Geometric Aspects of Statistical Learning Theory — A Preliminary Draft

Shahar Mendelson

Contents

Part I

Preliminaries

Chapter 1

Basic function spaces

In what follows, we will be interested in various properties of normed spaces consisting of functions defined on a probability space (Ω, μ). The simplest norms we shall encounter are the L_p norms.

Definition 1.1. *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 let

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

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

To ease notation, we 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 functionals $\| \ \|_{L_p}$ are indeed norms. Moreover, because (Ω, μ) is a probability space, the L_p norms form a hierarchy:

$$\|f\|_{L_p} \leq \|f\|_{L_q} \quad \text{if} \quad 1 \leq p \leq q \leq \infty. \tag{1.1}$$

The space of measurable functions with a finite L_p norm is denoted by $L_p(\mu)$. Again, we often omit the dependence on the measure and write L_p instead of $L_p(\mu)$. Also, observe that by (1.1), if $1 \leq p \leq q \leq \infty$, then $L_q \subset L_p$.

Another useful feature of the L_p norm is that it could be computed using tail integration:

Exercise 1. *Show that for $1 \leq p < \infty$,*

$$\|f\|_{L_p}^p = p \int_0^\infty t^{p-1} Pr(|f| > t) dt. \tag{1.2}$$

A significant part of these notes is devoted to the question of preservation of structure through random sampling. In this chapter we present a very naive version of this question, and for a single function/random variable, and use it to describe some of the important features of the L_p norm.

Let f be a function on (Ω, μ) and let X be distributed according to μ. One is given a sample $X_1, ..., X_N$, consisting of N independent copies of the random variable X which is distributed according to the underlying measure μ. One would like to see when the values

$$(f(X_1), ..., f(X_N))$$

can be used to reconstruct $\|f\|_{L_p(\mu)}$, and for that reconstruction to be valid with high probability with respect to the given sample. The simplest attempt of reconstruction is by using the empirical mean

$$\frac{1}{N}\sum_{i=1}^{N} |f|^p(X_i),$$

and the reconstruction is successful if there are absolute constants c and C such that, with high probability,

$$c\|f\|^p_{L_p(\mu)} \leq \frac{1}{N}\sum_{i=1}^{N} |f|^p(X_i) \leq C\|f\|^p_{L_p(\mu)}, \tag{1.3}$$

which is an *isomorphic* reconstruction; when $c = 1 - \varepsilon$ and $C = 1 + \varepsilon$ for $0 < \varepsilon < 1$, (1.3) is an *almost isometric* reconstruction.

In this chapter we shall not explore the question of whether the empirical means can by used to reconstruct $\|f\|_{L_p}$. Rather, let us present two examples of cases in which this question is of interest. In both cases the central object is $\mathbb{R}^d$, though viewed in different ways. As a result, the question of L_p norm preservation via empirical means is completely different.

Remark 1.2. *The fact that the empirical mean have been chosen as a preliminary reconstruction option should not lead the reader to expect it to be a wise choice. In fact, one of the main aspects of this presentation is that using the empirical mean is, at least from certain aspects, rather useless: unless f is particularly nice, the empirical mean is typically close to $\|f\|^p_{L_p}$ only for samples that belong to an event that is not very big. To obtain estimates in the very high confidence regime, a more sophisticated recovery methods are called for.*

1.1 Two examples

The space $\mathbb{R}^d$ is featured in two natural ways in this presentation: sometimes it serves as a space of functions on the probability space $\{1, ..., d\}$ and sometimes as a probability space, endowed with some natural measure; in that case one is interested in functions defined on $\mathbb{R}^d$.

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

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

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

It follows that

$$\|v\|_p = \left(\sum_{i=1}^d |v_i|^p\right)^{1/p} = d^{1/p}\left(\frac{1}{d}\sum_{i=1}^d |v|^p(i)\right)^{1/p} = d^{1/p}\|v\|_{L_p^d}, \tag{1.4}$$

and $\|\ \|_{L_p^d}$ is just a re-scaling of the ℓ_p norm $\|\ \|_p$. The space $(\mathbb{R}^d, \|\ \|_{L_p^d})$ is denoted by L_p^d, and the unit ball in ℓ_p^d is a 'scaled-down' copy of the unit ball in L_p^d:

$$B_p^d = \{v \in \mathbb{R}^d : \|v\|_p \leq 1\} = d^{-1/p}\{v : \|v\|_{L_p^d} \leq 1\} = d^{-1/p}B(L_p^d). \tag{1.5}$$

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

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

Sampling in this case consists of a selection of a subset $I \subset \{1, ..., d\}$ and one has access to the values $f(i)$ for $i \in I$. Clearly, there are various ways of selecting a subset in a reasonable way, for example, by using *selectors*: let $\delta_1, ..., \delta_d$ be independent, $\{0, 1\}$-valued random variables with mean δ, and set $I = \{i : \delta_i = 1\}$. Another natural possibility is to select a subset uniformly from all subsets of a fixed cardinality, etc.

Intuitively, the ideal behaviour of a reasonable sampling method is that for a typical selection of $I \subset \{1, ..., d\}$,

$$\left(\frac{1}{d}\sum_{i=1}^d |v_i|^p\right)^{1/p} \sim \left(\frac{1}{|I|}\sum_{i=1}^{|I|} |v_i|^p\right)^{1/p}; \tag{1.6}$$

that is, $\|v\|_{L_p^d} \sim \|v\|_{L_p^I}$. Of course, understanding when that equivalence holds, and obtaining quantitative estimates on the equivalence constants and the probability (relative to the choice of I) for which (1.3) holds, is, at this point, a long way away.

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

Let us turn to the other role that $\mathbb{R}^d$ plays—as a probability space. Let μ be a probability measure on $\mathbb{R}^d$ and let X be a random vector, distributed according to μ. Hence, $(\mathbb{R}^d, \mu)$ is a probability space, and one may consider the L_p space of functions on $(\mathbb{R}^d, \mu)$. For the time being, our focus is on a rather particular choice on functions: the subspace of $L_p(\mu)$ containing the linear functionals on $\mathbb{R}^d$.

Clearly, there is a natural correspondence between $\mathbb{R}^d$ and those linear functionals: each $v \in \mathbb{R}^d$ defines a linear functional by

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

Having said that, it is far from obvious that f_v actually belongs to L_p. For that to happen, $\|f\|_{L_p}$ must be finite, and here

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

Here, independent sampling means that N points, $X_1, ..., X_N$, each distributed according to X are selected independently, and in the context of (1.3), one would like to ensure that

$$\frac{1}{N}\sum_{i=1}^{N} |\langle X_i, v\rangle|^p \sim \mathbb{E}|\langle X, v\rangle|^p.$$

Despite the natural correspondence between points in $\mathbb{R}^d$ and the linear functionals, there is no reason to expect any connection between $\|v\|_p$ and $\|f_v\|_{L_p}$. Although both $\| \ \|_p$ and $\| \ \|_{L_p}$ are norms on $\mathbb{R}^d$, these norms can be totally different.

In what follows we abuse notation and write $\|v\|_{L_p}$ instead of $\|f_v\|_{L_p}$, and let us begin by exploring the case $p = 2$, in which both norms are endowed by inner products.

Definition 1.3. *A measure μ (resp. a random vector X) on $\mathbb{R}^d$ is isotropic if it is symmetric and for every $v \in \mathbb{R}^d$,*

$$\|v\|^2_{L_2(\mu)} = \int_{\mathbb{R}^d} |\langle v, x\rangle|^2 d\mu(x) = \|v\|_2^2;$$

in other words, $\mathbb{E}|\langle v, X\rangle|^2 = \|v\|_2^2$.

Hence, when the measure is isotopic, not only does each $f_v \in L_2(\mu)$, but the mapping $v \to f_v$ is an isometric embedding of $(\mathbb{R}^d, \| \ \|_2)$ in $L_2(\mu)$. Of course, this does not mean that $f_v \in L_p(\mu)$ for $p > 2$.

Exercise 2. *Give an example of an isotropic random vector on $\mathbb{R}^d$ and some $v \in \mathbb{R}^d$ such that $\langle X, v\rangle \notin L_p$ for any $p > 2$.*

In general, if every f_v belongs to $L_2(\mu)$ then the inner product in L_2 endows an alternative inner product of $\mathbb{R}^d$. Set $X = (x_1, ..., x_d)$ to be a random vector on $\mathbb{R}^d$, let $e_1, ..., e_d$ be the standard basis in $\mathbb{R}^d$ and define

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

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

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

is an inner product on $\mathbb{R}^d$. The unit ball of the norm that inner product endows on $\mathbb{R}^d$ is the ellipsoid

$$\left\{ v \in \mathbb{R}^d : \langle \mathrm{Cov}(X)v, v\rangle \leq 1 \right\}.$$

In the special case of an isotropic measure, $\mathrm{Cov}(X) = Id$ is the identity matrix, $\langle v, u\rangle = [v, u]$; the ℓ_2 inner product $\langle \cdot, \cdot \rangle$ and the inner product $[\cdot, \cdot]$ endowed by L_2 coincide on $\mathbb{R}^d$; and the ellipsoid is the standard Euclidean ball. However, in general, $\| \ \|_{L_2(\mu)}$ and $\| \ \|_2$ are different norms.

The difference between $\| \ \|_{L_p(\mu)}$ and $\| \ \|_p$ is even more obvious when considering the standard gaussian vector G, whose density is proportional to $c\exp(-\|t\|_2^2/2)$. Clearly, G has

the same distribution as $(g_1, ..., g_d)$ where the g_i's are independent, standard gaussian random variables. Observe that for every $v \in \mathbb{R}^d$,

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

implying that G is an isotropic random vector, and in particular, the L_2 norm it endows on $\mathbb{R}^d$ coincides with the standard Euclidean norm $\| \ \|_2$. We show in what follows that for every $v \in \mathbb{R}^d$ and any $1 \leq p < \infty$,

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

implying that $\{v : \|\langle G, v\rangle\|_{L_p} \leq 1\} \sim \sqrt{p} B_2^d$. However, that set is a very different from the unit ball in ℓ_p^d.

Remark 1.4. *L_p norm preservation (or reconstruction) in the gaussian case via the empirical means implies that with high probability, if $G_1, ..., G_N$ are independent copies of G,*

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

which is a completely different question than (1.6), although both setups deal with $\mathbb{R}^d$.

1.2 Weak L_p spaces

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

$$\|f\|_{L_{p,\infty}(\mu)} = \inf\left\{A > 0 : \sup_{t \in \mathbb{R}^+} t^p Pr\left(|f| > tA\right) \leq 1\right\}, \tag{1.7}$$

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

The weak L_p space consists of all the functions for which $\|f\|_{L_{p,\infty}} < \infty$, and it is denoted by $L_{p,\infty}$.

Remark 1.5. *It should stressed that $\|f\|_{L_{p,\infty}}$ is actually not a norm, as it does not satisfy the triangle inequality. Having said that, we will keep referring to it as the weak L_p norm and denote it by $\| \ \|_{L_{p,\infty}}$ as if it were a norm.*

Exercise 3. *Show that indeed, $\| \ \|_{L_{p,\infty}}$ need not satisfy the triangle inequality.*

There is a true difference between the L_p norm and the weak L_p norm. Indeed, as was noted previously,

$$\|f\|_{L_p}^p = p\int_0^\infty t^{p-1} Pr(|f| > t) dt. \tag{1.8}$$

Moreover, a straightforward application of Chebyshev's inequality shows that

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

and in particular,

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

i.e., $\|f\|_{L_{p,\infty}} \le \|f\|_{L_p}$. However, if $f \in L_{p,\infty}$ then its tail probability $Pr(\{|f| > t\})$ decays faster than $\sim 1/t^p$, but that does not ensure integrability as in (1.8).

Exercise 4. *Construct an example of a function on (Ω, μ) that belong to $L_{p,\infty}$ but not to L_p.*

Although the weak L_p norm is indeed weaker than the L_p norm, the next lemma shows that it is only slightly weaker.

Lemma 1.6. *If $1 \le p < q < \infty$, then*

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

Proof. Clearly, $\sup_{t\in\mathbb{R}^+} t^q Pr(|f| > tA) \le 1$ if and only if $\sup_{t\in\mathbb{R}^+} t^q Pr(|f| > t) \le A^q$. Hence, by (1.8),

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

■

Chapter 2

Orlicz norms and maximal inequalities

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

Definition 2.1. *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*

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

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

Example 2.2. *Let $\Phi(t) = |t|^p$ for $1 \leq p < \infty$. Then $\mathbb{E}(|f|^p/C^p) \leq 1$ when $\|f\|_{L_p} \leq C$, implying that $\|f\|_\Phi = \|f\|_{L_p}$ and $L_\Phi = L_p$.*

Lemma 2.3. *Let Φ be an even, convex function that is increasing in $\mathbb{R}^+$ and satisfies $\Phi(0) = 0$. 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 $C > 0$, $\mathbb{E}\Phi(f/C) \leq 1$. Now, since Φ is even, it follows from Jensen's inequality that 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, Φ is convex and increasing, and since $\Phi(0) = 0$ and $\Phi \not\equiv 0$, one has 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, one has to establish the triangle inequality. Let $f, h \in L_\Phi$ and set $\alpha > \|f\|_\Phi$ and $\beta > \|h\|_\Phi$. Let us show that $\alpha + \beta$ is a 'legal candidate' in the definition of $\|f + h\|_\Phi$, i.e., that

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

To that end, 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 the expectation on both sides,

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

■

An important choice of a family of Orlicz norms is $\Phi_\alpha(t)=\exp(|t|^\alpha)-1$ for $1\leq\alpha\leq 2$. The corresponding norms are called ψ_α norms and they play a significant role in what follows.

2.1 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. The ψ_α norms capture an exponential tail decay and thus 'live' between the L_p spaces for $1\leq p<\infty$ and L_∞.

Definition 2.4. *Let $1\leq\alpha\leq 2$. The ψ_α norm of $f:\Omega\to\mathbb{R}$ is*

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

The space of all 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. Observe that for $C>0$,

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

implying that $X\in L_{\psi_\beta}$ but $X\notin L_{\psi_\alpha}$ for $\alpha>\beta$.

Corollary 2.5. *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 $\|g\|_{\psi_2}\leq c\sigma$.*

Because the belief is that random variables with densities $c_\alpha\exp(-|t|^\alpha)$ are a good example of ψ_α random variables, let us explore their moments growths and tail decays.

Note that for $t>e$,

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

where the last inequality follows by comparing to an appropriate geometric progression.

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

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

where

$$c_\alpha = 1/2 \int_0^\infty \exp(-u^\alpha) du = 1/2 \int_0^\infty u^{(1/\alpha)-1} \exp(-u) du.$$

Recall the definition of the Gamma function

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

and thus,

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

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 m,

$$\sqrt{2\pi m}\left(\frac{m}{e}\right)^m \leq m! \leq e\sqrt{2\pi m}\left(\frac{m}{e}\right)^m, \tag{2.1}$$

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

$$c_1 p^{1/\alpha} \leq \|X\|_{L_p} \leq c_2 p^{1/\alpha} \tag{2.2}$$

for suitable absolute constants c_1 and c_2.

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

Theorem 2.6. *Each one of the following three conditions implies the other two:*

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

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

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

Moreover, for (1) $\Longrightarrow$ (2) *one may select* $L_1 = L_2$, *for* (2) $\Longrightarrow$ (3) *one may select* $L_3 = 2eL_2$ *and for* (3) $\Longrightarrow$ (1) *one may select* $L_1 = 2e^2 L_3$ *(though all these choices are not optimal).*

Theorem 2.6 implies that if $f \in L_{\psi_\alpha}$ then it also belongs to L_q for every $1 \leq q < \infty$; 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 2.6 requires an observation that was used previously: there is an absolute constant c which satisfies that for every $q \geq 1$ and any $1 \leq \alpha \leq 2$,

$$q \int_1^\infty u^{q-1} \exp(-u^\alpha) du \leq c^q \cdot q^{q/\alpha}. \tag{2.3}$$

Proof of Theorem 2.6. (1) $\Longrightarrow$ (2) is an immediate outcome of Chebyshev's inequality: for $t \geq 0$,

$$Pr(|f| \geq L_2 t) = Pr\left(\exp(|f/L_1|^\alpha) \geq \exp((L_2 t/L_1)^\alpha)\right) \leq 2\exp(-|t|^\alpha)$$

once one selects $L_2 \geq L_1$.

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

$$\begin{aligned}\mathbb{E}|f|^q =& q\int_0^\infty t^{q-1}Pr(|f|>t)dt = L_2^q q\int_0^\infty u^{q-1}Pr(|f|>L_2u)du \\ \leq& L_2^q\left(1+q\int_1^\infty u^{q-1}\exp(-u^\alpha)du\right).\end{aligned}$$

Applying (2.3) one has that $(\mathbb{E}|f|^q)^{1/q} \leq L_2 \cdot cq^{1/\alpha}$, thus verifying (3) for any $L_3 \geq cL_2$.

