_1, ..., X_N$. That *empirical measure* is usually denoted by μ_N and is defined by

$$\mu_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{X_i}, \tag{2.4}$$

and δ_{X_i} are point-masses: $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise.

It is important to keep in mind that μ_N is random – every realization of $X_1, ..., X_N$ leads to a (potentially) different measure (2.4). Informally put, the L_p norm of f is preserved if, with high probability relative to $X_1, ..., X_N$ (i.e., with respect to the product measure $\mu^{\otimes N}$), $\|f\|_{L_p(\mu_N)} \sim \|f\|_{L_p(\mu)}$.

2.1 Two examples

The space $\mathbb{R}^M$ will be featured in two natural ways in our discussion: sometimes it will serve as a space of functions on the probability space $\{1, ..., M\}$ and sometimes it will be the probability space, endowed with some natural measure; in that case we will be interested in functions defined on $\mathbb{R}^M$.

$\mathbb{R}^M$ as a space of functions

The vector space $\mathbb{R}^M$ can be identified as a space of functions: each $v \in \mathbb{R}^M$ corresponds to a function defined on $\Omega = \{1, ..., M\}$ by setting $f(i) = v_i$ for $1 \leq i \leq M$. Let μ be the uniform probability measure on $\{1, ..., M\}$, i.e., the measure that assigns the weight of $1/M$ to each $i \in \Omega$. Therefore,

$$\|f\|_{L_p^M} = \left(\frac{1}{M} \sum_{i=1}^{M} |v_i|^p \right)^{1/p} = \left(\int_\Omega |f(i)|^p d\mu(i) \right)^{1/p}.$$

It follows that

$$\|v\|_p = \left(\sum_{i=1}^M |v_i|^p\right)^{1/p} = M^{1/p}\left(\frac{1}{M}\sum_{i=1}^M |v|^p(i)\right)^{1/p} = M^{1/p}\|v\|_{L_p^M}, \quad (2.5)$$

and $\| \ \|_{L_p^M}$ is just a re-scaling of $\| \ \|_p$. We will denote by L_p^M to be $\mathbb{R}^M$ endowed with the L_p^M norm, and the unit ball satisfies

$$B_p^M = \{v \in \mathbb{R}^M : \|v\|_p \leq 1\} = M^{-1/p}B(L_p^m) = M^{-1/p}\{v : \|v\|_{L_p^M} \leq 1\}. \quad (2.6)$$

Also, for $I \subset \{1, ..., M\}$ we set

$$\|v\|_{L_p^I} = \left(\frac{1}{|I|}\sum_{i\in I} |v_i|^p\right)^{1/p}.$$

Remark 2.1.1 *When we mentioned the benchmark behaviour of sampling, we were hoping to see that for a random selection of $I \subset \{1, ..., M\}$, $\sum_{i\in I} |v_i|^p \sim (|I|/M)\|v\|_p^p$. In other words, that*

$$\left(\frac{1}{M}\sum_{i=1}^M |v_i|^p\right)^{1/p} = \|v\|_{L_p^M} \sim \|v\|_{L_p^I} = \left(\frac{1}{|I|}\sum_{i=1}^{|I|} |v_i|^p\right)^{1/p}. \quad (2.7)$$

This is an instance of the benchmark behaviour of sampling: that with high probability

$$\|f\|_{L_p(\mu_N)} = \left(\frac{1}{N}\sum_{i=1}^N |f(X_i)|^p\right)^{1/p} \sim \left(\int_\Omega |f(x)|^p d\mu(x)\right)^{1/p} = \|f\|_{L_p(\mu)}.$$

$\mathbb{R}^M$ as a probability space

Let us turn to the case $\Omega = \mathbb{R}^M$. Let μ be a probability measure on $\mathbb{R}^M$ and let X be a random vector, distributed according to μ. Hence, $(\mathbb{R}^M, \mu)$ is a probability space, and we may consider the L_p space of functions on $(\mathbb{R}^M, \mu)$. We will focus on a rather particular choice on functions: linear functionals on $\mathbb{R}^M$.

There is a natural correspondence between $\mathbb{R}^M$ and those linear functionals: each $v \in \mathbb{R}^M$ defines a linear functional by

$$f_v(x) = \langle v, x\rangle = \sum_{i=1}^M v_i x_i.$$

Having said that, it is far from obvious that f_v actually belongs to L_p. For that to happen, it is necessary that

$$\|f_v\|_{L_p} = \big(\mathbb{E}|\langle X, v\rangle|^p\big)^{1/p} = \left(\int_\Omega \left|\sum_{i=1}^M v_i x_i\right|^p d\mu(x)\right)^{1/p} < \infty.$$

Moreover, there is no reason to expect any connection whatsoever between $\|v\|_p$ and $\|f_v\|_{L_p}$, even when $f_v \in L_p$ for every $v \in \mathbb{R}^M$. If that happens, both $\|\ \|_p$ and $\|\ \|_{L_p}$ define norms on $\mathbb{R}^M$, but as will show below, these norms can be totally different.