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

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

Since $\exp(q) \geq q^q/q!$ one has $(q^q/q!)^{1/q} \leq e$, 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 2.6 is that in a similar fashion to (2.2), a finite ψ_α norm is actually equivalent to a tempered growth of moments: the L_q norm does not grow faster than $\sim q^{1/\alpha}$ (though unlike (2.2), the lower estimate on the L_q norms need not be true). Moreover, it follows that there are absolute constants c_1 and c_2 such that, for every $1 \leq \alpha \leq 2$,

$$c_1 \sup_{q\geq 1}\frac{\|f\|_{L_q(\mu)}}{q^{1/\alpha}} \leq \|f\|_{\psi_\alpha} \leq c_2 \sup_{q\geq 1}\frac{\|f\|_{L_q(\mu)}}{q^{1/\alpha}}. \tag{2.4}$$

Remark 2.7. *Recall that there is a real difference between the L_p norm and the weak L_p norm as the latter is determined by a certain tail decay property. The situation is different when it comes to ψ_α norms: the 'weak-space', characterized by a faster tail decay than $\exp(-|t|^\alpha)$, actually coincides with having a finite ψ_α norm.*

Exercise 5. *Show that the is an absolute constant C, such that for every mean-zero random variable,*

$$\|X\|_{\psi_2} \leq C\mathbb{E}\exp(|X|^2).$$

2.2 Subgaussian random variables

A significant fact (and a source of much confusion) is that there is a substantial difference between the assumption that a function has a finite norm—say, $\|f\|_{L_p} < \infty$ or $\|f\|_{\psi_\alpha} < \infty$—, and assuming that two different norms of f are equivalent. The information that can be derived from the two assumptions is of a completely different nature. In the context of these notes, the most important notion of norm equivalence is called *subgaussian*[1].

[1]It should be noted that in some places, being called "subgaussian" means having a bounded ψ_2 norm.

Definition 2.8. *A function f defined on the probability space (Ω, μ) is L-subgaussian if $\|f\|_{\psi_2} \leq L\|f\|_{L_2}$.*

Clearly, the important factor in Definition 2.8 is the identity of the equivalence constant L. Later we study L-subgaussian classes of functions, that is, classes that consist of functions that share the same equivalence constant.

The origin of the name "subgaussian" is the characterization of the tail behaviour of functions with a finite ψ_2 norm. Indeed, recall 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 is L subgaussian then $\|f\|_{\psi_2} \leq L\|f\|_{L_2}$, implying that

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

In other words, the tail of f is dominated by the tail of a centred gaussian random variable with variance $\sim (L\|f\|_{L_2})^2$.

Moreover, if f is centred and L-subgaussian, then thanks to Theorem 2.6,

$$\|f\|_{\psi_2} \sim \sup_{q\geq 1} \frac{\|f\|_{L_q}}{\sqrt{q}}.$$

Thus, for a suitable absolute constant c_1 and every $q \geq 2$,

$$\|f\|_{L_2} \leq \|f\|_{L_q} \leq c_1\sqrt{q}\|f\|_{\psi_2} \leq c_1 L\sqrt{q}\|f\|_{L_2},$$

implying that all the L_2 and L_q norms of f are equivalent with the equivalence constant $\sim L\sqrt{q}$.

Exercise 6. *Show that if f is centred and L-subgaussian then*

$$\|f\|_{L_2} \leq c(L)\|f\|_{L_1}$$

for a constant c that depends only on L.

It turns out that L-subgaussian functions/random variable appear frequently and in natural situations.

- **A gaussian is subgaussian:** Let g be a centred gaussian random variable with variance σ^2. Thus, $\|g\|_{L_2} = \sigma$ and $Pr(|g| \geq t\sigma) \leq c\exp(-t^2)$. Applying Theorem 2.6, $\|g\|_{\psi_2} \lesssim \sigma$, and g is L-subgaussian for L that is an absolute constant.

- **Stability under tensorization:** Let $Z_1, ..., Z_d$ be independent, centred L-subgaussian variables with variance 1. Let $x = (x_1, ..., x_d) \in \mathbb{R}^d$ and put $Z_x = \sum_{i=1}^d x_i Z_i$. Clearly, $\|Z_x\|_{L_2} = \|x\|_2$. We show in what follows that there is an absolute constant C such that

$$\|Z_x\|_{\psi_2} \leq C\left(\sum_{i=1}^d x_i\|Z_i\|_{\psi_2}^2\right)^{1/2},$$

  implying that $\|Z_x\|_{\psi_2} \leq cL\|x\|_2$ for an absolute constant c. Thus, for every $x \in \mathbb{R}^d$ the random variable Z_x is L-subgaussian, with a constant that is independent of the dimension d 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 2.6, $\|Z\|_{\psi_2} \leq ca$ for an absolute constant c.

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

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

  Since the ψ_2 is stable under tensorization (which still has to be proved), one has that $\|Z_x\|_{\psi_2} \leq c\|x\|_2$, implying that Z_x is L-subgaussian with a constant that is independent of the dimension d and the specific choice of the vector x.

  Let us now give a direct proof of that fact—the so-called Höffding inequality.

Lemma 2.9. *Let $\varepsilon_1, ..., \varepsilon_d$ be independent, symmetric, $\{-1, 1\}$-valued random variables and let $x = (x_1, ..., x_d) \in \mathbb{R}^d$. Then, for every $t > 0$,*

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

The proof of Lemma 2.9 is based on an argument that is used frequently in what follows—obtaining tail estimates using the moment generating function.

Proof. Since the random variables ε_i, $1 \leq i \leq d$ are symmetric and $\{-1, 1\}$-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 $\exp(t) + \exp(-t) \leq 2\exp(-t^2/2)$. Therefore, by the independence of $\varepsilon_1, ..., \varepsilon_d$,

$$\begin{aligned} Pr\left(\sum_{i=1}^d \varepsilon_i x_i \geq t\right) &= Pr\left(\exp\left(\lambda \sum_{i=1}^d \varepsilon_i x_i\right) \geq \exp(\lambda t)\right) \\ &\leq \exp(-\lambda t) \cdot \mathbb{E}\exp\left(\lambda \sum_{i=1}^d \varepsilon_i x_i\right) \\ &= \exp(-\lambda t) \prod_{i=1}^d \mathbb{E}\exp(\lambda \varepsilon_i x) \leq \exp(-\lambda t) \prod_{i=1}^d \exp\left(\lambda^2 x_i^2/2\right) \\ &= \exp\left(-\lambda t + (\lambda^2/2)\sum_{i=1}^d x_i^2\right). \end{aligned}$$

Optimizing the choice of λ one has that

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

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

■

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

Theorem 2.10. *There exist absolute constants c_1 and c_2 for which the following holds. For any integer d, any $x \in \mathbb{R}^d$ and any $1 \leq p < \infty$,*

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

Exercise 7. *Construct an infinite-dimensional space E, consisting of functions defined on a probability space (Ω, μ) such that E is closed in $L_2(\mu)$ and on which all the $L_p(\mu)$ norms are equivalent: for $p \geq 1$ there are constants c_p and C_p that depend only on p such that for any $f \in E$,*

$$c_p \|f\|_{L_2} \leq \|f\|_{L_p} \leq C_p \|f\|_{L_2}.$$

Hint: construct a sequence of symmetric, $\{-1, 1\}$-valued that are independent $\varepsilon_i : [0, 1] \to \mathbb{R}$. Then use Theorem 2.10.

2.3 Maximal inequalities

The study of the supremum of a collection of random variable $\{Z_t : t \in T\}$ has been the subject of extensive study over the years (see, for example [**?**]). We refer the reader to Talagrand's treasured manuscript [**?**] which is the most comprehensive one on this topic.

For now, let us consider a more modest goal:

Question 2.11. *Let T be a finite set. Is there a simple way of obtaining (possibly crude) estimates on*

$$Pr(\sup_{t \in T} Z_t \geq u)?$$

As always, let us begin with an example. Assume that $|T| = \{1, ..., m\}$ and that each Z_i is a standard gaussian random variable. Of course, there is likely to be a substantial difference in the behaviour of the supremum, depending on the correlation between the random variables Z_i. In the simplest of situations, the random variables Z_t are independent, and one has that

$$Pr\left(\exists 1 \leq i \leq m \ : \ |Z_i| \geq u\right) = 1 - Pr^m(|g| \leq u) = 1 - (1 - Pr(|g| > u))^m.$$

Clearly, $(1 - Pr(|g| > u))^m = 1/2$ when $Pr(|Z| > u) \sim 1/m$; hence, setting $u = c_1\sqrt{\log m}$,

$$Pr\left(\max_{1 \leq i \leq m} |Z_i| \geq c_1 \sqrt{\log m}\right) \geq 1/2.$$

On the other hand, if $u \gtrsim \sqrt{\log m}$ then

$$Pr\left(\exists 1 \leq i \leq m \ : \ |Z_i| \geq u\right) \leq mPr(|g| \geq u) \leq \exp(\log m - u^2/2) \leq \exp(-u^2/4).$$

Exercise 8. *Show that there is an absolute constant c such that for every $p \geq 1$ and integer m,*

$$\left(\mathbb{E} \left(\max_{1 \leq i \leq m} |Z_i| \right)^p \right)^{1/p} \leq c(\sqrt{\log m} + \sqrt{p})$$

This example happens to be more indicative than what one would expect: the upper estimate on $Pr(\max |Z_i| \geq u)$ is based on the union bound, applied in a rather direct way. It implies that

$$\mathbb{E} \sup_{1 \leq i \leq m} |Z_i| \leq c\sqrt{\log m}$$

and that for larger values of u than $\sqrt{\log m}$, the $\max_{1 \leq i \leq m} |Z_i|$ decays as fast as a single gaussian.

The reverse inequality is a different story: independence is used to generate the lower bound and if the variables Z_i were not independent, there would have been no reason to expect that $\mathbb{E} \max_{1 \leq i \leq m} |Z_i| \sim \sqrt{\log m}$—for example, if $Z_1 = Z_2 = ... = Z_m = g$, the lower bound is just an absolute constant.

In later parts of these notes we will present *generic chaining*: a general mechanism, introduced by Michel Talagrand, which leads to upper bounds on $\mathbb{E} \sup_{t \in T} Z_t$. The method is based on identifying natural metric structures d_p endowed on T by the process, followed by the study of the geometry of the metric spaces (T, d_p).

Although chaining leads, in many cases, to satisfactory upper bounds, obtaining matching lower bounds turns out to be a formidable task which is still far from being fully understood.

At this point, targeting modest goals and armed with a rather limited set of tools, one can still derive the following general maximal inequalities which are used extensively in what follows.

Theorem 2.12. *There exists an absolute constant c_1 for which the following hold. If $Z_1, ..., Z_m$ are random variables, then*

(1) *For any $p \geq \log m$, $(\mathbb{E} \max_{1 \leq i \leq m} |Z_i|^p)^{1/p} \leq e \max_{1 \leq i \leq m} \|Z_i\|_{L_p}$.*

(2) *For every $1 \leq \alpha \leq 2$, $\| \max_{1 \leq i \leq m} Z_i \|_{\psi_\alpha} \leq c_1 \log^{1/\alpha} m \cdot \max_{1 \leq i \leq m} \|Z_i\|_{\psi_\alpha}$.*

The proof of Theorem 2.12 is based on the following observation: if $p \geq c \log m$ then $(\mathbb{R}^m, \| \ \|_\infty)$ and $(\mathbb{R}^m, \| \ \|_p)$ are equivalent; in other words, for every $x \in \mathbb{R}^m$,

$$\|x\|_\infty \leq \|x\|_p \leq e\|x\|_\infty. \tag{2.5}$$

Indeed, the first inequality is just the natural hierarchy of the ℓ_p norms. For the second one, assume without loss of generality that $x_i \geq 0$ and $x_1 \geq x_2 \geq ... \geq x_m$. Since $m^{1/p} \leq m^{1/\log m} \leq e$,

$$\|x\|_p = \Big(\sum_{i=1}^m x_i^p\Big)^{1/p} \leq m^{1/p} x_1 \leq e\|x\|_\infty.$$

Proof of Theorem 2.12. For every realization of $Z_1, ..., Z_m$, let $\mathcal{Z} = (Z_1, ..., Z_m)$. Thus,

$$\max_{1\le i\le m} |Z_i| = \|\mathcal{Z}\|_\infty \le \big(\sum_{i=1}^m |Z_i|^p\big)^{1/p},$$

and

$$\mathbb{E}\max_{1\le i\le m} |Z_i|^p \le \sum_{i=1}^m \mathbb{E}|Z_i|^p \le m\cdot \max_{1\le i\le m} \|Z\|_{L_p}^p.$$

The first part follows because $m^{1/p}\le e$.

Turning to the second part, set $K = \max_{1\le i\le m}\|Z_i\|_{\psi_\alpha}$. Recall that for every $p\ge 1$, $\|Z\|_{L_p}\le Lp^{1/\alpha}\|Z\|_{\psi_\alpha}$. Therefore, by the first part, for $p\ge \log m$,

$$\|\max_{1\le i\le m} Z_i\|_{L_p} = \left(\mathbb{E}\max_{1\le i\le m}|Z_i|^p\right)^{1/p} \le e\max_{1\le i\le m}\|Z_i\|_{L_p} \le eLp^{1/\alpha}K.$$

For $p\le \log^{1/\alpha} m$,

$$Pr(\max_{1\le i\le m}|Z_i|>u) \le \sum_{i=1}^m Pr(|Z_i|\ge u) \le m\exp(-cu^\alpha/K^\alpha) \le \exp(-c_1u^\alpha/K^\alpha),$$

provided that $u\ge c_2K\log^{1/\alpha} m$. Therefore, setting $u_0 = c_3K\log^{1/\alpha} m$, it follows that

$$\mathbb{E}\max_{1\le i\le m}|Z_i|^p = \int_0^\infty pu^{p-1}Pr(\max_{1\le i\le m}|Z_i|>u)du \le u_0^p + \int_{u_0}^\infty pu^{p-1}\exp(-c_1u^\alpha/K^\alpha)du$$
$$\le u_0^p + \int_{u_0}^\infty \exp(-c_4u^\alpha/K^\alpha)du,$$

provided that $c_1u^\alpha/K^\alpha \ge 2(\log p + (p-1)u)$ for any $u\ge u_0$. Since $p\le \log m$, that is the case for a suitable choice of the absolute constant c_3; hence,

$$\|\max_{1\le i\le m}|Z_i|\|_{L_p} \le cu_0 \sim K\log^{1/\alpha} m,$$

and the claim follows by recalling that $\|X\|_{\psi_\alpha}\sim \sup_{p\ge 1}\|X\|_{L_p}/p^{1/\alpha}$. ∎

It is instructive to see that the ψ_α estimate from Theorem 2.12 hides the true picture, because $\|X\|_{\psi_\alpha}$ is equivalent to the *largest* ratio $\|X\|_{L_p}/p^{1/\alpha}$. While there are values of p for which this ratio is attained, there could be others for which $\|X\|_{L_p}$ is significantly smaller than $p^{1/\alpha}\|X\|_{\psi_\alpha}$. The estimate on the maximum of $Z_1, ..., Z_m$ is one such example: the ψ_α estimate implies that for every $p\ge 1$,

$$\|\max_{1\le i\le m} Z_i\|_{L_p} \le cp^{1/\alpha}\cdot\|\max_{1\le i\le m} Z_i\|_{\psi_\alpha} \lesssim p^{1/\alpha}\log^{1/\alpha} m \max_{1\le i\le m}\|Z_i\|_{\psi_\alpha},$$

but in reality, the situation is much better. For $p\le \log m$ we have that

$$\|\max_{1\le i\le m} Z_i\|_{L_p} \le \log^{1/\alpha} m \max_{1\le i\le m}\|Z_i\|_{\psi_\alpha}$$

(which, as indicated by the gaussian example, cannot be improved even for $p=2$). In other words, the estimate for $\|\max_{1\le i\le m} Z_i\|_{L_2}$ remains stable up to $p=\log m$. Only at that point does $\|\max_{1\le i\le m} Z_i\|_{L_p}$ begin to grow like a ψ_α random variable. The actual estimate is summarized in the following corollary:

Corollary 2.13. *There exists an absolute constant c_1 for which the following hold. If $Z_1, ..., Z_m$ are random variables, then for every $p \geq 1$,*

$$\| \max_{1 \leq i \leq m} |Z_i| \|_{L_p} \leq c_1 \max\{\log^{1/\alpha} m, \log^{1/\alpha} p\} \cdot \max_{1 \leq i \leq m} \|Z_i\|_{\psi_\alpha}.$$

Chapter 3

Independent random variables

As the proof of Lemma 2.9 shows, one may generate tail estimates for a sum of independent random variables based on the behaviour of the moment generating function: if $Z_1, ..., Z_N$ are independent random variables, then for any $\lambda > 0$,

$$Pr\left(\sum_{i=1}^{N} Z_i \geq t\right) = Pr\left(\exp\left(\lambda \sum_{i=1}^{N} Z_i\right) \geq \exp(\lambda t)\right)$$
$$\leq \exp(-\lambda t)\mathbb{E}\exp\left(\lambda \sum_{i=1}^{N} Z_i\right) = \exp(-\lambda t)\prod_{i=1}^{N} \mathbb{E}\exp(\lambda Z_i). \tag{3.1}$$

Therefore, by estimating each $\mathbb{E}\exp(\lambda Z_i)$ and optimizing the choice of λ one may derive nontrivial tail bounds.

In what follows we show how certain assumptions on the random variable Z can be used to bound its moment generating function $\mathbb{E}\exp(\lambda Z)$.

3.1 The sum of independent ψ_2 random variables

Let $Z_1, ..., Z_N$ be independent, centred random variables. If $Z_i \in L_{\psi_2}$ and $a = (a_1, ..., a_N) \in \mathbb{R}^N$, what can be said about the tail behaviour of the random variable $\sum_{i=1}^{N} a_i Z_i$?

We answer this question using two different arguments. The first is based on estimates on the moment generating function of Z_i.

Lemma 3.1. *There is an absolute constant c for which the following holds. If $Z \in L_{\psi_2}$ is a centred random variable then for any $\lambda > 0$,*

$$\mathbb{E}\exp(\lambda Z) \leq \exp(c\lambda^2 \|Z\|_{\psi_2}^2).$$

The proof requires a simple idea that appears frequently in what follows—symmetrization, which allows one to replace a mean-zero random variable Z with its symmetric counterpart, εZ, where ε is a symmetric, $\{-1, 1\}$-valued random variable that is independent of Z.

Lemma 3.2. *Let ϕ be a convex function and let Z be a mean-zero random variable. Then*

$$\mathbb{E}\phi(Z) \leq \mathbb{E}\phi(2\varepsilon Z).$$

Proof. Let Z_1, Z_2 be independent copies of Z. By Jensen's inequality,

$$\mathbb{E}_{Z_1}\phi(Z_1) = \mathbb{E}_{Z_1}\phi(Z_1 - \mathbb{E}_{Z_2}Z_2) \leq \mathbb{E}_{Z_1}\mathbb{E}_{Z_2}\phi(Z_1 - Z_2) = (*). \tag{3.2}$$

The random variable $Z_1 - Z_2$ is symmetric, and in particular it has the same distribution as $-1 \cdot (Z_1 - Z_2)$. Therefore, $\mathbb{E}_{Z_1}\mathbb{E}_{Z_2}\phi(Z_1 - Z_2) = \mathbb{E}_{Z_1}\mathbb{E}_{Z_2}\phi(\varepsilon(Z_1 - Z_2))$ for every realization of the symmetric random variable ε that is independent of Z_1 and Z_2. Taking the expectation with respect to ε and using the convexity of ϕ, Fubini's Theorem, and that Z_1 and Z_2 have the same distribution as Z,

$$(*) = \mathbb{E}_\varepsilon\mathbb{E}_{Z_1}\mathbb{E}_{Z_2}\phi(\varepsilon(Z_1 - Z_2)) \leq \mathbb{E}_\varepsilon\mathbb{E}_{Z_1}\mathbb{E}_{Z_2} \cdot \frac{1}{2}\left(\phi(2\varepsilon Z_1) + \phi(2\varepsilon Z_2)\right) = \mathbb{E}\phi(2\varepsilon Z). \tag{3.3}$$

■

In particular, if Z is a mean-zero random variable then $\mathbb{E}\exp(\lambda Z) \leq \mathbb{E}\exp(2\lambda\varepsilon Z)$, and so from here on we may assume, if needed, that Z is a symmetric random variable.

Exercise 9. *Use the same argument as in Lemma 3.2 to show the following: if $Z_1, ..., Z_N$ are independent, mean-zero random variables, then for any $a_1, ..., z_N \in \mathbb{R}$,*

$$\mathbb{E}\phi\left(\sum_{i=1}^N a_i Z_i\right) \leq \mathbb{E}\phi\left(2\sum_{i=1}^N \varepsilon_i Z_i\right), \tag{3.4}$$

where $(\varepsilon_i)_{i=1}^N$ are independent, symmetric, $\{-1, 1\}$-valued random variables that are also independent of $Z_1, .., Z_N$.

Proof of Lemma 3.1. Applying Lemma 3.2 to the convex function $t \to \exp(\lambda t)$,

$$\mathbb{E}\exp(\lambda Z) \leq \mathbb{E}\exp(2\lambda\varepsilon Z) = \mathbb{E}\left(1 + 2\lambda\varepsilon Z + \sum_{j\geq 2}(2\lambda)^j\frac{(\varepsilon Z)^j}{j!}\right) = (*).$$

It is straightforward to verify (e.g., by the Monotone Convergence Theorem for the positive and negative parts of εZ) that

$$(*) \leq 1 + \sum_{j\geq 2}(2\lambda)^{2j}\frac{\mathbb{E}Z^{2j}}{(2j)!} \leq 1 + \sum_{j\geq 2}(2\lambda)^{2j}\frac{\|Z\|_{\psi_2}^j(2cj)^j}{j!j^j};$$

indeed, recall that $\|Z\|_{L_p} \leq c\sqrt{p}\|Z\|_{\psi_2}$ and that $(2j)! \leq j!j^j$. Therefore,

$$\mathbb{E}\exp(\lambda Z) \leq 1 + \sum_{j\geq 2}\frac{1}{j!}\cdot\left(c_1\lambda^2\|Z\|_{\psi_2}^2\right)^j \leq \exp(c_1\lambda^2\|Z\|_{\psi_2}^2).$$

■

Lemma 3.1 leads to the wanted estimate on $\|\sum_{i=1}^N a_i Z_i\|_{\psi_2}$.

Corollary 3.3. *There is an absolute constant c for which the following holds. Let $Z_1, ..., Z_N$ be independent, centred, ψ_2 random variables, and let $a_1, ..., a_N \in \mathbb{R}$. Then*

$$\left\|\sum_{i=1}^N a_i Z_i\right\|_{\psi_2} \leq c\left(\sum_{i=1}^N a_i^2\|Z\|_{\psi_2}^2\right)^{1/2}.$$

Proof. Since a norm is a convex function of its argument, it follows from Exercise 3.4 that one may assume without loss of generality that $a_i \geq 0$ and that each Z_i is symmetric. Fix $\lambda > 0$ to be named later and observe that by Lemma 3.1,

$$\begin{aligned}\mathbb{E}\exp\left(\lambda\sum_{i=1}^N a_i Z_i\right) &= \prod_{i=1}^N \mathbb{E}\exp\left(\lambda a_i Z_i\right) \leq \prod_{i=1}^N \exp\left(c\lambda^2 a_i^2 \|Z_i\|_{\psi_2}^2\right) \\ &= \exp\left(c\lambda^2 \sum_{i=1}^N a_i^2 \|Z_i\|_2^2\right).\end{aligned}$$

Therefore,

$$Pr\left(\sum_{i=1}^N a_i Z_i > t\right) \leq \exp(-\lambda t)\mathbb{E}\exp\left(\lambda\sum_{i=1}^N a_i Z_i\right) \leq \exp\left(-\lambda t + c\lambda^2 \sum_{i=1}^N a_i^2 \|Z_i\|_2^2\right).$$

Setting $\lambda = t/2c\sum_{i=1}^N a_i^2\|Z_i\|_2^2$, it follows that for every $t > 0$,

$$Pr\left(\sum_{i=1}^N a_i Z_i > t\right) \leq \exp\left(-\frac{t^2}{2c\sum_{i=1}^N a_i^2 \|Z_i\|_2^2}\right),$$

and recalling that each Z_i is symmetric, it is evident that

$$Pr\left(\left|\sum_{i=1}^N a_i Z_i\right| > t \cdot \left(c_1 \sum_{i=1}^N a_i^2 \|Z_i\|_{\psi_2}^2\right)^{1/2}\right) \leq 2\exp\left(-t^2\right).$$

The claim follows from the characterization Theorem 2.6. ∎

Remark 3.4. *The argument used in the proof of Corollary 3.3 is almost identical to the proof of Lemma 2.9. This should not come as a surprise, as the latter is a particular instance of the former. Moreover, Corollary 3.3 leads to a more general version of Khintchine's inequality (Theorem 2.10): if $\mathcal{Z} = (Z_1, ..., Z_N)$ that has independent, mean-zero, variance 1 coordinates that satisfy $\max \|Z_i\|_{\psi_2} \leq M$ then*

$$\left\|\sum_{i=1}^N a_i Z_i\right\|_{L_p} \leq cM\sqrt{p}\left\|\sum_{i=1}^N a_i Z_i\right\|_{L_2}.$$

Let us present a different proof of Corollary 3.3, which is based on a comparison argument. The idea is to find $X_1, ..., X_N$ that on the one hand, 'dominate' $Z_1, ..., Z_N$ in an appropriate sense, and at the same time, one can compute (or at least estimate) $\|\sum_{i=1}^N a_i X_i\|_{L_p}$ directly.

Lemma 3.5. *Let $X_1, ..., X_N$ be symmetric, independent random variables and assume that for every $1 \leq i \leq N$ and any $1 \leq p \leq q$, $\|Z_i\|_{L_p} \leq \|X_i\|_{L_p}$. Then for every $a \in \mathbb{R}^N$,*

$$\left\|\sum_{i=1}^N a_i Z_i\right\|_{L_q} \leq c_1 L\left\|\sum_{i=1}^N a_i X_i\right\|_{L_q}.$$

Proof. Without loss of generality, assume that q is an even integer, and by a symmetrization argument assume that $Z_1, ..., Z_N$ are symmetric. Therefore,

$$\mathbb{E}\left(\sum_{i=1}^N a_i Z_i\right)^q = \mathbb{E}\sum c_{\vec{\beta}} \prod_{i=1}^N a_i^{\beta_i} Z_i^{\beta_i} = \sum \prod_{i=1}^N a_i^{\beta_i} \mathbb{E} Z_i^{\beta_i},$$

with the sum taken over all choices of $\vec{\beta} = (\beta_1, ..., \beta_N) \in \{0, ..., q\}$ that sum to q, and $c_{\vec{\beta}}$ is the appropriate multinomial coefficient. Since $Z_1, ..., Z_N$ are symmetric, each product does not vanish only when $\beta_1, ..., \beta_N$ are even, and in that case,

$$\prod_{i=1}^N a_i^{\beta_i} \mathbb{E} Z_i^{\beta_i} \le \prod_{i=1}^N a_i^{\beta_i} L^{\beta_i} \mathbb{E} X_i^{\beta_i}.$$

Therefore,

$$\sum \prod_{i=1}^N a_i^{\beta_i} \mathbb{E} Z_i^{\beta_i} \le L^q \sum \prod_{i=1}^N a_i^{\beta_i} \mathbb{E} X_i^{\beta_i} = L^q \mathbb{E}(\sum_{i=1}^N a_i X_i)^q.$$

■

Proof of Corollary 3.3—version 2. Recall that if $Z_i \in L_{\psi_2}$ then for every $p \ge 1$,

$$\|Z_i\|_{L_p} \le c_0 \sqrt{p} \|Z_i\|_{\psi_2} \le c_1 \|Z_i\|_{\psi_2} \|g\|_{L_p}$$

for a standard gaussian variable g. Therefore, if $g_1, ..., g_N$ are independent standard gaussian random variables and selecting $X_i = c_1 \|Z_i\|_{\psi_2} g_i$ in Lemma 3.5, it is evident that for every $p \ge 1$ and every $a \in \mathbb{R}^N$,

$$\left\| \sum_{i=1}^N a_i Z_i \right\|_{L_p} \le \left\| \sum_{i=1}^N a_i X_i \right\|_{L_p} = c_1 \left\| \sum_{i=1}^N a_i \|Z_i\|_{\psi_2} g_i \right\|_{L_p}.$$

Finally, using the rotation invariance of the standard gaussian vector, it is evident that $\sum_{i=1}^N a_i \|Z_i\|_{\psi_2} g_i$ has the same distribution as $(\sum_{i=1}^N a_i^2 \|Z_i\|_{\psi_2}^2)^{1/2} g$; therefore,

$$\left\| \sum_{i=1}^N a_i \|Z_i\|_{\psi_2} g_i \right\|_{L_p} \le c_2 \left(\sum_{i=1}^N a_i^2 \|Z_i\|_{\psi_2}^2 \right)^{1/2} \sqrt{p},$$

and the claim follows. ■

3.2 Bernstein type inequalities

Let us return to more applications of (3.1), leading to *Bernstein type inequalities.*

Lemma 3.6. *Let Z be a mean-zero random variable, and assume that there are constants M and σ such that, for every integer $p \ge 2$,*

$$\mathbb{E}|Z|^p \le p! \cdot M^{p-2} \sigma^2.$$

Then for every $0 < \lambda \le 1/2M$,

$$\mathbb{E} \exp(\lambda Z) \le 1 + 2\lambda^2 \sigma^2 \le \exp(2\lambda x).$$

Before proving the lemma, let us examine the condition on the growth rate of the moments of Z—which fits two standard situations.

- Let Z be a bounded random variable and set $M = \|Z\|_{L_\infty}$ and $\sigma^2 = \mathbb{E}Z^2$. Observe that $\mathbb{E}Z^p \leq \|Z\|_{L_\infty}^{p-2}\mathbb{E}Z^2 = M^{p-2}\sigma^2$, which is far better than the required condition (there is no additional factor of $p!$).
- Let $Z \in L_{\psi_1}$ and recall that by Theorem 2.6, $\|Z\|_{L_p} \leq cp\|Z\|_{\psi_1}$ for a suitable absolute constant c. Since $p^p \leq e^p \cdot p!$ (e.g. by induction), one may select $\sigma = M = ec\|Z\|_{\psi_1}$.

Proof. Using Taylor's expansion, $\exp(x) = \sum_{i=0}^{\infty} x^p/p!$, and by a standard argument, (e.g. the monotone convergence theorem applied to the positive and negative parts of Z),

$$\mathbb{E}\exp(\lambda Z) = \mathbb{E}\sum_{p=0}^{\infty}\frac{(\lambda Z)^p}{p!} = 1 + \sum_{p=1}^{\infty}\frac{\lambda^p\mathbb{E}Z^p}{p!} = (*).$$

Recall that $\mathbb{E}Z = 0$ and that $\mathbb{E}|Z|^p \leq p!M^{p-2}\sigma^2$; therefore, since $\lambda M \leq 1/2$,

$$(*) \leq 1 + \frac{\sigma^2}{M^2}\sum_{p=2}^{\infty}(\lambda M)^p \leq 1 + 2\lambda^2\sigma^2.$$

The claim follows because $1 + x^2/2 \leq \exp(x)$. ∎

With Lemma 3.6 in place, one can derive various versions of Bernstein's inequality.

Theorem 3.7. *Let $Z_1, ..., Z_N$ be independent, mean-zero random variables and assume that there are constants M and σ_i, $i = 1, ..., N$, such that, for every $1 \leq i \leq N$, $\mathbb{E}|Z_i|^p \leq p!M^{p-2}\sigma_i^2$. If $S^2 = \sum_{i=1}^N \sigma_i^2$ then for every $t > 0$,*

$$Pr\left(\sum_{i=1}^N Z_i > t\right) \leq \exp\left(-\min\left\{\frac{t^2}{8S^2}, \frac{t}{4M}\right\}\right).$$

Proof. Combining (3.1) and Lemma 3.6, it follows that for $0 < \lambda \leq 1/2M$,

$$\prod_{i=1}^N \mathbb{E}\exp(\lambda Z_i) \leq \prod_{i=1}^N \exp(2\lambda^2\sigma_i^2) = \exp(2\lambda^2 S^2).$$

Therefore,

$$Pr\left(\sum_{i=1}^N Z_i > t\right) \leq \exp(-\lambda t + 2\lambda^2 S^2) \leq \exp(-\lambda t/2),$$

provided that $2\lambda^2 S^2 \leq \lambda t/2$, i.e., $\lambda \leq t/4S^2$. Let

$$\lambda = \min\left\{\frac{t}{4S^2}, \frac{1}{2M}\right\}$$

and thus

$$Pr\left(\sum_{i=1}^N Z_i > t\right) \leq \exp\left(-\min\left\{\frac{t^2}{8S^2}, \frac{t}{4M}\right\}\right).$$

∎

Remark 3.8. *The constants appearing in Theorem 3.7 are not optimal, though this will be of no importance in what follows. Because of that, from here on we will replace the constants appearing in Theorem 3.7 with an unspecified constant c.*

Corollary 3.9. *There exists an absolute constant c for which the following holds. Let $Z_1, ..., Z_N$ be independent, mean-zero random variables.*

- *If $Z_1, ..., Z_N$ are bounded almost surely by M (i.e., $\max_{1\leq i\leq N} \|Z_i\|_{L_\infty} \leq M$), and $\sigma_i^2 = \mathbb{E}Z_i^2$, then for every $t > 0$,*

$$Pr\left(\sum_{i=1}^N Z_i > t\right) \leq \exp\left(-c\min\left\{\frac{t^2}{\sum_{i=1}^N \sigma_i^2}, \frac{t}{M}\right\}\right).$$

- *If $Z_1, ..., Z_N$ are uniformly bounded in L_{ψ_1}, i.e., if $\max_{1\leq i\leq M} \|Z_i\|_{\psi_1} \leq M$, then*

$$Pr\left(\sum_{i=1}^N Z_i > t\right) \leq \exp\left(-c\min\left\{\frac{t^2}{NM^2}, \frac{t}{M}\right\}\right).$$

In particular, if $Z_1, ..., Z_N$ are also identically distributed as Z then

$$Pr\left(\frac{1}{N}\sum_{i=1}^N Z_i > t\right) \leq \exp\left(-cN\min\left\{\frac{t^2}{\|Z\|_{\psi_1}^2}, \frac{t}{\|Z\|_{\psi_1}}\right\}\right).$$

3.3 Sum of squares of ψ_2 random variables

For reasons that will become clear in what follows, there is a special interest in the behaviour of a sum of squares of independent random variables

$$\frac{1}{N}\sum_{i=1}^N Z_i^2.$$

One can establish a rather accurate description of the behaviour of this average, highlighting the difference between the upper estimate and the lower one and way the estimates are connected with various properties of the random variables Z_i. However, for the time being, let us obtain a two-sided estimate of the from

$$Pr\left(\left|\frac{1}{N}\sum_{i=1}^N Z_i^2 - \mathbb{E}Z^2\right| > t\right)$$

when $Z_1, ..., Z_N$ are independent copies of a random variable Z that is L-subgaussian; that is, it satisfies $\|Z\|_{\psi_2} \leq L\|Z\|_{L_2}$.