In what follows we will abuse notation and write $\|v\|_{L_p}$ instead of $\|f_v\|_{L_p}$.

Recall that if μ is isotropic then for every $v \in \mathbb{R}^M$, $\|f_v\|_{L_2}^2 = \mathbb{E}|\langle v, X\rangle|^2 = \|v\|_2^2$. Hence, not only does each $f_v \in L_2(\mu)$, but the mapping $v \to f_v$ is an isometric embedding of $(\mathbb{R}^M, \|\ \|_2)$ in $L_2(\mu)$. Of course, this does not mean that $f_v \in L_p(\mu)$ for $p > 2$.

If every f_v belongs to $L_2(\mu)$ then the inner product in L_2 endows an alternative inner product of $\mathbb{R}^M$. Indeed, set $X = (x_1, ..., x_M)$ be a random vector on $\mathbb{R}^M$, let $e_1, ..., e_M$ be the standard basis in $\mathbb{R}^M$ and define

$$[e_i, e_j] = \int_{\mathbb{R}^M} f_{e_i}(x) f_{e_j}(x) d\mu(x) = \int_{\mathbb{R}^M} \langle e_i, x\rangle \cdot \langle e_j, x\rangle d\mu(x) = \int_{\mathbb{R}^M} x_i x_j d\mu(x).$$

It follows that the matrix $([e_i, e_j])_{i,j}$ is the covariance matrix of the random vector X and

$$[v, u] = \sum_{i,j} v_i u_j [e_i, e_j] = \big\langle \mathrm{Cov}(X)v, u\big\rangle$$

is an inner product on $\mathbb{R}^M$ and the unit ball of the norm this inner product endows on $\mathbb{R}^M$ is the ellipsoid $\{v \in \mathbb{R}^M : \big\langle \mathrm{Cov}(X)v, v\big\rangle \le 1\}$. In the special case of an isotropic measure, $\mathrm{Cov}(X) = Id_M$ is the identity matrix, $\big\langle v, u\big\rangle = [v, u]$ and the ellipsoid is the standard Euclidean ball. However, in general, $\|\ \|_{L_2(\mu)}$ and $\|\ \|_2$ are different norms.

The difference is even more obvious when examining the L_p norm of a standard gaussian vector G, whose density is $c\exp(-\|t\|_2^2/2)$ for a normalizing constant c. Clearly, G has the same distribution as $(g_1, ..., g_M)$ where the g_i's are independent, standard gaussian random variables. Observe that for every $v \in \mathbb{R}^M$,

$$\mathbb{E}\big\langle G, v\big\rangle^2 = \mathbb{E}\sum_{i,j} g_i g_j v_i v_j = \sum_{i=1}^M v_i^2 \mathbb{E} g_i^2 = \|v\|_2^2,$$

implying that G is an isotropic random vector, and in particular, the L_2 norm it endows on $\mathbb{R}^M$ coincides with the standard Euclidean norm $\| \ \|_2$.

We will show in what follows that for every $v \in \mathbb{R}^M$ and any $1 \leq p < \infty$,

$$\left(\mathbb{E}|\langle G, v\rangle|^p\right)^{1/p} = \left(\mathbb{E}\left|\sum_{i=1}^M g_i v_i\right|^p\right)^{1/p} \sim \sqrt{p}\|v\|_2.$$

Therefore, the corresponding L_p unit ball is equivalent to $\sqrt{p}B_2^M$, and is a very different set from B_p^M, the unit ball of $(\mathbb{R}^M, \| \ \|_p)$ or from $B(L_p^M) = M^{-1/p}B_p^M$, the L_p ball from (2.6).

Let us see what the benchmark behaviour of sampling one would like to have in this case. Let $G_1, ..., G_N$ be independent copies of G, and observe that sampling preserves the L_p norm if with high probability,

$$\left(\mathbb{E}|\langle G, v\rangle|^p\right)^{1/p} \sim \left(\frac{1}{N}\sum_{i=1}^M |\langle G_i, v\rangle|^p\right)^{1/p},$$

which is a completely different question than (2.7), although both setups deal with $\mathbb{R}^M$.

2.2 Weak L_p spaces

Let f be a measurable function defined on the probability space (Ω, μ). Denote

$$\|f\|_{L_{p,\infty}(\mu)} = \inf\left\{A > 0 : \sup_{t\in\mathbb{R}^+} t^p Pr\left(|f| > tA\right) \leq 1\right\}, \tag{2.8}$$

and at times we will omit the underlying measure μ and denote the norm by $\|f\|_{L_{p,\infty}}$.

It should stressed that $\|f\|_{L_{p,\infty}}$ is actually not a norm – it does not satisfy the triangle inequality. Despite that, we will keep referring to it as the weak L_p norm. The weak L_p space consists of all the measurable functions that satisfy $\|f\|_{L_{p,\infty}} < \infty$, and it is denoted by $L_{p,\infty}$.

There is a true difference between the L_p norm and the weak L_p norm as can be seen from the following two basic facts. As we noted earlier,

$$\|f\|_{L_p}^p = p\int_0^\infty t^{p-1} Pr(|f| > t)dt. \tag{2.9}$$

Also, by Chebyshev's inequality

$$Pr(|f| > t\|f\|_{L_p}) \leq \frac{1}{t^p};$$

and in particular,

$$\sup_{t\in\mathbb{R}^+} t^p Pr(|f| > t\|f\|_{L_p}) \leq 1.$$

These facts illustrate the gap between an L_p space and a weak L_p space: if $f \in L_{p,\infty}$ then its tail $\{|f| > t\}$ decays faster than $\sim 1/t^p$, but that does not ensure integrability as in (2.9); hence, for any probability measure μ,

$$\|f\|_{L_{p,\infty}} \leq \|f\|_{L_p}.$$

In contrast, it is straightforward to construct examples of functions on (Ω, μ) that belong to $L_{p,\infty}$ but not to L_p. As such, the weak L_p norm is as its name suggests: weaker than the L_p norm, though as the next lemma shows, it is only slightly weaker.

Lemma 2.2.1 *If $1 \leq p < q < \infty$, then*

$$\|f\|_{L_p} \leq \left(1 + \frac{p}{q-p}\right)^{1/p} \|f\|_{L_{q,\infty}}.$$

Proof. Note that $\sup_{t\in\mathbb{R}^+} t^q Pr(|f| > tA) \leq 1$ if and only if $\sup_{t\in\mathbb{R}^+} t^q Pr(|f| > t) \leq A^q$. Hence, by (2.9),

$$\begin{aligned}
\|f\|_{L_p}^p =& p\int_0^\infty t^{p-1} Pr(|f| \geq t)dt \\
\leq& p\int_0^{\|f\|_{L_{q,\infty}}} t^{p-1} + p\int_{\|f\|_{L_{q,\infty}}}^\infty t^{p-1-q} \cdot t^q Pr_\mu(|f| \geq t)dt \\
\leq& \|f\|_{L_{q,\infty}}^p \left(1 + p\int_0^\infty t^{p-1-q} dt\right) = \|f\|_{L_{q,\infty}}^p \left(1 + \frac{p}{q-p}\right).
\end{aligned}$$

■

Example. Consider the weak counterpart of the L_p space L_p^M, that is, the weak L_p space that consists of functions defined on $\Omega = \{1, ..., M\}$ endowed with the uniform probability measure. Using the natural identification between functions on Ω and vectors in $\mathbb{R}^M$ it is evident that for $v \in \mathbb{R}^M$ and $t > 0$, $Pr(|v| > t) = M^{-1}|\{i : |v_i| > t\}|$. By the definition of the weak L_p norm,

$$M^{-1}|\{i : |v_i| \geq t\}| \leq \left(\frac{\|v\|_{L_{p,\infty}^M}}{t}\right)^p. \tag{2.10}$$

This observation may be formulated using a very important notion: the monotone nonincreasing rearrangement of the coordinates of a vector.

Definition 2.2.2 *For every $v \in \mathbb{R}^M$, let $(v_i^*)_{i=1}^M$ denote the monotone nonincreasing rearrangement of $(|v_i|)_{i=1}^M$; that is, $(v_i^*)_{i=1}^M$ is a permutation of $(|v_i|)_{i=1}^M$ for which $v_1^* \geq v_2^* \geq ... \geq 0$.*

Equation (2.10) implies that for every $1 \leq k \leq M$,

$$v_k^* \leq \left(\frac{M}{k}\right)^{1/p} \|v\|_{L_{p,\infty}^M} = \frac{\|v\|_p}{k^{1/p}},$$

an observation that will appear frequently in what follows.

2.3 Orlicz norms

Let us turn to a very important family of function spaces – the so called *Orlicz spaces*, defined on the probability space (Ω, μ).

Definition 2.3.1 *Let $\Phi \not\equiv 0$ be an even, convex function that is increasing in $\mathbb{R}^+$ and satisfies $\Phi(0) = 0$. For $f : \Omega \to \mathbb{R}$, set*

$$\|f\|_\Phi = \inf \left\{C > 0 : \mathbb{E}\Phi(f/C) \leq 1\right\},$$

and denote by L_Φ the set of all (measurable) functions that satisfy $\|f\|_\Phi < \infty$.