The starting point is the straightforward observation that $\|Z^2\|_{\psi_1} = \|Z\|_{\psi_2}^2$ which follows from the definition of the ψ_α norms. Also, because $\|\cdot\|_{\psi_\alpha}$ are norms for $1 \leq \alpha \leq 2$,

$$\|Z^2 - \mathbb{E}Z^2\|_{\psi_1} \leq \|Z^2\|_{\psi_1} + \|\mathbb{E}Z^2\|_{\psi_1} \leq 2\|Z^2\|_{\psi_1};$$

indeed, by the definition of the ψ_α norm and Jensen's inequality it is evident that for a random variable Y, $\|\mathbb{E}Y\|_{\psi_\alpha} \leq \|Y\|_{\psi_\alpha}$.

Thus, the random variable $Z^2 - \mathbb{E}Z^2$ is centred and $\|Z^2 - \mathbb{E}Z^2\|_{\psi_1} \leq 2\|Z\|_{\psi_2}^2$. Applying the ψ_1 version of Bernstein's inequality,

$$Pr\left(\left|\frac{1}{N}\sum_{i=1}^N Z_i^2 - \mathbb{E}Z^2\right| > t\right) \leq 2\exp\left(-cN\min\left\{\frac{t^2}{\|Z\|_{\psi_2}^4}, \frac{t}{\|Z\|_{\psi_2}^2}\right\}\right).$$

Setting $t = \varepsilon\mathbb{E}Z^2$, it follows that with probability at least

$$1 - 2\exp\left(-cN\min\left\{\varepsilon^2\left(\frac{\|Z\|_{L_2}}{\|Z\|_{\psi_2}}\right)^4, \varepsilon\left(\frac{\|Z\|_{L_2}}{\|Z\|_{\psi_2}}\right)^2\right\}\right), \tag{3.5}$$

$$(1-\varepsilon)\mathbb{E}Z^2 \leq \frac{1}{N}\sum_{i=1}^N Z_i^2 \leq (1+\varepsilon)\mathbb{E}Z^2. \tag{3.6}$$

Now recall that $\|Z\|_{\psi_2}/\|Z\|_{L_2} \leq L$ for some $L \geq 1$. Hence, if $0 < \varepsilon < 1$ then (3.5) becomes

$$1 - 2\exp\left(-cN\min\left\{\frac{\varepsilon^2}{L^4}, \frac{\varepsilon}{L^2}\right\}\right) = 1 - 2\exp(-c_1(L)\varepsilon^2 N). \tag{3.7}$$

Thus, on a high probability event, $N^{-1}\sum_{i=1}^N Z_i^2$ is almost-isometrically equivalent to $\mathbb{E}Z^2$.

As an example, let X be an isotropic, L-subgaussian random vector in $\mathbb{R}^d$.

Definition 3.10. *A random vector X taking values in $\mathbb{R}^d$ is L-subgaussian if it is symmetric and for every $t \in \mathbb{R}^d$,*

$$\|\langle X, t\rangle\|_{\psi_2} \leq L\|\langle X, t\rangle\|_{L_2}.$$

Thus, for any $t \in \mathbb{R}^d$, the random variable $Z = \langle t, X\rangle$ is exactly as described previously: it is mean-zero by the symmetry of X; it satisfies $\mathbb{E}\langle X, t\rangle^2 = \|t\|_2^2$ because X is isotropic; and $\|\langle t, X\rangle\|_{\psi_2} \leq L\|\langle X, t\rangle\|_{L_2}$ because X is an L-subgaussian random vector.

Let $T \subset \mathbb{R}^n$ be a finite set. Let $X_1, ..., X_N$ be independent copies of X, and consider the random matrix

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

The following result is the celebrated Johnson-Lindenstrauss embedding lemma:

Lemma 3.11. *For $L \geq 1$ there exist constants c_1 and c_2 that depend only on L and for which the following holds. If $T \subset \mathbb{R}^d$ is a finite set, $0 < \varepsilon < 1$ and $N \geq c_1\varepsilon^{-2}\log|T|$, then with probability at least $1 - 2\exp(-c_2\varepsilon^2 N)$, for every $x, y \in T$,*

$$(1-\varepsilon)\|x-y\|_2^2 \leq \|\Gamma(x-y)\|_2^2 \leq (1+\varepsilon)\|x-y\|_2^2.$$

In other words, Lemma 3.11 implies that the random operator Γ almost preserves distances in T, as long as there is 'enough randomness'—sufficiently many independent copies of the random vector X, which serve as the rows of the random matrix Γ.

Remark 3.12. *As will be clarified later,* $\log|T|$ *is a rather crude measure of complexity for a set, and one may obtain better estimates on the number of rows required (i.e., the number of sample points/linear measurements needed), as well as improved understanding on the way a subgaussian operator acts on a set. Moreover, we show that the lower estimate*

$$(1-\varepsilon)\|x-y\|_2^2 \leq \|\Gamma(x-y)\|_2^2$$

holds in far more general situations than the one considered here.

Proof. Observe that for every $t_1, t_2 \in T$,

$$\|\Gamma(t_1-t_2)\|_2^2 = \frac{1}{N}\sum_{i=1}^N \langle X_i, t_1-t_2\rangle^2$$

and obviously, $\mathbb{E}\langle X_i, t_1-t_2\rangle^2 = \|t_1-t_2\|_2^2$. Hence, for $0<\varepsilon<1$, with probability at least $1-2\exp(-c(L)\varepsilon^2 N)$,

$$(1+\varepsilon)\|t_1-t_2\|_2^2 \leq \|\Gamma(t_1-t_2)\|_2^2 \leq (1+\varepsilon)\|t_1-t_2\|_2^2. \tag{3.8}$$

The number of different pairs of distinct points in T is at most $|T|^2$, and by the union bound, with probability at least

$$1-2|T|^2\exp(-c(L)\varepsilon^2 N),$$

for every $t_i \neq t_j$, $t_i, t_j \in T$, one has

$$(1+\varepsilon)\|t_i-t_j\|_2^2 \leq \|\Gamma(t_i-t_j)\|_2^2 \leq (1+\varepsilon)\|t_i-t_j\|_2^2. \tag{3.9}$$

Therefore, if $N \geq c_1(L)\varepsilon^{-2}\log|T|$ then (3.9) holds with probability at least $1-2\exp(-c_2(L)\varepsilon^2 N)$. ■

3.4 Bennett's Inequality

Let us return to the first part of Corollary 3.9, which deals with bounded random variables. As can be immediately seen from the estimates on the moment generating function in that case, there is some room for maneuvering, since $\mathbb{E}|Z|^p \leq M^{p-2}\sigma^2$—without the additional factor of $p!$. This observation is at the heart of the following improvement to Bernstein's inequality in the bounded case, called *Bennett's inequality.*

Before formulating and proving that inequality, let us improve the estimate on the moment generating function of a bounded random variable.

Lemma 3.13. *Let* Z *be a centred random variable. If* $\|Z\|_{L_\infty} \leq M$ *then for every* $\lambda > 0$*,*

$$\mathbb{E}\exp(\lambda Z) \leq 1 + \frac{\mathbb{E}Z^2}{M^2}\left(\exp(\lambda M) - 1 - \lambda M\right).$$

Proof. Using the same argument as in Lemma 3.6 and recalling that $\mathbb{E}Z = 0$ and that $\mathbb{E}|Z|^p \leq M^{p-2}\mathbb{E}Z^2$,

$$\mathbb{E}\exp(\lambda Z) \leq 1 + \frac{\mathbb{E}Z^2}{M^2}\sum_{p=2}^\infty \frac{\lambda^p M^p}{p!} = 1 + \frac{\mathbb{E}Z^2}{M^2}\left(\exp(\lambda M) - 1 - \lambda M\right),$$

where the final step follows from Taylor's expansion of $\exp(x)$ and the choice $x = \lambda M$. ■

Corollary 3.14. *Let $Z_1, ..., Z_N$ be independent mean-zero random variables. Assume that $\max_{1\leq i\leq M} \|Z_i\|_{L_\infty} \leq M$, set $S^2 = \sum_{i=1}^N \mathbb{E}Z_i^2$ and put*

$$\Phi(x) = (1+x)\log(1+x) - x.$$

Then, for every $t > 0$,

$$Pr\left(\sum_{i=1}^N Z_i \geq t\right) \leq \exp\left(-\frac{S^2}{M^2}\Phi\left(\frac{tM}{S^2}\right)\right).$$

Proof. Combining (3.1) and Lemma 3.13, it follows that

$$\exp(-t\lambda)\prod_{i=1}^N \mathbb{E}\exp(\lambda Z_i) \leq \exp\left(-\lambda t + \frac{S^2}{M^2}\left(\exp(\lambda M) - 1 - \lambda M\right)\right) = (*).$$

One may optimize the choice of λ and verify that the minimum is attained for

$$\lambda = \frac{1}{M}\log\left(\frac{tM}{S^2} + 1\right),$$

which yields the desired bound. ∎

To understand the meaning of Corollary 3.14, one should study the behaviour of the function $\Phi(x)$. It is straightforward to verify that for $x \geq 1$, $\Phi(x) \geq (1/2)x\log(1+x)$, and that for $0 < x < 1$, $\Phi(x) \geq x^2/4$. Thus, the tail behaviour changes according to the value of t: when $0 < t \leq S^2/M$, the sum exhibits a subgaussian behaviour, and when $t > S^2/M$ the tail behaviour is better than $\sim \exp(-t)$: thanks to the extra logarithmic term, the tail is actually similar to that of a Poisson random variable.

As an example, let $Z_1, ..., Z_N$ be identically distributed according to the bounded, centred random variable Z. Then $S^2 = N\mathbb{E}Z^2 = N\sigma^2$ and therefore,

$$Pr\left(\frac{1}{N}\sum_{i=1}^N Z_i > t\right) \leq \exp\left(-\frac{Nt}{M} - \left(\frac{Nt}{M} + \frac{N\sigma^2}{M^2}\right)\log\left(1 + \frac{Mt}{\sigma^2}\right)\right).$$

If $0 < t \leq N\sigma^2/M$ then the tail is smaller than

$$\exp(-t^2/2N\sigma^2), \tag{3.10}$$

while for every $t > 0$, it is smaller than

$$\exp\left(-\frac{t}{M}\left(\log\left(1 + \frac{Mt}{N\sigma^2}\right)\right)\right). \tag{3.11}$$

One very useful application of Bennett's inequality is for the sum of independent selectors – which are simply $\{0,1\}$-valued random variables. Indeed, let $(\delta_i)_{i=1}^N$ be independent, taking values in $\{0,1\}$ and assume that for every $1 \leq i \leq N$, $\mathbb{E}\delta_i = \delta$. Set $Z_i = \delta_i - \delta$ and note that $\mathbb{E}Z_i = 0$, $\|Z_i\|_\infty = 1$ and $\mathbb{E}Z_i^2 = \delta - \delta^2 \geq \delta/4$ provided that $\delta \leq 3/4$. Therefore, applying (3.10) and (3.11) to the random variables Z_i and then to the random variables $-Z_i$, the following is evident: for $t \leq N\delta/4$,

$$Pr\left(\left|\sum_{i=1}^N (\delta_i - \delta)\right| > t\right) \leq 2\exp\left(-c\frac{t^2}{N\delta}\right) \tag{3.12}$$

and for $t \geq N\delta/4$,

$$Pr\left(\left|\sum_{i=1}^{N}(\delta_i - \delta)\right| > t\right) \leq 2\exp\left(-ct\left(\log\left(1 + \frac{t}{N\delta}\right) - 1\right)\right). \tag{3.13}$$

Thus, with probability at least $1 - 2\exp(-c_1\delta N)$,

$$\frac{1}{2}\delta N \leq |\{i : \delta_i = 1\}| \leq \frac{3}{2}\delta N \tag{3.14}$$

and for $u \geq 2$, with probability at least $1 - 2\exp(-c_1\delta N u \log u)$,

$$|\{i : \delta_i = 1\}| \leq u\delta N.$$

3.5 Symmetrization of Empirical Processes

We end this chapter with a symmetrization argument, due to Giné and Zinn [**?**], which shows that $\mathbb{E}\sup_{f\in F}|N^{-1}\sum_{i=1}^{N} f(X_i) - \mathbb{E}f|$ is equivalent to $N^{-1}\mathbb{E}\sup_{f\in F}|\sum_{i=1}^{N}\varepsilon_i f(X_i)|$. This generalizes the simple symmetrization arguments described previously, and it is one of the main technical tools used in the study of empirical processes.

Theorem 3.15. *There are absolute constants C_1 and C_2 for which the following holds. Let F be a class of functions on (Ω, μ) and let X be distributed according to μ. If $X_1, ..., X_N$ are independent copies of X then*

$$\mathbb{E}\sup_{f\in F}\left|\frac{1}{N}\sum_{i=1}^{N} f(X_i) - \mathbb{E}f\right| \leq \frac{C_1}{N}\mathbb{E}\sup_{f\in F}\left|\sum_{i=1}^{N}\varepsilon_i f(X_i)\right| \leq C_2\mathbb{E}\sup_{f\in F}\left|\frac{1}{N}\sum_{i=1}^{N} f(X_i) - \mathbb{E}f\right| + \frac{r}{\sqrt{N}},$$

where $r = \sup_{f\in F}|\mathbb{E}f(X)|$.