As an example, let $\Phi(t) = |t|^p$ for $1 \leq p < \infty$ and observe that $\mathbb{E}(|f|^p/C^p) \leq 1$ when $\|f\|_{L_p} \leq C$; therefore, $\|f\|_\Phi = \|f\|_{L_p}$ and $L_\Phi = L_p$.

Lemma 2.3.2 *Let Φ be as in Definition 2.3.1. Then $\| \ \|_\Phi$ is a norm on the space L_Φ.*

Proof. Clearly, $\| \ \|_\Phi$ is positive homogeneous and for every $f \in L_\Phi$, $\|f\|_\Phi \geq 0$. Observe that if $\|f\|_\Phi = 0$ then for every $C > 0$ $\mathbb{E}\Phi(f/C) \leq 1$. Since Φ is even then by Jensen's inequality, for every $C > 0$,

$$\Phi(\mathbb{E}|f|/C) \leq \mathbb{E}\Phi(|f|/C) = \mathbb{E}\Phi(f/C) \leq 1.$$

On the other hand, it is straightforward to verify that $\lim_{t\to\infty} \Phi(t) = \infty$. Therefore, if $\mathbb{E}|f| \neq 0$ and C is small enough, then $\Phi(\mathbb{E}|f|/C) > 1$, which is impossible – implying that $\mathbb{E}|f| = 0$ and that $f = 0$ almost surely.

Finally, let us turn to the triangle inequality. Let $f, h \in L_\Phi$ and set $\alpha > \|f\|_\Phi$ and $\beta > \|h\|_\Phi$. We need to show that $\alpha + \beta$ is a 'legal candidate' in the definition of $\|f + h\|_\Phi$, i.e., that

$$\mathbb{E}\Phi\left(\frac{f+h}{\alpha+\beta}\right) \leq 1.$$

Note that

$$\frac{f+h}{\alpha+\beta} = \frac{\alpha}{\alpha+\beta} \cdot \frac{f}{\alpha} + \frac{\beta}{\alpha+\beta} \cdot \frac{h}{\beta}$$

which is a convex combination of f/α and h/β. Hence,

$$\Phi\left(\frac{\alpha}{\alpha+\beta} \cdot \frac{f}{\alpha} + \frac{\beta}{\alpha+\beta} \cdot \frac{h}{\beta}\right) \leq \frac{\alpha}{\alpha+\beta}\Phi\left(\frac{f}{\alpha}\right) + \frac{\beta}{\alpha+\beta}\Phi\left(\frac{h}{\beta}\right)$$

and taking then expectation,

$$\mathbb{E}\Phi\left(\frac{f+h}{\alpha+\beta}\right) \leq \frac{\alpha}{\alpha+\beta} + \frac{\beta}{\alpha+\beta} = 1.$$

■

Another important choice of Φ is $\Phi(t) = \exp(|t|^\alpha) - 1$ for $1 \leq \alpha \leq 2$. The corresponding norms are called the ψ_α norms.

2.3.1 The Orlicz ψ_α norms

Recall that the natural hierarchy of L_p (probability) spaces, implies that for $1 \leq p \leq q \leq \infty$, $\|f\|_{L_p} \leq \|f\|_{L_q}$. And, by Chebyshev's inequality, functions with a finite L_p norm exhibit a polynomial tail decay. As it happens, the ψ_α norms capture an exponential tail decay and the corresponding normed space L_{ψ_α} 'lives' between all the L_p spaces for $1 \leq p < \infty$ and L_∞.

Definition 2.3.3 *Let $1 \leq \alpha \leq 2$. The ψ_α norm of $f : \Omega \to \mathbb{R}$ is*

$$\|f\|_{\psi_\alpha} = \inf\left\{C > 0 : \mathbb{E}\exp(|f/C|^\alpha) \leq 2\right\},$$

and the space of all measurable functions with a finite ψ_α norm is denoted by L_{ψ_α}.

The most natural example of a function (or random variable) X that belongs to L_{ψ_α} is the one with density $c_\beta \exp(-|t|^\beta)$ for some $1 \leq \beta \leq 2$, where c_β is an appropriate normalization constant. Indeed, observe that for $C > 0$,

$$\mathbb{E}\exp(|X|^\alpha/C^\alpha) = 2c_\beta \int_0^\infty \exp(-t^\beta + t^\alpha/C^\alpha)dt.$$

Thus, it is evident that $X \in L_{\psi_\beta}$ but $X \notin L_{\psi_\alpha}$ for $\alpha > \beta$.

Corollary 2.3.4 *If X has density $\sim \exp(-(|t|/L)^\alpha)$ then $\|X\|_{\psi_\alpha} \leq cL$ for an absolute constant c. In particular, if g is a centred gaussian random variable with variance σ^2 then $\|g\|_{\psi_2} \leq c\sigma$.*

It is a reasonable speculation that random variables with densities $c_\alpha \exp(-|t|^\alpha)$ are a good example of ψ_α random variables, and this speculation will be verified below. Let us see what other properties such random variables have – specifically, their tail behaviour and the growth of their moments.

Two obvious properties come to mind: tail estimates, and moments. Note that for $t > e$,

$$\begin{aligned} Pr(|X| > t) =& 2c_\alpha \int_t^\infty \exp(-u^\alpha)du = 2c_\alpha \sum_{j=0}^\infty \int_{2^j t}^{2^{j+1}t} \exp(-u^\alpha)du \\ \leq& 2c_\alpha \sum_{j=0}^\infty 2^j t \exp(-2^{\alpha j} t^\alpha) \leq 4c_\alpha \exp(-t^\alpha), \end{aligned}$$

by comparing the sum to a suitable geometric progression.

As for the moments of X, following a change of variables $t^\alpha \to u$, o

$$\mathbb{E}|X|^p = 2c_\alpha \int_0^\infty t^p \exp(t^\alpha)dt = (2c_\alpha/\alpha) \int_0^\infty u^{(p+1/\alpha)-1} \exp(-u)du,$$

and, of course, $c_\alpha = 1/2 \int_0^\infty \exp(-u^\alpha)du = (1/2\alpha) \int_0^\infty u^{(1/\alpha)-1} \exp(-u)du$. Recall the definition of the Gamma function

$$\Gamma(x) = \int_0^\infty u^{x-1} \exp(-u)du,$$

and thus,

$$\mathbb{E}|X|^p = \frac{\Gamma(\frac{p+1}{\alpha})}{\Gamma(\frac{1}{\alpha})}.$$

Using the natural hierarchy of the L_p norms it suffices to consider the case in which $p + 1/\alpha = m + 1$ for an integer m. Recall that $\Gamma(m + 1) = m!$ and that by Stirling's approximation, for every integer p,

$$\sqrt{2\pi p} \left(\frac{p}{e}\right)^p \leq p! \leq e\sqrt{2\pi p} \left(\frac{p}{e}\right)^p, \tag{2.11}$$

Hence, for $1 \leq \beta \leq 2$,

$$c_1 p^{1/\alpha} \leq \|X\|_{L_p} \leq c_2 p^{1/\alpha} \tag{2.12}$$

for suitable absolute constants c_1 and c_2.

It turns out that such tail estimates and moment growth condition actually characterize the ψ_α norms:

Theorem 2.3.5 *Each one of the following three conditions implies the other two:*

(1) $\mathbb{E}\exp(|f/L_1|^\alpha) \le 2$,

(2) $Pr(|f| \ge L_2 t) \le 2\exp(-|t|^\alpha)$ *for every* $t \ge 1$,

(3) for every $q \ge 1$, $\|f\|_{L_q(\mu)} \le L_3 q^{1/\alpha}$.

Moreover, for (1) $\Longrightarrow$ (2) *one may select* $L_1 = L_2$, *for* (2) $\Longrightarrow$ (3) *one may select* $L_3 = c_2 L_2$ *and for* (3) $\Longrightarrow$ (1) *one may select* $L_1 = c_3 L_3$ *for absolute constants* c_2 *and* c_3.

Theorem 2.3.5 implies that if $f \in L_{\psi_\alpha}$ then it also belongs to L_p for every $1 \le p < \infty$; however, such functions need not be bounded. Thus, the L_{ψ_α} hierarchy "lives" between all the L_p spaces and L_∞.

The proof of Theorem 2.3.5 requires an observation that was already used above: that there is an absolute constant c that satisfies that for every $p \ge 1$,

$$p\int_1^\infty u^{p-1}\exp(-u^\alpha)du \le c^p \cdot p^{p/\alpha}. \tag{2.13}$$

Proof of Theorem 2.3.5. (1) $\Longrightarrow$ (2) is an immediate outcome of Chebyshev's inequality: for $t \ge 0$,

$$Pr(|f| \ge L_2 t) = Pr(\exp(|f/L_1|^\alpha) \ge \exp((L_2 t/L_1)^\alpha) \le 2\exp(-|t|^\alpha)$$

once one selects $L_2 \ge L_1$.

Turning to (2) $\Longrightarrow$ (3), one may use tail integration:

$$\begin{aligned}\mathbb{E}|f|^q =& q\int_0^\infty t^{q-1}Pr(|f|>t)dt = L_2^q q\int_0^\infty u^{q-1}Pr(|f|>L_2 u)du\\ \le& L_2^q\left(1+q\int_1^\infty u^{q-1}\exp(-u^\alpha)du\right).\end{aligned}$$

Applying (2.13) one has that $(\mathbb{E}|f|^q)^{1/q} \le L_2 \cdot c_2 q^{1/\alpha}$, thus verifying (3) for any $L_3 \ge c_2 L_2$.