Proof. Let $(Y_i)_{i=1}^{N}$ be an independent copy of $(X_i)_{i=1}^{N}$. Note that

$$\mathbb{E}_X\sup_{f\in F}\left|\frac{1}{N}\sum_{i=1}^{N} f(X_i) - \mathbb{E}_Y f\right| = \mathbb{E}_X\sup_{f\in F}\left|\frac{1}{N}\sum_{i=1}^{N} f(X_i) - \mathbb{E}_Y f - \mathbb{E}_Y\left(\frac{1}{N}\sum_{i=1}^{N} f(Y_i) - \mathbb{E}_Y f\right)\right|$$

Conditioning on $X_1, ..., X_N$ followed by Jensen's inequality with respect to $\mathbb{E}_Y$ and then Fubini's Theorem, one has

$$\begin{aligned}\mathbb{E}\sup_{f\in F}\left|\frac{1}{N}\sum_{i=1}^{N} f(X_i) - \mathbb{E}f\right| &\leq \frac{1}{N}\mathbb{E}_X\mathbb{E}_Y\sup_{f\in F}\left|\sum_{i=1}^{N} f(X_i) - \sum_{i=1}^{N} f(Y_i)\right| \\ &= \frac{1}{N}\mathbb{E}_X\mathbb{E}_Y\sup_{f\in F}\left|\sum_{i=1}^{N}\varepsilon_i\left(f(X_i) - f(Y_i)\right)\right|,\end{aligned}$$

where the final equality holds for every $(\varepsilon_i)_{i=1}^{n} \in \{-1, 1\}^N$. Taking the expectation with respect to $(\varepsilon_i)_{i=1}^{N}$ and by the triangle inequality,

$$\frac{1}{N}\mathbb{E}_X\mathbb{E}_Y\mathbb{E}_\varepsilon\sup_{f\in F}\left|\sum_{i=1}^{N}\varepsilon_i\left(f(X_i) - f(Y_i)\right)\right| \leq \frac{2}{N}\mathbb{E}_X\mathbb{E}_\varepsilon\sup_{f\in F}\left|\sum_{i=1}^{N}\varepsilon_i f(X_i)\right| = \frac{2}{N}\mathbb{E}\sup_{f\in F}\left|\sum_{i=1}^{N}\varepsilon_i f(X_i)\right|.$$

To prove the upper bound, one may apply the triangle inequality,

$$\frac{1}{N}\mathbb{E}_X\mathbb{E}_\varepsilon \sup_{f\in F}\left|\sum_{i=1}^N \varepsilon_i f(X_i)\right| \le \frac{1}{N}\mathbb{E}_X\mathbb{E}_\varepsilon \sup_{f\in F}\left|\sum_{i=1}^N \varepsilon_i \left(f(X_i)-\mathbb{E}f\right)\right| + \left|\sup_{f\in F}\mathbb{E}f\right|\cdot \mathbb{E}_\varepsilon\left|\frac{1}{N}\sum_{i=1}^N \varepsilon_i\right|.$$

Let $(Z_i)_{i=1}^N$ be the stochastic process defined by $Z_i(f) = f(X_i) - \mathbb{E}f$ and let $(W_i)_{i=1}^N$ be an independent copy of $(Z_i)_{i=1}^N$. For every $f \in F$, $\mathbb{E}W_i(f) = 0$, thus,

$$\begin{aligned}\mathbb{E}_X\mathbb{E}_\varepsilon \sup_{f\in F}\left|\sum_{i=1}^N \varepsilon_i \left(f(X_i)-\mathbb{E}f\right)\right| &= \mathbb{E}_Z\mathbb{E}_\varepsilon \sup_{f\in F}\left|\sum_{i=1}^N \varepsilon_i Z_i(f)\right| \\ &= \mathbb{E}_\varepsilon\mathbb{E}_Z \sup_{f\in F}\left|\sum_{i=1}^N \varepsilon_i \left(Z_i(f)-\mathbb{E}_W W_i(f)\right)\right|.\end{aligned}$$

For every realization of the Bernoulli random variables $(\varepsilon_i)_{i=1}^n$ and by Jensen's inequality conditioned on Z_i,

$$\mathbb{E}_Z \sup_{f\in F}\left|\sum_{i=1}^N \varepsilon_i \left(Z_i(f)-\mathbb{E}_W W_i(f)\right)\right| \le \mathbb{E}_Z\mathbb{E}_W \sup_{f\in F}\left|\sum_{i=1}^N \varepsilon_i \left(Z_i(f)-W_i(f)\right)\right|,$$

which is invariant for any selection of signs $(\varepsilon_i)_{i=1}^N$. Therefore,

$$\begin{aligned}\mathbb{E}_\varepsilon\mathbb{E}_Z \sup_{f\in F}\left|\sum_{i=1}^N \varepsilon_i \left(Z_i(f)-\mathbb{E}_W W_i(f)\right)\right| &\le \mathbb{E}_Z\mathbb{E}_W \sup_{f\in F}\left|\sum_{i=1}^N \left(Z_i(f)-W_i(f)\right)\right| \\ &\le 2\mathbb{E}_Z \sup_{f\in F}\left|\sum_{i=1}^n Z_i(f)\right|,\end{aligned}$$

as claimed. ∎

Next, let us turn to an 'in-probability' symmetrization argument. The proof can be found in [**?**].

Theorem 3.16. *Let $(Z_i)_{i=1}^N$ be an iid stochastic process, that is mean-zero (i.e., for every $f \in F$, $\mathbb{E}Z(f) = 0$). For every $1 \le i \le N$, set $h_i : F \to \mathbb{R}$ to be an arbitrary function. Then, for every $x > 0$*

$$\begin{aligned}\left(1-\frac{4N}{x^2}\sup_{f\in F}\operatorname{var}\left(Z_1(f)\right)\right) Pr\left(\sup_{f\in F}\left|\sum_{i=1}^N Z_i(f)\right| > x\right) \\ \le 2Pr\left(\sup_{f\in F}\left|\sum_{i=1}^N \varepsilon_i\left(Z_i(f)-h_i(f)\right)\right| > \frac{x}{4}\right).\end{aligned}$$

Let us consider its implications to the standard empirical process. Set $Z_i(f) = f(X_i) - \mathbb{E}f$ and put $h_i(f) = -\mathbb{E}f$. Also, set $v^2 = \sup_{f\in F}\operatorname{var}(f)$, and note that if $x \ge 2\sqrt{2}\sqrt{N}v$ then

$$1-\frac{4N}{x^2}\sup_{f\in F}\operatorname{var}\left(Z_1(f)\right) \ge 1/2.$$

Therefore, for such a choice of x,

$$Pr\left(\sup_{f\in F}\left|\sum_{i=1}^{N}(f(X_i)-\mathbb{E}f)\right|>x\right)\leq 4Pr\left(\sup_{f\in F}\left|\sum_{i=1}^{N}\varepsilon_i f(X_i)\right|>\frac{x}{4}\right).$$

Now, fix any $\varepsilon>0$ and let $x=N\varepsilon$. If $N\geq 8v^2/\varepsilon^2$ then

$$Pr\left(\sup_{f\in F}\left|\frac{1}{N}\sum_{i=1}^{N}f(X_i)-\mathbb{E}f\right|>\varepsilon\right)\leq 4Pr\left(\sup_{f\in F}\left|\sum_{i=1}^{N}\varepsilon_i f(X_i)\right|>\frac{N\varepsilon}{4}\right). \tag{3.15}$$

Theorem 3.15 and Theorem 3.16 give a different way of addressing questions related to empirical processes. Instead of analyzing the collection of the random variables $\{Z_f : f\in F\}$, where each Z_f is the empirical mean (centred or not) of f, one can study the behaviour of a random subset of $\mathbb{R}^N$, the *random coordinate projection*, defined for a given F and a sample $\sigma=(X_i)_{i=1}^N$ by

$$P_\sigma F=\left\{(f(X_i))_{i=1}^N \ : \ f\in F\right\}.$$

By a symmetrization argument, the expectation of the supremum of the empirical process indexed by F is equivalent to

$$\mathbb{E}_X\left(\mathbb{E}_\varepsilon\sup_{v\in P_\sigma F}\left|\sum_{i=1}^{N}\varepsilon_i v_i\right|\right),$$

and understanding objects like $\mathbb{E}_\varepsilon\sup_{t\in T}\left|\sum_{i=1}^N\varepsilon_i t_i\right|$ and the way they depend on the indexing set T takes one a significant step forward towards a sharp analysis of empirical processes.

A significant part in these notes is the long journey towards a better understanding of the connections between $\mathbb{E}_\varepsilon\sup_{t\in T}\left|\sum_{i=1}^N\varepsilon_i t_i\right|$ and the geometry of the indexing set T. A crucial part of this journey is the study of how much of the structure of the indexing class F is reflected in the geometry of a typical $P_\sigma F$.

Chapter 4

Vectors with iid coordinates

One of the most significant objects that appears in what follows is a random vector in $\mathbb{R}^N$ whose coordinates are independent copies of a fixed random variable. This is a natural feature of sampling: given a function f one 'sees' the sample $(f(X_i))_{i=1}^N$ and uses the sample to address statistical questions on properties of f. In other words, the belief behind sampling is that with high probability $(f(X_i))_{i=1}^N$ encodes information that is important in the context of statistical questions.

Of course, another way of looking at the same object is as a random vector in $\mathbb{R}^N$ with iid coordinates. If the idea behind sampling is indeed true, then the location of a typical realization of the random vector $(f(X_i))_{i=1}^N$ must contain important information. That leads us to the following fundamental question:

Question 4.1. *If Z is a random variable and $Z_1, ..., Z_d$ are independent copies of Z, what is the typical location of the random vector $\mathcal{Z} = (Z_1, ..., Z_d)$?*

As outlined in Section 1.1, the two notions of location we focus on are when $\mathcal{Z}$ is viewed as a function on $\mathbb{R}^d$ endowed with the natural probability measure; the other is when $\mathcal{Z}$ endows a probability measure on $\mathbb{R}^d$, and we study the linear functionals $\langle t, \cdot \rangle$ defined on that probability space. Specifically,

- What is the typical behaviour of $\|\mathcal{Z}\|$ for various natural norms on $\mathbb{R}^d$, like L_p^d and ψ_α^d? And does $\mathcal{Z}$ necessarily have many 'nontrivial coordinates'?
- Given $t \in \mathbb{R}^d$, what are the moments of the linear form $\langle \mathcal{Z}, t \rangle$? In a more geometric language, what is the structure of the set

$$B(L_p(\mathcal{Z})) = \{t \in \mathbb{R}^d : \|\langle \mathcal{Z}, t \rangle\|_{L_p} \leq 1\}?$$

Both notions of location will turn out to be very important in what follows, and at times, the study of the location can be highly nontrivial. At this point, the machinery at our disposal is somewhat limited, and Question 4.1 will accompany us for a while. Still, some very useful facts can still be derived, and we present some of them in what follows.

To get a better feeling of what is going on, we study two special cases: when Z is the standard gaussian random variable, implying that $\mathcal{Z} = (g_1, ..., g_d)$ is the standard gaussian random vector in $\mathbb{R}^d$ (denoted in what follows by G; and when Z is a symmetric, $\{-1, 1\}$-valued random variable, implying that $\mathcal{Z} = (\varepsilon_1, ..., \varepsilon_d)$—the uniform distribution on the d-dimensional combinatorial cube. In what follows we refer to this vector as the (standard)

Bernoulli vector and denote it by $\mathcal{E}$. Before getting one's hopes too high, even in these very special cases we will be far from a complete answer.

4.1 The gaussian vector in $\mathbb{R}^d$

The standard gaussian vector is of extreme significance for the theory we present. We do not present all its basic properties in this presentation.

A very useful fact is that the gaussian random vector is rotation-invariant, meaning that for any measurable set A and any orthogonal matrix O, $Pr(G \in A) = Pr(OG \in A)$ i.e., the distribution of G and of OG coincide.

Two facts immediately follow from this observation:

- For $t \in \mathbb{R}^d$, $\langle t, G\rangle$ has the same distribution at $\|t\|_2 g$: indeed, let O be the orthogonal matrix that satisfies $O^{-1}t = e_1\|t\|_2$. Since G and OG have the same distribution then $\langle t, G\rangle$ is distributed as $\langle t, OG\rangle = \langle O^{-1}t, G\rangle = \|t\|_2 g_1$, as required.
- Let $\mathcal{X} = G/\|G\|_2$. Then $\mathcal{X}$ is a probability measure on the sphere S^{d-1} which is invariant to rotations: for every orthogonal operator O and a measurable set $A \subset S^{d-1}$,

$$Pr(\mathcal{X} \in OA) = Pr(O^{-1}G/\|O^{-1}G\|_2 \in A) = Pr(G/\|G\|_2 \in A),$$

  because G and $\|G\|_2$ are rotation invariant. A deep fact is that the there is only one probability measure on the sphere that is rotation invariant—the surface area. This is an example of the notion of the Haar measure.

Thanks to the first observation, we can identify the L_p structure endowed on $\mathbb{R}^d$ by G: for any $t \in \mathbb{R}^d$, $\langle t, G\rangle$ has the same distribution as $g_1\|t\|_2$, and therefore, by the standard estimate on moments of a gaussian random variable, $\|\langle t, G\rangle\|_{L_p} = \|t\|_2\|g\|_{L_p} = c\sqrt{p}\|t\|_2$. We thus have that the L_p unit ball endowed on $\mathbb{R}^d$ by the gaussian vector,

$$B(L_p(G)) = \left\{t \in \mathbb{R}^d : \|\langle t, G\rangle\|_{L_p} \le 1\right\} = \frac{1}{c\sqrt{p}} B_2^d$$

and is just a re-scaling of the standard Euclidean ball.

A far more subtle question is the 'location' in $\mathbb{R}^d$ of a typical realization of G. Let us start by a simple question: examining the Euclidean norm of G. It turns out that typically, G 'lives' in some well specified shell in $\mathbb{R}^d$. While this may sound strange at first, it becomes straightforward once presented in a more probabilistic language: the event $\{\alpha \le \|G\|_2 \le \beta\}$ is simply $\{\alpha^2 \le \sum_{i=1}^d g_i^2 \le \beta^2\}$, and we have obtained sharp estimates on the sum of independent subgaussian random variables in Chapter 3.3. Since $\|g^2\|_{\psi_1} = \|g\|_{\psi_2}^2$ it follows by the ψ_1 version of Bernstein's inequality that with probability at least $1 - 2\exp(-c_0 d \min\{u^2, u\})$,

$$\left|\frac{\|G\|_2^2}{d} - 1\right| \le u;$$

In other words,

$$\|G\|_2 \in (\sqrt{d} - c_1\sqrt{m}, \sqrt{d} + c_1\sqrt{m}) \tag{4.1}$$

with probability at least $1 - 2\exp(-c_2 m)$.

Even this sharp shell bound does not give precise enough information on the location of G: as noted previously, from the perspective of sampling, the vector $(\sqrt{d}, 0, ..., 0)$ is (at least intuitively) terrible because it is 'peaky', while $(1, ..., 1)$ is great—being 'well spread'. However, both these vectors belong to the shell (4.1).

To complement the shell picture, we need to derive more information on the distribution of the coordinates of G, and show that a typical G has to be well spread. Recall that in the sense that a proportional number of its coordinates are of the order of $\|G\|_2/\sqrt{d}$,

Lemma 4.2. *There exists absolute constants c_1 and c_2 for which the following holds: with probability at least $1 - 2\exp(-c_1 d)$*

$$\left| \left\{ i : |g_i| \geq c_2 \frac{\|G\|_2}{\sqrt{d}} \right\} \right| \geq 0.98d.$$

One useful fact is that a standard gaussian variable satisfies a small-ball property: for a suitable absolute constant κ, $Pr(|g| \geq \kappa) \geq 0.99$. Therefore, with probability at least $1 - 2\exp(-c_0 d)$,

$$|\{i : |g_i| \geq \kappa\}| \geq 0.98d,$$

which means that typically, G is well-spread: there is $\sim d$ coordinates larger than $\sim \|G\|_2/\sqrt{d} \geq c_1$ which by (4.1) holds with probability at least $1 - 2\exp(-c_2 d)$.

Note that the coordinate small-ball estimate does not exclude the fact that many of the coordinates of G are big — say that there a constant number of them is of the order of $\sqrt{d}$. To exclude that option one has to obtain an upper estimate on $\|G\|$ for norms that are more restrictive that the ℓ_2 one.

Lemma 4.3. *There are absolute constants c_1 and c_2 for which the following holds: $\|E\|G\|_{\psi_2^d} \leq c_1$, and for $u > 4$, $Pr(\|G\|_{\psi_2^d} \geq u) \leq \exp(-c_2 u^2 \log d)$.*

Before we prove Lemma 4.3, let us explain its meaning. As observed previously, with probability at least $1 - 2\exp(-c_0 d)$, $d/2 \leq \|G\|_2 \leq 2d$ and G has $d/2$ coordinates that are larger than a suitable absolute constant c_1, and, in fact, that a proportional number of the coordinates belong to an interval $[c_2, c_3]$. Now, by Lemma 4.3 we have that with super-polynomial probability,

$$g_i^* \leq c_2 \sqrt{\log(ed/i)},$$

showing that the coordinates of $(g_i)_{i=1}^d$ cannot be too big: $\|G\|\infty \leq c_3\sqrt{\log d}$ and for any $k \leq d$, $\sum_{i \leq k} (g_i^*)^2 \leq c_3 k \log(ed/k)$.

Remark 4.4. *Recall that $G/\|G\|_2$ is distributed as the uniform measure of S^{d-1}. Therefore, we have that 'most' of the mass is concentrated on regular vectors:*

$$|\{i : |x_i| \in [c_5/\sqrt{d}, c_6/\sqrt{d}]\}| \geq c_7 d, \quad \text{and} \quad |x_i^*| \leq c_8 \sqrt{\frac{\log(ed/i)}{d}}.$$

Before we proceed with the proof, let us mention the following standard fact about binomial coefficients:

Lemma 4.5. *If $1 \leq m \leq N$ then*

$$\binom{N}{m} \leq \sum_{k=1}^{m} \binom{N}{k} \leq \left(\frac{eN}{m}\right)^m.$$

Proof. Observe that

$$\sum_{k=1}^{m}\binom{N}{k} = \Big(\frac{N}{m}\Big)^m \cdot \sum_{k=1}^{m}\binom{N}{k}\Big(\frac{m}{N}\Big)^m \cdot 1^{N-k} \le \Big(\frac{N}{m}\Big)^m \cdot \sum_{k=1}^{N}\binom{N}{k}\Big(\frac{m}{N}\Big)^m \cdot 1^{N-k}$$
$$= \Big(\frac{N}{m}\Big)^m \Big(1+\frac{m}{N}\Big)^N \le \Big(\frac{eN}{m}\Big)^m,$$

as claimed. ∎

Proof. Recall that for a random variable X on a probability space we have that $\|X\|_{\psi_2} \le C\mathbb{E}\exp(|X|^2)$. Consider the probability space $\{1,...,d\}$ endowed with the uniform measure and put $X = G/R$ for a fixed realization of G. Therefore,

$$\|X\|_{\psi_2^d} \le \mathbb{E}\exp(|G|^2/R^2) = \frac{1}{d}\sum_{i=1}^{d}\exp(g_i^2/R^2).$$

Hence, taking the expectation with respect to G, if we set $R = \|g\|_{\psi_2}$, it follows that

$$\mathbb{E}\|X\|_{\psi_2^d} \le \frac{1}{d}\sum_{i=1}^{d}\mathbb{E}\exp(|g_i|^2/R^2) \le 2,$$

implying that $\mathbb{E}\|G\|_{\psi_2^d} \le 2R = 2\|g\|_{\psi_2} \le 10$.