Finally, to show that (3) $\Longrightarrow$ (1), recall that $\exp(x) = 1 + \sum_{q\ge 1} x^q/q!$. By the monotone convergence theorem,

$$\mathbb{E}\exp(|f/L_1|^\alpha) = 1 + \sum_{q\ge 1}\frac{\mathbb{E}|f/L_1|^{\alpha q}}{q!} \le 1 + \sum_{q\ge 1}\left(\frac{L_3}{L_1}\right)^{\alpha q}\cdot\frac{q^q}{q!}\alpha^q = (*).$$

Since $\exp(q) \geq q^q/q!$ it follows that $(q^q/q!)^{1/q} \leq e$ and

$$(*) \leq 1 + \sum_{q \geq 1} \left(\frac{L_3^\alpha e \alpha}{L_1^\alpha} \right)^q \leq 2$$

provided that $L_1 \geq 2e^2 L_3$. ■

The most significant outcome of Theorem 2.3.5 is that in a similar fashion to (2.12), a finite ψ_α norm is actually equivalent to a tempered growth of moments: the L_p norm does not grow faster than $\sim p^{1/\alpha}$ (though unlike (2.12), the lower estimate on the L_p norms need not be true). Moreover, it follows that there are absolute constants c_1 and c_2 for which, for every $1 \leq \alpha \leq 2$,

$$c_1 \sup_{p \geq 1} \frac{\|f\|_{L_p}}{p^{1/\alpha}} \leq \|f\|_{\psi_\alpha} \leq c_2 \sup_{p \geq 1} \frac{\|f\|_{L_p}}{p^{1/\alpha}}. \tag{2.14}$$

Remark 2.3.6 *Recall that there is a real difference between the L_p norm and the weak L_p norm: the former is defined by a finite p-th moment, while the latter by a certain tail decay. The situation is different when it comes to ψ_α norms: the 'weak-space', characterized by having a faster tail decay than $\exp(-c|t|^\alpha)$, actually coincides with having a finite ψ_α norm.*

2.3.2 Subgaussian random variables

As we will explain in great detail later, there is a substantial difference between the assumption that a function has a finite norm – say, with respect to one of the normed spaces we have considered thus far, and assuming that two different norms of f are 'close'. The most significant notion of *norm equivalence* we will use is called *subgaussian*.

Definition 2.3.7 *A function f on (Ω, μ) is L-subgaussian if $\|f\|_{\psi_2} \leq L\|f\|_{L_2}$.*

The important factor in Definition 2.3.7 is the constant L. The reason is that $\|f\|_{\psi_2} \sim \sup_{p \geq 2} \|f\|_{L_p}/\sqrt{p}$ and therefore, if $\|f\|_{\psi_2} < \infty$ then $\|f\|_{L_2} \leq c\|f\|_{\psi_2}$ for an absolute constant c. The reverse inequality is what leads to nontrivial information on f and that information depends on the value of L.

The name subgaussian comes from the tail characterization of the ψ_2 norm: we know that for a suitable absolute constant c and every $t > 0$,

$$Pr(|f| \geq ct\|f\|_{\psi_2}) \leq 2\exp(-t^2);$$

but if $\|f\|_{\psi_2} \leq L\|f\|_{L_2}$ then $Pr(|f| \geq cLt\|f\|_{L_2}) \leq 2\exp(-t^2)$, implying that

$$Pr(|f| \geq t) \leq 2\exp(-ct^2/L^2\|f\|_{L_2}^2). \tag{2.15}$$

And, (2.15) means that the tail of f is dominated by the tail of a centred gaussian random variable with variance $\sim L\|f\|_{L_2}$.

Another observation is that for a suitable absolute constant c_1 and every $p \geq 2$,

$$\|f\|_{L_2} \leq \|f\|_{L_p} \leq c_1\sqrt{p}\|f\|_{\psi_2} \leq c_1 L\sqrt{p}\|f\|_{L_2},$$

implying that the L_2 and L_p norms of f are equivalent with the equivalence constant of the order of $L\sqrt{p}$.

[ADD AS EX: EQUIVALENCE BETWEEN L_2 AND L_1].

L-subgaussian random variable are happen to be rather natural objects and appear more frequently than one would expect:

- Let g be a centred gaussian random variable with variance σ^2. Thus,$\|g\|_{L_2} = \sigma$ and we saw that $\|g\|_{\psi_2} \leq c\sigma$. Therefore, g is L-subgaussian for L that is an absolute constant, and in particular, does not depend on σ.

- Let $g_1, ..., g_M$ be independent, centred gaussian variables with variance 1. Let $x = (x_1, ..., x_M) \in \mathbb{R}^M$ and put $Z_x = \sum_{i=1}^M x_i g_i$. Since the gaussian vector $G = (g_1, ..., g_M)$ is invariant under rotation, Z is distributed as $\|x\|_2 g$; in particular, $\|Z\|_{L_2} = \|x\|_2$ and $\|Z\|_{\psi_2} \leq c\|x\|_2$ for an absolute constant c. Thus, for every $x \in \mathbb{R}^M$ the random variable Z_x is L-subgaussian, with a constant L that is independent of the dimension M and of the vector x.

- Let $a > 0$ and set Z to be a symmetric, $\{-a, a\}$-valued random variable. It follows that $\|Z\|_{L_2} = a$, and $Pr(|Z| \geq ta) = 0$ for any $t > 1$. Therefore, by Theorem 2.3.5, $\|Z\|_{\psi_2} \leq ca$ for an absolute constant c.

- Let $\varepsilon_1, ..., \varepsilon_M$ be independent, symmetric, $\{-1, 1\}$-valued random variables. Let $x = (x_1, ..., x_M) \in \mathbb{R}^M$ and put $Z_x = \sum_{i=1}^M \varepsilon_i x_i$. Note that

  $$\|Z\|_{L_2}^2 = \mathbb{E}\sum_{i,j=1}^M x_i x_j \varepsilon_i \varepsilon_j = \sum_{i=1}^M x_i^2 = \|x\|_2^2.$$

  Let us show that $\|Z\|_{\psi_2} \lesssim \|x\|_2$, implying that Z_x is L-subgaussian for an absolute constant L. In particular, L is independent of the dimension M and the vector x. The proof is based on Höffding's inequality.

Lemma 2.3.8 *Let $\varepsilon_1, ..., \varepsilon_M$ be independent, symmetric, $\{-1, 1\}$-valued random variables and let $x = (x_1, ..., x_M) \in \mathbb{R}^M$. Then, for every $t > 0$,*

$$Pr\left(|\sum_{i=1}^M \varepsilon_i x_i| > t\|x\|_2\right) \leq 2\exp(-t^2/2).$$

The proof of Lemma 2.3.8 is based on an argument that will be used again several times – obtaining tail estimates using the moment general function.

Proof. Since the ε_i's are symmetric, $\{-1, 1\}$-valued, then for every $x_i \in \mathbb{R}$ and $\lambda > 0$,

$$\mathbb{E}\exp(\lambda\varepsilon_i x_i) = \frac{1}{2}\exp(\lambda x_i) + \frac{1}{2}\exp(-\lambda x_i) \leq \exp(\lambda^2 x_i^2/2),$$

because $\exp(t) + \exp(-t) \leq 2\exp(-t^2/2)$. The random variables $\varepsilon_1, ..., \varepsilon_M$ are independent and thus

$$Pr\left(\sum_{i=1}^M \varepsilon_i x_i \geq t\right) = Pr\left(\exp(\lambda\sum_{i=1}^M \varepsilon_i x_i) \geq \exp(\lambda t)\right) \leq \exp(-\lambda t)\mathbb{E}\exp(\lambda\sum_{i=1}^M \varepsilon_i x_i)$$

$$= \exp(-\lambda t)\prod_{i=1}^M \mathbb{E}\exp(\lambda\varepsilon_i x) \leq \exp(-\lambda t)\prod_{i=1}^M \exp(\lambda^2 x_i^2/2)$$

$$= \exp(-\lambda t + (\lambda^2/2)\sum_{i=1}^M x_i^2).$$

Optimizing the choice of λ,

$$Pr\left(\sum_{i=1}^M \varepsilon_i x_i > t\|x\|_2\right) \leq \exp(-t^2/2),$$

and the claim follows because $\sum_{i=1}^M \varepsilon_i x_i$ is a symmetric random variable. ■

The fact that $\sum_{i=1}^M \varepsilon_i x_i$ is L-subgaussian for a constant that is independent of M and of x is a reformulation of a classical result known as *Khintchine'a inequality.*

Theorem 2.3.9 *There exist absolute constants c_1 and c_2 for which the following holds. For any integer M, any $x \in \mathbb{R}^M$ and any $1 \leq p < \infty$,*

$$c_1\|\sum_{i=1}^M \varepsilon_i x_i\|_{L_2} \leq \|\sum_{i=1}^M \varepsilon_i x_i\|_{L_p} \leq c_2\sqrt{p}\|\sum_{i=1}^M \varepsilon_i x_i\|_{L_2}.$$