Next, let us estimate the monotone rearrangement of the coordinate of G. By Lemma 4.5 and the tail estimate of the gaussian,

$$Pr(g_i^* \ge t) \le \binom{d}{i} Pr^i(|g| > t) \le \Big(\frac{ed}{i}\Big)^i Pr(|g| > t) \le 2\exp(i(\log(ed/i) - t^2/2)).$$

If we set $t = u\sqrt{\log(ed/i)}$ for $u > 4$, then

$$Pr(g_i^* \ge u\sqrt{\log(ed/i)}) \le 2\exp(-(u^2/4)\cdot i\log(ed/i)).$$

The claim follows by taking the union bound over $1 \le i \le d$. ∎

4.2 The Bernoulli vector in $\mathbb{R}^d$

Let $\varepsilon_1,...,\varepsilon_d$ be d independent, symmetric $\{-1,1\}$-valued random variables. Let $\mathcal{E} = (\varepsilon_1,...,\varepsilon_d)$, where in what follows we do not specify the dimension of the underlying space.

While for the gaussian vector G the structure of linear functionals $\langle G, t\rangle$ is simple but the study of the coordinate structure of $\mathcal{G}$ required some work, the situation for $\mathcal{E}$ is quite the opposite: the coordinate structure of $\mathcal{E}$ is very simple since every realization is just a point in the combinatorial cube, but the behaviour of linear forms is more complex.

For $t \in \mathbb{R}^d$, let

$$\mathcal{E}_t = \langle \mathcal{E}, t\rangle = \sum_{i=1}^{d}\varepsilon_i t_i.$$

An immediate outcome of Höffding's inequality (Lemma 2.9) is that $\mathcal{E}_t$ is a subgaussian random vector:

$$\|\mathcal{E}_t\|_{\psi_2} \le c\|\mathcal{E}_t\|_{L_2} = c\|t\|_2,$$

and the constant c is independent of t or of the dimension d. Therefore, the ψ_2 and L_2 norms are equivalent in $\mathbb{R}^d$, with the standard identification of each $t \in \mathbb{R}^d$ with $\mathcal{E}_t$. Moreover, by the characterization of the ψ_2 norm, one has that for every $p \geq 2$ and every $t \in \mathbb{R}^d$,

$$\|\langle \mathcal{E}, t\rangle\|_{L_p} \leq c_1\sqrt{p}\|\langle \mathcal{E}, t\rangle\|_{L_2} = c_1\sqrt{p}\|t\|_2. \tag{4.2}$$

While (4.2) is straightforward, it is far from sharp. It leads to the same estimate as for the standard gaussian vector in $\mathbb{R}^d$, although clearly there are vectors $t \in \mathbb{R}^d$ for which $\langle \mathcal{E}, t\rangle$ is 'much nicer' than $\langle G, t\rangle$. For example, if $t = e_1$, then $\mathcal{E}_t = \varepsilon_1$ which is a bounded random variable, and $\|\langle \mathcal{E}, t\rangle\|_{L_p} \leq 1$ for every p, while $\langle G, t\rangle = g_1$, implying that $\|\langle G, t\rangle\|_{L_p} \sim \sqrt{p}$.

The fact is that contrary to the rotationally invariant gaussian $\langle G, t\rangle$, $\langle \mathcal{E}, t\rangle$ is 'direction dependent', and at the same time its moment growth is always better than that of the corresponding gaussian. And the corresponding $L_p(\mathcal{E})$ balls endowed on $\mathbb{R}^d$ will be larger sets than $L_p(\mathcal{G})$ as the next lemma shows:

Lemma 4.6. *For every $t \in \mathbb{R}^d$ and every $p \geq 1$, $\|\langle \mathcal{E}, t\rangle\|_{L_p} \leq \|\langle G, t\rangle\|_{L_p}$*

Proof. Clearly, $\sum_{i=1}^d \varepsilon_i t_i = \frac{1}{\mathbb{E}|g_1|}\sum_{i=1}^d \varepsilon_i \mathbb{E}|g_i| t_i$. Therefore, by Jensen inequality,

$$\begin{aligned}\mathbb{E}\left|\sum_{i=1}^d \varepsilon_i t_i\right|^p &= \left(\frac{1}{\mathbb{E}|g_1|}\right)^p \mathbb{E}\left|\sum_{i=1}^d \varepsilon_i \mathbb{E}|g_i| t_i\right|^p \leq \left(\frac{1}{\mathbb{E}|g_1|}\right)^p \mathbb{E}_g\mathbb{E}_\varepsilon\left|\sum_{i=1}^d \varepsilon_i |g_i| t_i\right|^p \\ &= \left(\frac{1}{\mathbb{E}|g_1|}\right)^p \mathbb{E}\left|\sum_{i=1}^d g_i t_i\right|^p,\end{aligned}$$

because $(g_i)_{i=1}^d$ and $(\varepsilon_i g_i)_{i=1}^d$ have the same distribution. ∎

A geometric way of formulating Lemma 4.6 is that the L_p unit balls endowed on $\mathbb{R}^d$ by G and $\mathcal{E}$ imply that

$$\frac{1}{\sqrt{p}} B_2^d = B_p(G) \subset \frac{1}{\mathbb{E}|g|} B_p(\mathcal{E}).$$

The different behaviour of linear forms $\langle \mathcal{E}, t\rangle$ and $\langle G, t\rangle$ plays a crucial role in what follows. And at this point, let us quantify the moment growth of each linear form (i.e., one dimensional marginal) $\langle \mathcal{E}, t\rangle$ and as an outcome, identify $B_p(\mathcal{E})$.

In what follows let $(t_i^*)_{i=1}^d$ be the nonincreasing rearrangement of $(|t_i|)_{i=1}^d$ and set

$$\|t\|_{\mathcal{E},p} = \sum_{i=1}^p t_i^* + c\sqrt{p}\Big(\sum_{i>p}(t_i^*)^2\Big)^{1/2}.$$

The $\|\ \|_{\mathcal{E},p}$ is a 'mixture' of two norms—the ℓ_1 norm and an appropriate multiple of the ℓ_2 one; such mixtures are studied in Interpolation Theory (see e.g. [**?**]). Let us mention that at this point using $\|\ \|_{\mathcal{E},p}$ is a terrible abuse of notation, since at this point, it is not clear that $\|\ \|_{\mathcal{E},p}$ has anything to do with a norm.

The motivation behind the definition is the main result of this section:

Theorem 4.7. *There exist absolute constants c_1 and c_2 for which for any $t \in \mathbb{R}^d$,*

$$c_1\|t\|_{\mathcal{E},p} \leq \|\langle t, \mathcal{E}\rangle\|_{L_p} \leq c_2\|t\|_{\mathcal{E},p}$$

Moreover, set K to be the convex hull of B_1^d and $(1/\sqrt{p})B_2^d$. Then, there are absolute constants c_3 and c_4 such that

$$c_3 K \subset B_p(\mathcal{E}) \subset c_4 K.$$

Note that indeed K is a larger set than $B(L_p(G))$, which is $(c/\sqrt{p})B_2^d$. In contrast $B_p(\mathcal{E})$ is attained by taking the convex hull of what is effectively $B(L_p(G))$ with B_1^d.

Remark 4.8. *It is instructive to see for what directions $t \in S^{d-1}$ $\|\langle t, G\rangle\|_{L_p}$ is equivalent to $\|\langle t, \mathcal{E}\rangle\|_{L_p}$: since it always holds that*

$$\sum_{i\le p} t_i^* \le \sqrt{p}\big(\sum_{i\le p}(t_i^*)^2\big)^{1/2}$$

then equivalence is satisfies when either

- $\sum_{i\le p} t_i^* \ge c\sqrt{p}\big(\sum_{i\le p}(t_i^*)^2\big)^{1/2}$, *or when*
- $\big(\sum_{i>p}(t_i^*)^2\big)^{1/2} \ge c\big(\sum_{i\le p}(t_i^*)^2\big)^{1/2}$.

We will return to this fact when we explore empirical processes where we explore vectors t of the form $t = (f(X_i))_{i=1}^N$.

Let us begin the proof of the second part of Theorem 4.7 assuming that its first part is. The proof requires the application of some basic notions from the theory of normed spaces which will be used again in what follows.

If $K \subset \mathbb{R}^d$ is a convex body (that is, a bounded, convex, centrally-symmetric subset of $\mathbb{R}^d$ with a nonempty interior). It is standard to verify that K is the unit ball of a norm on $\mathbb{R}^d$; and, if we set

$$K^\circ = \{t : \langle x, t\rangle \le 1 \text{ for all } x \in K\}$$

then K° is the unit ball of its dual norm. For example, if $1 \le p \le \infty$ and $K = B_p^d$—the unit ball of $(\mathbb{R}^d, \|\ \|_p)$—then $K^\circ = B_q^d$, where q is the conjugate index of p.

K° is called the *polar body* of K.

The following facts are standard and their proofs are left to the reader as an exercise.

Lemma 4.9. *Let K, K_1 and K_2 be convex bodies in $\mathbb{R}^d$.*

(1) *For $c > 0$, $(cK)^\circ = c^{-1}K^\circ$.*

(2) *If $c, C > 0$ and $cK_1 \subset K_2 \subset CK_1$ then $C^{-1}K_1^\circ \subset K_2^\circ \subset c^{-1}K_1^\circ$.*

(3) $(\mathrm{conv}(K_1 \cup K_2))^\circ = K_1^\circ \cap K_2^\circ$.

Exercise 10. *Prove Lemma 4.9.*

The claim is that the unit ball of the norm $B_p(\mathcal{E})$ endowed on $\mathbb{R}^d$ is equivalent to $\mathrm{conv}\left(B_1^p \cup \frac{1}{\sqrt{p}}B_2^d\right)$. By Lemma 4.9, it suffices to show that the dual ball is equivalent to

$$\left(\mathrm{conv}\left(B_1^p \cup \frac{1}{\sqrt{p}}B_2^d\right)\right)^\circ = B_\infty^d \cap \sqrt{p}B_2^p;$$

that is, there are absolute constants c and C such that for every $t \in \mathbb{R}^d$

$$c\|\langle t, \mathcal{E}\rangle\|_{L_p} \le \sup_{x \in B_\infty^d \cap \sqrt{p}B_2^p} \langle t, x\rangle \le C\|\langle t, \mathcal{E}\rangle\|_{L_p}.$$

Now, if the first part is to be believed, all that is left is demonstrate the following:

Lemma 4.10. *There exist absolute constants c_1 and c_2 such that for every $t \in \mathbb{R}^d$,*

$$c_1\|t\|_{\mathcal{E},p} \le \sup_{x \in B_\infty^d \cap \sqrt{p}B_2^p} \langle t, x\rangle \le c_2\|t\|_{\mathcal{E},p}. \tag{4.3}$$

Proof. Fix $t \in \mathbb{R}^d$ and without loss of generality assume that $(t_i)_{i=1}^d$ is non-increasing and nonnegative. Fix $x \in B_\infty^d \cap \sqrt{p}B_2^p$ and note that by the $\ell_1 - \ell_\infty$ Hölder inequality and the Cauchy-Schwarz inequality,

$$\begin{aligned}\langle x, t\rangle &= \sum_{i\le p} t_i x_i + \sum_{i>p} t_i x_i \le \|x\|_\infty \sum_{i=1}^p t_i + \|x\|_2 \big(\sum_{i>p} t_i^2\big)^{1/2} \\ &\le \sum_{i=1}^p t_i + \sqrt{p}\big(\sum_{i>p} t_i^2\big)^{1/2} \le \|t\|_{\mathcal{E},p}.\end{aligned}$$

Taking the supremum with respect to $x \in B_\infty^d \cap \sqrt{p}B_2^p$ it follows that the left-hand side of (4.3) holds with $c_2 = 1$.

As for the right-hand side, let $R^2 = \sum_{i>p} t_i^2$, and consider two cases. Let if $\sum_{i=1}^p t_i \ge \sqrt{p}R$, then set $x = \sum_{i=1}^p e_i$. Hence, $x \in B_p^d \cap \sqrt{p}B_2^d$, and

$$\langle x, t\rangle = \sum_{i=1}^p t_i \ge \frac{1}{2}\left(\sum_{i=1}^p t_i + \sqrt{p}R\right) = \frac{1}{2}\|t\|_{\mathcal{E},p}.$$

Otherwise, $\sum_{i=1}^p t_i < \sqrt{p}R$, and set

$$x = \frac{\sqrt{p}}{R}\sum_{i>p} t_i e_i.$$

Then

$$\|x\|_\infty \le \frac{\sqrt{p}}{R}t_p \le \frac{\sqrt{p}}{R}\cdot\frac{1}{p}\sum_{i=1}^p t_i \le 1$$

and

$$\langle x, t\rangle \ge \sqrt{p}R \ge \frac{1}{2}\|t\|_{\mathcal{E},p},$$

showing that the right-hand side holds with constant $c_1 = 1/2$. ∎

Next, we shall establish the first part of Theorem 4.7, by considering the upper and lower estimates separately.

Lemma 4.11. *There exists an absolute constant c for which the following holds. For every $t \in \mathbb{R}^d$ and every $p \ge 2$,*

$$\|\langle \mathcal{E}, t\rangle\|_{L_p} \le \sum_{i=1}^p t_i^* + c\sqrt{p}\left(\sum_{i>p}(t_i^*)^2\right)^{1/2},$$

Proof. By the L_p triangle inequality, followed by the triangle inequality for the first term and (4.2) for the second — applied to $(t_i^*)_{i>p}$, it is evident that

$$\left(\mathbb{E}\left|\sum_{i=1}^{d}\varepsilon_i t_i\right|^p\right)^{1/p} \le \left(\mathbb{E}\left|\sum_{i=1}^{p}\varepsilon_i t_i^*\right|^p\right)^{1/p} + \left(\mathbb{E}\left|\sum_{i>p}\varepsilon_i t_i^*\right|^p\right)^{1/p}$$
$$\le \sum_{i=1}^{p} t_i^* + c\sqrt{p}\left(\sum_{i>p}(t_i^*)^2\right)^{1/2}. \tag{4.4}$$

■

While the proof of Lemma 4.11 is simple, it is far from obvious that this upper bound is optimal. That requires a lower bound on $\|\langle \mathcal{E}, t\rangle\|_{L_p}$, and obtaining such a lower bound is a considerably harder task.

Lemma 4.12. *There exists an absolute constant c for which the following holds. For every $t \in \mathbb{R}^d$ and every $p \ge 2$,*

$$\left(\mathbb{E}\left|\sum_{i=1}^{d}\varepsilon_i t_i\right|^p\right)^{1/p} \ge c\left(\sum_{i=1}^{p} t_i^* + \sqrt{p}\left(\sum_{i>p}(t_i^*)^2\right)^{1/2}\right). \tag{4.5}$$

The proof of Theorem 4.12 requires some preparation.

Lemma 4.13. *Let $a \in \mathbb{R}^d$. If $\|a\|_\infty \le \|a\|_2/16\sqrt{s}$, then there is a decomposition of $\{1,...,d\}$ to coordinate blocks $I_1, ...I_s$, and for every $1 \le j \le s$,*

$$\frac{\|a\|_2}{4\sqrt{s}} \le \left(\sum_{i\in I_j}|a_i|^2\right)^{1/2} \le \frac{\|a\|_2}{2\sqrt{s}}$$

Proof. Without loss of generality, we may assume that $a_1 \ge a_2 \ge ... \ge a_d \ge 0$. Let j_1 be the smallest integer for which $\sum_{i=1}^{j_1} a_i^2 \ge \frac{\|a\|_2^2}{16s}$. Since $a_1 \le \|a\|_2/16\sqrt{s}$ it follows that $j_1 > 1$. Moreover, since $\sum_{i=1}^{j_1-1} a_i^2 < \frac{\|a\|_2^2}{/}16s$, it is evident that

$$\sum_{i=1}^{j_1} a_i^2 < a_1^2 + \frac{\|a\|_2^2}{16s} \le \frac{\|a\|_2^2}{8s}$$

and setting $I_1 = \{1, ..., j_1\}$,

$$\frac{\|a\|_2}{4\sqrt{s}} \le \left(\sum_{i\in I_1} a_i^2\right)^{1/2} \le \frac{\|a\|_2}{2\sqrt{s}},$$

as required.