Thus far we have observed that both $\sum_{i=1}^M \varepsilon_i x_i$ and $\sum_{i=1}^M g_i x_i$ are L-subgaussian for an absolute constant L, and these facts are actually a rather general phenomenon. Indeed, let $v_1, ..., v_M$ be independent, mean-zero, variance 1 random variables and set $V = (v_1, ..., v_M)$. Note that for every $x \in \mathbb{R}^M$, $Z_x = \langle x, V \rangle$ satisfies

$$\|Z\|_{L_2}^2 = \mathbb{E}\left(\sum_{i=1}^M x_i v_i\right)^2 = \|x\|_2^2,$$

and we will show later (see Corollary 3.1.4) that

$$\|\sum_{i=1}^M x_i v_i\|_{\psi_2} \leq C\left(\sum_{i=1}^M x_i^2 \|v_i\|_{\psi_2}^2\right)^{1/2}$$

for a suitable absolute constant C. Therefore, if $\|v_i\|_{\psi_2} \leq L\|v_i\|_{L_2}$ for every $1 \leq i \leq M$, then for every $x \in \mathbb{R}^M$, $\|\langle V, x\rangle\|_{\psi_2} \leq CL\|\langle V, x\rangle\|_{L_2}$.

2.3.3 Example: the ψ_α norm for {1,...,M}

Let us turn to an example of a different flavour. Let $\Omega = \{1, ..., M\}$ endowed with the uniform probability measure and denote the corresponding ψ_α norm by $\| \ \|_{\psi_\alpha^M}$.

Theorem 2.3.5 and the fact that the probability measure of set is its normalized cardinality show that the $\|v\|_{\psi_\alpha^M}$ norm is determined by the rate of decay of the monotone rearrangement $(v_i^*)_{i=1}^M$.

Lemma 2.3.10 *There exists absolute constants c_1 and c_2 for which, for every $v \in \mathbb{R}^M$,*

$$c_1 \max_{1\leq k\leq M} \frac{v_k^*}{\log^{1/\alpha}(eN/k)} \leq \|v\|_{\psi_\alpha^M} \leq c_2 \max_{1\leq k\leq M} \frac{v_k^*}{\log^{1/\alpha}(eN/k)}.$$

where we use the natural identification of functions with vectors in $\mathbb{R}^M$ and $(v_i^*)_{i=1}^M$ is the nonincreasing rearrangement of $(|v_i|)_{i=1}^M$.

Proof. Applying Theorem 2.3.5 it is evident that for every $t \geq 1$,

$$|\{i : |v_i| \geq tc_0\|v\|_{\psi_\alpha^M}\}| = MPr(|v| \geq tc_0\|v\|_{\psi_\alpha^M}) \leq M\exp(-t^\alpha)$$

where c_0 is an absolute constant. Thus, for $t = \log^{1/\alpha}(eM/k)$,

$$|\{i : |v_i| \geq c_0\|v\|_{\psi_\alpha^M} \log^{1/\alpha}(eM/k)\}| \leq k/e. \tag{2.16}$$

Equation (2.16) implies that for every $k \le M$,

$$v_k^* \le c_0 \|v\|_{\psi_\alpha^M} \log^{1/\alpha}(eM/k), \tag{2.17}$$

that is,

$$\|v\|_{\psi_\alpha^M} \ge c_0^{-1} \sup_{1\le k\le M} \frac{v_k^*}{\log^{1/\alpha}(eM/k)}.$$

Turning to the reverse direction, let $\beta > 0$ and set

$$B = \beta \max_{1\le k\le M} \frac{v_k^*}{\log^{1/\alpha}(eM/k)}.$$

Thus, for every $1 \le k \le M$, $v_k^*/B \le \beta^{-1} \log^{1/\alpha}(eM/k)$, and

$$\begin{aligned}\mathbb{E}\exp(|v/B|^\alpha) =& \frac{1}{M}\sum_{k=1}^M \exp(|v_k/B|^\alpha) = \frac{1}{M}\sum_{k=1}^M \exp(|v_k^*/B|^\alpha)\\ \le& \frac{1}{M}\sum_{k=1}^M \exp\left(\beta^{-\alpha}\log(eM/k)\right) = \frac{1}{M}\sum_{k=1}^M \left(\frac{eM}{k}\right)^{1/\beta^\alpha}.\end{aligned}$$

Set $\delta = 1/\beta^\alpha$ and observe that for an integer k, $1/k^\delta \le \int_{k-1}^k x^{-\delta}dx$. Therefore,

$$\sum_{k=1}^M \left(\frac{1}{k}\right)^\delta \le 1 + \int_1^M x^{-\delta}dx \le 1 + \frac{M^{1-\delta}}{1-\delta}.$$

Hence,

$$\frac{1}{M}\sum_{k=1}^M \left(\frac{eM}{k}\right)^\delta \le \frac{e^\delta}{M^{1-\delta}} + \frac{e^\delta}{1-\delta} < 2,$$

provided that $M \ge 2$ and $\delta \le \delta_0$; i.e, $\beta \ge c_1$ for a suitable absolute constant c_1. Therefore, by the definition of the ψ_α^M norm,

$$\|v\|_{\psi_\alpha^M} \le c_1 \max_{1\le k\le M} \frac{v_k^*}{\log^{1/\alpha}(eM/k)}.$$

■