We continue along the same lines and construct k coordinate blocks: the process terminates when $\sum_{i=j_k}^{d} a_i^2 \le \frac{\|a\|_2^2}{16s}$, implying that

$$k\frac{\|a\|_2^2}{4s} \ge \sum_{\ell=1}^{k}\sum_{i\in I_\ell} a_i^2 \ge \frac{\|a\|_2^2}{4};$$

hence, $k \ge s$, as claimed. ■

Proof of Lemma 4.12. Without loss of generality we may assume that $t_1 \geq t_2 \geq ... \geq t_d \geq 0$. First, consider the case $\sum_{i=1}^p t_i \geq \frac{\sqrt{p}}{16}(\sum_{i>p} t_i^2)^{1/2}$. Clearly, with probability at least 2^{-p}, $\varepsilon_1 = \varepsilon_2 = ... = \varepsilon_p = 1$, and in particular,

$$\mathbb{E}\left|\sum_{i=1}^p \varepsilon_i t_i\right|^p \geq 2^{-p}\left(\sum_{i=1}^p t_i\right)^p.$$

Moreover, by Jensen's inequality and since $\mathbb{E}\varepsilon_i = 0$,

$$\mathbb{E}\left|\sum_{i=1}^d \varepsilon_i t_i\right|^p = \mathbb{E}_{(\varepsilon_i)_{i\leq p}}\mathbb{E}_{(\varepsilon_i)_{i>p}}\left|\sum_{i=1}^d \varepsilon_i t_i\right|^p \geq \mathbb{E}_{(\varepsilon_i)_{i\leq p}}\mathbb{E}\left|\sum_{i=1}^p \varepsilon_i t_i\right|^p, \tag{4.6}$$

implying that

$$\|\langle \mathcal{E}, t\rangle\|_{L_p} \geq \frac{1}{2}\sum_{i=1}^p t_i \geq c\left(\sum_{i=1}^p t_i + \sqrt{p}\left(\sum_{i>p} t_i^2\right)^{1/2}\right).$$

Next, assume that $\sum_{i=1}^p t_i < \frac{\sqrt{p}}{16}(\sum_{i>p} t_i^2)^{1/2}$ and consider the vector $a = (t_i)_{i\geq p}$. Observe that

$$\|a\|_\infty \leq \frac{1}{p}\sum_{i=1}^p t_i \leq \frac{\|a\|_2}{16\sqrt{p}},$$

and the condition of Lemma 4.13 holds with $s = p$. Let $I_1, ..., I_p$ be the decomposition of $\{p+1, ..., d\}$ guaranteed by that Lemma. Recall that there is an absolute constant c_1 for which

$$\|\sum_{i\in I_j} \varepsilon_i t_i\|_{L_p} \leq c_1\sqrt{p}\|\sum_{i\in I_j} \varepsilon_i t_i\|_{L_2} = c_1\sqrt{p}(\sum_{i\in I_j} t_i^2)^{1/2};$$

thus, by the Paley-Zygmund inequality, there are absolute constants c_2 and c_3 such that

$$Pr\left(\left|\sum_{i\in I_j} \varepsilon_i t_i\right| \geq c_2\left(\sum_{i\in I_j} t_i^2\right)^{1/2}\right) \geq c_3.$$

Note that the p random variables $\sum_{i\in I_j} \varepsilon_j t_i$ are independent and symmetric. Therefore, with probability at least $(c_3/2)^p$, for every $1 \leq j \leq p$,

$$\sum_{i\in I_j} \varepsilon_i t_i \geq c_2\left(\sum_{i\in I_j} t_i^2\right)^{1/2} \geq \frac{c_4}{\sqrt{p}}\left(\sum_{i>p} t_i^2\right)^{1/2}$$

for a suitable absolute constant c_4. Thus,

$$\left(\mathbb{E}\left|\sum_{i>p} \varepsilon_i t_i\right|^p\right)^{1/p} \geq c_5\sqrt{p}\left(\sum_{i>p} t_i^2\right)^{1/2},$$

and the claim follows using a similar argument to (4.6). ∎

Part II

An Introduction to Statistical Learning Theory

Chapter 5

Introduction

The underlying theme of these notes is *structure*, and specifically, the way randomness may be used to expose hidden structures in sets. Using randomness to expose structure has been one of the central ideas of *Asymptotic Geometric Analysis*, an area devoted to the study of convex sets in $\mathbb{R}^d$. 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, are areas that focus on *prediction* and *estimation* problems. Roughly put, a random sample is used to generate an approximation of an unknown random variable by (wisely) selecting a function 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.

To give some indication of why problems involving sampling are connected with structure, let us begin by describing a toy example: selecting randomly a subset of the coordinates of a single vector in $\mathbb{R}^d$. This example captures many of the issues one has to contend with in what follows, though, obviously, difficulties will have to be addressed not just for a single vector in $\mathbb{R}^d$ (or for a single function), but rather uniformly—for an infinite family of vectors/functions.

Contrary to what one might 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 the outcome of a sampling procedure may be totally distorted. To illustrate this observation let $v \in \mathbb{R}^d$ (and d is very large). One has access to set of N coordinates $I \subset \{1, ..., d\}$ selected randomly. and the hope is that the sampled vector $(v_i)_{i\in I}$ 'inherits' the significant properties of the vector $(v_i)_{i=1}^d$—for example, that the ℓ_2 norm of v,

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

is reflected by the values $(v_i)_{i\in I}$.

Clearly, the best outcome one can hope for is when v is the constant vector, in which case, for every $I \subset \{1, ..., d\}$, if $P_I v = (v_i)_{i\in I}$ then

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

Therefore, if $|I|$ is very close to N (as would be the case in any reasonable of choosing a random

subset I), a successful outcome of sampling is that $\|P_I v\|_2$ is 'very close' to $(|N|/d)^{1/2}\|v\|_2$; in other words, sampling shrinks the Euclidean norm by a factor that is very close to $(|I|/d)^{1/2}$.

Now, consider the following vectors

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

Both vectors belong to the Euclidean unit sphere, but respond in very different ways to a choice of a random subset of their coordinates. Indeed, for any reasonable definition of a random subset $I \subset \{1, ..., d\}$, and even if N is large, the typical outcome is that the first coordinate is not be selected. Therefore, $\|P_I v_1\|_2 = 0$, which is very far from the benchmark value of $\sqrt{N/d}$. In contrast, since v_2 is a constant vector, $\|P_I v_2\|_2 = \sqrt{|I|/d}$.

The way the two vectors respond to sampling happens to be an outcome of their different structures. Although the two have the same Euclidean norm, the 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 a vector says absolutely nothing about the success of sampling. And, like-wise, having information on the vector being bounded with respect to other ℓ_p norms is equally useless. Meaningful information has to provide more information than just "being bounded".

Intuitively and somewhat inaccurately, independent sampling works reasonably well when the wanted property is captured by a 'large set' of Ω, 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 $\sqrt{|I|/d}\|v\|_2$ is that the set of coordinates

$$J_{\alpha,\beta} = \left\{ i : \|v\|_2 \frac{\alpha}{\sqrt{d}} \le |v_i| \le \|v\|_2 \frac{\beta}{\sqrt{d}} \right\} \tag{5.1}$$

has cardinality that is proportional to d, for α and β that are absolute constants. In that case, a typical random choice of $I \subset \{1, ..., d\}$ of cardinality N satisfies that $|I \cap J_{\alpha,\beta}| \sim N$, leading to the wanted outcome: the shrinking of the ℓ_2 norm by a factor of $(N/d)^{1/2}$. 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

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

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

$$\|v\|_p = \left(\sum_{i=1}^{d} |v_i|^p \right)^{1/p} = \frac{1}{\sqrt{2}} \left(1 + \frac{d-1}{d^{p/2}} \right)^{1/p}, \tag{5.2}$$

which is of the order of 1. Set $\alpha = 1/2$ and $\beta = 2$, and thus $|J_{\alpha,\beta}| = d - 1$, implying, at least intuitively, that a random choice of coordinates does not 'collapse' the ℓ_2 norm. But as

[1] Although the terms "peaky" and "well-spread" are used rather freely, one should take care when using them. For example, is the vector $(1/\sqrt{2}, 1/\sqrt{2M}, ...1/\sqrt{2M})$ peaky or well-spread?

it happens, the same is not true for the ℓ_p norm for any $p > 2$. Indeed, for a typical subset $I \subset \{1, ..., d\}$, say of cardinality $N \ll d$,

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

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

The reason behind the collapse of the ℓ_p norm after sampling is simple—as relative to the p-norm, v is peaky: the main contribution to the ℓ_p norm comes from a single coordinate and all the other coordinates are negligible. In such a situation, sampling is useless. Upon some reflection it is straightforward to verify that the correct p-analogue of the set $J_{\alpha,\beta}$ is

$$\left\{i : \|v\|_p \frac{\alpha}{d^{1/p}} \le |v_i| \le \|v\|_p \frac{\beta}{d^{1/p}}\right\}, \tag{5.3}$$

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

Having large level sets like (5.1) or (5.3) has strong ties with the so-called *small-ball condition* and such a property helps 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 $\sqrt{d/|I|}(\sum_{i\in I} |v_i|^2)^{1/2}$ 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

$$\left\{i : |v_i| \ge t\frac{\|v\|_2}{\sqrt{d}}\right\} \tag{5.4}$$

for $t \ge 1$.

> One of the key observations used throughout these notes is 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.

5.1 A question

Let F be a class of functions defined on a probability space (Ω, μ). In a subtle twist that will become clearer in what follows, assume that very little is known about the measure μ. As a result, if X is distributed according to μ, there is no information on the $L_2(\mu)$ distance between any $f_1, f_2 \in F$; i.e.,

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

is not known. Instead of information on μ, one receives as data a sample $X_1, ..., X_N$, selected independently according to μ, and uses the empirical means

$$\left(\frac{1}{N}\sum_{i=1}^N |f_1 - f_2|^2(X_i)\right)^{1/2} \tag{5.5}$$

as a 'guess' of $\mathbb{E}|f_1(X) - f_2(X)|^2$.

Question 5.1. *Are the empirical means* (5.5) *a good guess of* $L_2(\mu)$ *distances? In a more geometric language, is the original* L_2 *structure of* F *preserved via the empirical means?*

Remark 5.2. *Clearly, a version of Question 5.1 is valid for any* $L_p(\mu)$ *rather than just for* $p = 2$.

As an example, let $\Omega = \mathbb{R}^d$ and set μ to be some probability measure on $\mathbb{R}^d$. For $T \subset \mathbb{R}^d$ let $F_T = \{\langle t, \cdot \rangle : t \in T\}$ be the class of linear functionals defined by T. As noted previously, each $t \in T$ has a two roles: as a vector in $\mathbb{R}^d$, and as a linear functional on $\mathbb{R}^d$. Clearly, for $u, v \in T$ one may easily compute various distances between u and v, for example, the ℓ_p^d distances,

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

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

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

Remark 5.3. *Let us stress again that there is no reason why* $\|u - v\|_p$ *should have anything to do with* $\|u - v\|_{L_p}$*; these are completely different objects.*

An assumption that is encountered frequently is that the measure μ is *isotropic*, i.e., it is symmetric and if for every $u \in \mathbb{R}^d$, $\|u\|_{L_2(\mu)}^2 = \mathbb{E} \langle u, X \rangle^2 = \|u\|_2^2$. Thus, when μ is isotropic, the L_2 norm endowed on linear functionals in $\mathbb{R}^d$ coincides with the $\| \ \|_2$ norm.

Given $X_1, ..., X_N$, let

$$\Phi\left((X_i)_{i=1}^N, f, h\right) = \frac{1}{N} \sum_{i=1}^N (f - h)^2(X_i), \tag{5.6}$$

and the hope is that with high probability, at least for most of the pairs in F,

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

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

Keeping in mind that there could be (and will be) a fundamental difference between upper estimates and lower ones, it makes sense to split Question 5.1 to two parts.

Question 5.4. *Given the functional* Φ*, find the best possible choices of* $0 < \delta_N < 1$, $A_N, B_N, r_N, r'_N > 0$ *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*

$$\Phi\left((X_i)_{i=1}^N, f, h\right) \geq A_N \|f - h\|_{L_2}^2, \tag{5.7}$$

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

$$\Phi\left((X_i)_{i=1}^N, f, h\right) \leq B_N \|f - h\|_{L_2}^2. \tag{5.8}$$

The optimal values of the parameters should be determined by the structure of F, the measure μ and the sample size N.

Remark 5.5. *One can ask a more fundamental question: what is the best choice of a functional Φ? Although* (5.6) *is the obvious candidate, is it the right one? As it happens, the answer to that question is a resounding "no", and far better alternatives are explored in what follows.*

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, when, with probability $1 - \delta_N$, if $\|f\|_{L_2} \geq r_N$ then

$$\Phi\left((X_i)_{i=1}^N, f, 0\right) \geq A_N \|f\|_{L_2}^2; \tag{5.9}$$

and if $\|f\|_{L_2} \geq r'_N$ then

$$\Phi\left((X_i)_{i=1}^N, f, 0\right) \leq B_N \|f\|_{L_2(\mu)}^2. \tag{5.10}$$

To put Question 5.4 in some context, let us present three examples of problems that may be resolved once the question is answered. The first example originates in asymptotic geometric analysis; the second is from random matrix theory; and the third one is from statistics/signal processing.

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

Let $T \subset (\mathbb{R}^d, \| \|_2)$ be a finite set. One 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 d, in a way that (almost) preserves the Euclidean distances between the points in T. Thus, the goal is to find a mapping $\psi : T \to \mathbb{R}^k$ which satisfies that for every $u, v \in T$, and $\varepsilon > 0$ as small as possible,

$$1 - \varepsilon \leq \frac{\|\psi(u) - \psi(v)\|_2}{\|u - v\|_2} \leq 1 + \varepsilon; \tag{5.11}$$

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

This problem has been studied extensively since the mid-80's, when Johnson and Lindenstrauss proved their celebrated lemma. They showed that with high probability, a correctly normalized random orthogonal projection[2] onto a k-dimensional subspace of $\mathbb{R}^d$ for $k = c\varepsilon^{-2} \log |T|$ satisfies (5.11).

Remark 5.6. *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 (5.11) over the last 30 years. One class of such operators that is of particular interest in the context of these embeddings consists of random matrices with independent rows.

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

Let X be an isotropic random vector in $\mathbb{R}^d$, set $X_1, ..., X_N$ to be independent copies of X, and define

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

i.e., the matrix whose rows are $X_1, ..., X_N$.

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

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

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 over a large collection of points is possible.

To see the connection the embedding problem has with Question 5.4, let $F_T = \{\langle t, \cdot \rangle : t \in T\}$. Note that by selecting Φ as in (5.6), Question 5.4 implies that for every $u, v \in T$,

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

and $\|f_u - f_v\|_{L_2}^2 = \mathbb{E}\langle u - v, X \rangle^2 = \|u - v\|_2^2$. Thus, (5.11) follows from a positive answer to Question 5.4 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 $r_N = r'_N = 0$.

An estimate for a finite set T was presented in Lemma 3.11, but the real reason why $N = c\varepsilon^{-2} \log |T|$ is a suitable choice might appear mysterious at this point; in fact, the technical argument used in the proof of Lemma 3.11 is far from the complete picture. Once the necessary machinery is developed it will become clear that the logarithm of the cardinality of a set is a actually a rather crude measure of the set's complexity, and considerably sharper alternatives can be established.

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^{d-1}:

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

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^d.

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

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

$$\lambda_{\min}^2 = \inf_{x \in S^{d-1}} \frac{1}{N} \sum_{i=1}^{N} \langle X_i, x \rangle^2,$$

corresponding to (5.10) and (5.9), respectively.

The fact that the upper estimate and the lower one have been separated will prove to be significant. For example, one may show that under minimal assumptions on X, and with high probability,

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

On the other hand,

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

is true, but only under more restrictive conditions. Moreover, these estimate hide what is the natural complexity parameter associated with the Euclidean unit sphere: $\sqrt{d}$. The 'error term' $\sqrt{d/N}$ happens to be the ratio between the 'complexity' of the indexing set—in this case, of the unit sphere S^{d-1}—and the square root of the cardinality of the given sample. This will prove to be a general phenomenon. Why $\sqrt{d}$ captures the complexity of sphere S^{d-1} has to be explained. And also, since the sphere is a very special set, one has to identify the right complexity parameters of more general classes of functions and their roles in 'error terms'.

Simple exact recovery

Let $T \subset \mathbb{R}^d$ and assume that some $t_0 \in T$ is selected but is kept concealed. The goal is to identify t_0, or, if that is impossible, to approximate it with respect to the ℓ_2 norm. To perform that task, the information one is provided consists 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 5.4 may be answered in this case. Specifically, that with probability at least $1 - \delta$, if $u, v \in T$ and $\|u - v\|_{L_2}^2 = \mathbb{E}\langle u - v, X \rangle^2 \geq r_N$ 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 $A_N > 0$. Clearly, on that event, if $\langle t, X_i \rangle = \langle t_0, X_i \rangle$ for $1 \leq i \leq$, then $\|t - t_0\|_{L_2}^2 \leq r_N$; moreover, if μ happens to be an isotropic measure, then

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

Finally, if $r_N = 0$ then no other point in T agrees with t_0 on the measurements. As a result, 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.

These observations have different names in different fields. In a naive version of *sparse recovery*, the set T consists of all the vectors in $\mathbb{R}^d$ that are s-sparse; that is, supported on at most s coordinates relative to the standard basis in $\mathbb{R}^d$. In learning theory, this is an example of a *realizable learning problem*, or a noise-free problem when the class of functions consists

of linear functionals in $\mathbb{R}^d$. 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 $\ker(\Gamma) \cap T$.

All these examples are explored in some detail in what follows. For now, their role is to convince the reader that Question 5.4 is a nontrivial question and has far-reaching implications in modern areas of mathematics, statistics, computer science and engineering—even if one only considers the restricted setup of classes of linear functionals in $\mathbb{R}^d$ and iid sampling is performed according to an isotropic measure. Because it is such a fundamental question, it should not be surprising that developing the machinery necessary for addressing it requires some effort. At the same time, as a learning problem, Question 5.1 is almost "as simple as it gets". One should keep in mind that general learning problems are far more complex and addressing them requires much more than an answer to Question 5.1.

5.2 A Learning problem

Let us turn to the "main event" of these notes: 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 and X is a random variable taking values in Ω which is distributed according to μ.

Let $\mathcal{Y}$ be a collection of *admissible targets*, consisting of the random variables from which the (unknown) target is selected. Obviously, one would like to keep that set as large as possible.

Thus, a learning problem is naturally associated with a triplet (F, X, Y), where the class F is known, but X and Y are not—other than, perhaps, some minimal assumptions on their general properties (e.g. that $Y \in \mathcal{Y}$).

Consider some (fixed but unknown) $Y \in \mathcal{Y}$. The goal in a learning problem is to find some $f \in F$ that is as close to Y as possible. For obvious reasons, the notion of similarity is up to the learner, and is calibrated using a *loss function*. In what follows we only consider the following collection of loss functions:

Definition 5.7. *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 is the squared loss $\ell(t) = t^2$, and it is the focus of these notes.

Given $y \in \mathbb{R}$ and $x \in \Omega$, the loss incurred by predicting $f(x)$ instead of y is $\ell(f(x) - y)$, and for the squared loss, it is $(f(x) - y)^2$. 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 (X, Y). We assume in what follows that the minimizer exists and is unique, which happens to be the case under rather minimal assumptions. That unique minimizer is denoted 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

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

This lack of information is the key difference between standard problems in approximation theory and the ones in statistical learning theory. In the latter, all the information one has access to, other than the identity of the class F, is a random sample $\mathcal{D}=(X_i,Y_i)_{i=1}^N$, selected according to the joint distribution (X,Y). Using that random sample the learner is expected to produce some $\hat{f}$ and ensure that for most samples, $\hat{f}$ approximates f^* in an appropriate sense.

Naturally, there are various notions of approximation one may consider. In all of them the objective is to make the error as small as possible and to do that with the highest confidence (i.e., probability estimate) possible. Here are a few important notions:

(1) $\hat{f}$ is selected from F, and one would like $\hat{f}$ to be close to f^* in the $L_2(\mu)$ sense; that is, ensure that with high probability with respect to the given sample,

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

$\mathcal{E}_e$ is called the *estimation error*.

(2) $\hat{f}$ is selected from F and one would like the risk of $\hat{f}$ to be almost the best possible in F. In other words, with high probability with respect to the given sample,

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

$\mathcal{E}_p$ is called the *prediction error*, and it is important to note that the constant in front of $\inf_{f\in F}R(f)$ is 1. The prediction problem becomes much easier if one is allowed to change it to a constant that is larger than 1.

(3) Procedures taking values in F are called *proper*. One can be allowed more freedom if 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,

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

$\mathcal{E}_{agg}$ is called the *aggregation error*[3] and such parameters are called *unrestricted*.

(4) It is possible to show that if the class F is 'too rich', the estimation error and prediction error are too big to be of any use, regardless of the way $\hat{f}$ is selected. Instead of restricting the problems one can reasonably address to small classes, an alternative is to invoke *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 by the learner. If two functions fit the random data to a similar extent, 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.

[3]The name "aggregation error" is not standard; it comes from problems involving finite classes of functions (or "dictionaries") consisting of reasonable estimators. One can show that by combining the estimators—and as a result leaving the original class—one can produce an even better estimator.

There are many questions that one could ask at this point, but the most fundamental one is probably this:

Question 5.8. *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 a learning procedure $\hat{f}$?*

The main goal of these notes is to address Question 5.8, or, rather, give the reader a flavour of what is needed for addressing it.

The accuracy/confidence tradeoff is the term used here to describe the way $\mathcal{E}_e$ and $\mathcal{E}_p$ depend on the confidence parameter δ, the sample size N and some features of F, X and Y. It is the notion more frequently used in statistical literature (though it is usually called the *error rate* of the problem). The emphasis is on the best tradeoff one can hope for when given a fixed sample size N, which is a natural point of view when data is expensive. At the same time, computer science literature uses the notion of *sample complexity*, in which one is given the wanted accuracy and confidence levels and has to produce a sample size N that suffices to ensure recovery with those accuracy and confidence levels.

To a certain extent the two notions are equivalent. The one subtle point that should be kept in mind is the information that the learning procedure requires as input: whether it is the sample size, or, rather, the wanted accuracy and confidence levels. We use both notions in various parts of this exposition.

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

Example 5.9. *Let μ be the standard gaussian measure on $\mathbb{R}^d$: i.e., the measure whose density is proportional to $\exp(-\|t\|_2^2/2)$ (in what follows we do not use any of the special properties of the gaussian measure—other than the fact that it is isotropic).*

Let $T = B_1^d = \{x : \sum_{i=1}^d |x_i| \leq 1\}$ be the unit ball of the normed space $\ell_1^d = (\mathbb{R}^d, \| \ \|_1)$ and set

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

As always, one may identify each $t \in T$ with the linear functional $f_t = \langle t, \cdot \rangle$.

Let $t_0 \in \mathbb{R}^d$ (not necessarily in B_1^d) and set $Y = \langle t_0, X \rangle + W$, for W that 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 $t \in T$,

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

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.

In an attempt to identify or approximate t_0 the data one is given consists of a random sample

$$(X_i, Y_i)_{i=1}^N = (X_i, \langle X_i, t_0 \rangle + W_i)_{i=1}^N$$

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 in the proper setup is to use that given sample to produce some $\hat{t} \in B_1^n$; thus, the learning procedure assigns to each sample $(X_i, Y_i)_{i=1}^N$ a function $\hat{f} = \langle \hat{t}, \cdot \rangle \in F$. The success of the procedure is measured, for a given confidence parameter δ, by

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

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

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

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

that the procedure achieves with confidence of $1-\delta$ with respect to the given samples $(X_i, Y_i)_{i=1}^N$.

In the unrestricted setup one is allowed to select $\hat{f} = \langle \hat{t}, \cdot \rangle$ for $\hat{t} \in \mathbb{R}^d$ that need not belong to B_1^d. The success of the procedure is measured by the tradeoff between the excess risk $R(\hat{f}) - R(f^*)$ and the confidence with which that prediction error can be achieved.

With Question 5.8 in mind, how should $\hat{t}$ be selected? How is the fact that $T = B_1^n$ reflected in the accuracy/confidence tradeoff of the estimation problem and of the prediction problem? And how would the tradeoff change for different X or Y—for example, if X happen to be more 'heavy tailed' than gaussian?

Even in the restricted setup of linear regression in B_1^d, giving a complete answer to all these questions is highly nontrivial. Doing so for an arbitrary problem seems to be asking for too much. Still, as the reader will discover, a highly satisfactory answer can be obtained in very general situations. That answer requires the development of a suitable machinery, which will be done gradually.

5.2.1 Some Definitions

As always, 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 collection of admissible targets.

Definition 5.10. *A set of admissible targets is minimal if it contains all targets of the form $\{f(X) + W_\sigma : f \in F,\ \sigma \geq 0\}$, where W_σ is a centred gaussian random variable with variance 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., targets 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 5.11. *Given a sample size N, a (proper) learning procedure is a collection of functions $\Phi_N : (\Omega \times \mathbb{R})^N \to F$. In other words, Φ_N assigns to each $(x_i, y_i)_{i=1}^N$ some $f \in F$. An* unrestricted procedure *is allowed to take values outside the class F.*

In what follows we ease notation and write Φ instead of Φ_N.

Remark 5.12. *For the time being, we shall focus our attention to proper procedures. We explain in what follows why it is essential in a generic learning problem to allow the procedure to take values outside F.*

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

Definition 5.13. *Given a class F, a set of admissible targets $\mathcal{Y}$ and an integer N, a procedure Φ performs with estimation accuracy $\mathcal{E}_e$ and confidence parameter δ if for any $Y \in \mathcal{Y}$, with probability at least $1 - \delta$,*

$$\|\Phi\left((X_i, Y_i)_{i=1}^N\right) - f^*\|_{L_2} \leq \mathcal{E}_e \tag{5.12}$$

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

$$R\left(\Phi\left((X_i, Y_i)_{i=1}^N\right)\right) \leq \inf_{f \in F} R(f) + \mathcal{E}_p. \tag{5.13}$$

The functions $\mathcal{E}_e$ and $\mathcal{E}_p$ depend on δ, N and some features of the known target Y (for example, the noise level $\|Y - f^*(X)\|_{L_2}$).

Alternatively, given $\varepsilon > 0$ and $0 < \delta < 1$, the sample complexity of the procedure Φ is the minimal sample size N_0 such that for any $N \geq N_0$, with probability $1 - \delta$,

$$\|\Phi\left((X_i, Y_i)_{i=1}^N\right) - f^*\|_{L_2} \leq \sqrt{\varepsilon} \;\text{ resp. }\; R\left(\Phi\left((X_i, Y_i)_{i=1}^N\right)\right) \leq \inf_{f \in F} R(f) + \varepsilon.$$

Follows Definition 5.13, the performance of a procedure is measured by its success when faced with any admissible target $Y \in \mathcal{Y}$ and that success is measured via the accuracy/confidence tradeoff (i.e., the error a procedure may guarantee and the probability with which it can guarantee it), or, equivalently, according to its sample complexity.

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 $\mathcal{Y}_{\min}$. 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 the notion is used here.

Definition 5.14. *Given a class of functions F on a probability space (Ω, μ), a set of admissible targets $\mathcal{Y}$ and a given sample size N, a procedure Φ performs with the γ-minimax accuracy for a confidence parameter δ if*

- *For every $Y \in \mathcal{Y}$, with probability at least $1-\delta$, (5.12) (resp. (5.13)) holds, with an estimation error $\mathcal{E}_e$ (resp. prediction error $\mathcal{E}_p$).*
- *If Ψ is any learning procedure, then there is some $Y \in \mathcal{Y}$ for which the event*

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

(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 only with probability that is smaller than $1-\delta$.

Because it is rather optimistic to hope that one can identify the best possible procedure, the parameter γ gives one some freedom. Thus, a procedure is optimal in the minimax sense if for a given degree of confidence, the accuracy with which it performs when faced with any $Y \in \mathcal{Y}$ is proportional to the theoretical limitations relative to that set of targets and for γ that is an absolute constant. Also, from here on all the sets of admissible targets $\mathcal{Y}$ considered contain at least $\mathcal{Y}_{\min}$ — i.e., the set of all realizable targets $Y = f(X)$ for $f \in F$, and targets consisting of independent, additive gaussian noise, that is $Y = f(X) + W$ where W is a centred gaussian random variable that is independent of X.

Unfortunately, we shall not discuss the question of the optimality and minimax rates in these notes. A more detailed exposition can be found in the appendix of [**?**] and in [**?**].

5.3 What estimates should one expect?

An intuitive way of viewing estimation and prediction problems is as structure preservation. In the case of the squared loss (which is the one we shall focus on here), one would like to compare $\mathbb{E}(f(X)-Y)^2$ to $\mathbb{E}(h(X)-Y)^2$ for every pair $(f,h) \in F$ and to be able to decide which of the two is bigger given only the values $(X_i, Y_i)_{i=1}^N$. Unfortunately, such a uniform comparison is likely to be impossible: even taking the most optimistic view of statistical recovery, one should expect inaccuracies caused by the incomplete information at one's disposal—as a sample does not capture the entire picture. Therefore, what is a more realistic goal is to be able to compare $\mathbb{E}(f(X)-Y)^2$ to $\mathbb{E}(h(X)-Y)^2$ when f and h are 'far enough', allowing for the true distance between f and h overcome the fluctuations caused by the data (and what is meant by 'far enough' has to be made explicit). Also, one should only be concerned about such comparisons when either f or h is the true minimizer of the functional $f \to \mathbb{E}(f(X)-Y)^2$: the goal is identify a function whose 'predictive capabilities' are comparable with the performance of the best function in the class. Naturally, one has to be able to do that without knowing beforehand which function is the best in the class, making the task more challenging.

As will become clearer, such a comparison is possible in a direct way only when the set of 'almost minimizers' in F has a small diameter; i.e., for an acceptable error of ε, the set

$$\{f \in F : \mathbb{E}(f(X)-Y)^2 \leq \mathbb{E}(f^*(X)-Y)^2 + \varepsilon\}$$

consists only of functions that are close to f^*. The difficulty in dealing with situations where the set of almost minimizers has a large diameter occurs because of massive 'statistical fluctuations' in estimates of $R(f) - R(f^*) = \mathbb{E}(f(X) - Y)^2 - \mathbb{E}(f^*(X) - Y)^2$ when f is an almost minimizer (meaning that $R(f) - R(f^*)$ is close to 0) but at the same time, $\|f - f^*\|_{L_2}$ is large. We refer to such issues as *geometric obstructions*; they are connected to the so-called *convexity condition* explored in Section 5.5.

Remark 5.15. *It should be stressed that the ability to accurately identify from data the smaller of* $\mathbb{E}(f(X)-Y)^2$ *or* $\mathbb{E}(h(X)-Y)^2$ *when* $h = f^*$ *and* f *is far from* f^* *is an approach that leads to a sufficient condition for recovery, resulting in an upper estimate on the sample complexity or on the accuracy/confidence tradeoff. Because it is enough to identify, for a typical sample, just one function that is close enough to* f^**, the fact that a method can be used to identify many such functions, is, at least in principle, a source of looseness.*

*The question of whether "separation analysis" is an optimal approach is still open. In almost any noisy learning scenario it leads to the best known bounds and in most cases, to a bound that is optimal in the minimax sense. A discussion on that issue can be found in [***?***].*

For now, let us assume that one is dealing with the simpler case, in which the set of 'almost minimizers' of the risk functional consists only of functions that are close to the true minimizer. To see how Question 5.8 can be addressed in such a scenario, let us attempt to speculate on when uniform separation is possible.

At the heart of this discussion is the following crucial observation: that the difficulty of learning problems is 'coded' in the geometry of a random subset of F that is naturally associated with F and $\sigma = (X_1, ..., X_N)$.

Definition 5.16. *Given a class* F *and a sample* $\sigma = (X_i)_{i=1}^N$*, the random coordinate projection associated with* F *and* σ *is the set*

$$P_\sigma F = \{(f(X_1), ..., f(X_N)) : f \in F\} \subset \mathbb{R}^N.$$

Example 5.17. *Let* $T \subset \mathbb{R}^d$ *and let* X *be a random vector in* $\mathbb{R}^d$*. If* $\Gamma = \sum_{i=1}^N \langle X_i, \cdot \rangle e_i : \mathbb{R}^d \to \mathbb{R}^N$ *is the random matrix whose rows are* $X_1, ..., X_N$*, and* $F_T = \{\langle t, \cdot \rangle : t \in T\}$*, then* $P_\sigma F_T = \Gamma T$*.*

The most important part of this presentation is the study of the way in which the difficulty of statistical problems involving the class F and the underlying distribution X is reflected in the geometry of the sets $P_\sigma F$. We will explain how statistical problems can be translated to questions on the structure of the sets $P_\sigma F$, and then solve these questions using the machinery of asymptotic geometric analysis—thus resolving the original statistical problem.

5.3.1 Two regimes

Let us try to split the description of a generic problem, consisting of a triplet (F, X, Y) (and under the assumption that there are no geometric obstructions) to two regimes: the *low-noise regime* and the *high-noise regime*, by making educated guesses on the reasons why a statistical

recovery procedure might make mistakes in each regime. We will make these guesses more concrete in the next section.

At the extreme end of low-noise problems is the (seemingly) simple case, in which the there is no noise at all: $Y \in F$, and in particular $f^*(X) = Y$. Continuing the separation analogy of the previous section, one has to be able to distinguish between

$$\mathbb{E}(f(X) - Y)^2 = \mathbb{E}(f(X) - f^*(X))^2 \quad \text{and} \quad \mathbb{E}(h(X) - f^*(X))^2,$$

using only the given sample as data. And, restricting to the important case, in which either $f = f^*$ or $h = f^*$, the question becomes distinguishing $\mathbb{E}(f(X) - f^*(X))^2$ from 0 uniformly in $f \in F$,—at least when $\|f - f^*\|_{L_2}$ is not too small.

As always, one has to contend with only having access to partial information in the form of a sample $(X_i, Y_i)_{i=1}^N = (X_i, f^*(X_i))_{i=1}^N$ and the natural problematic object is the corresponding *version space*:

Definition 5.18. *For $f^* \in F$ and a sample $\sigma = (X_i)_{i=1}^N$, the version space associated with F, f^* and σ is a random subset of F defined by*

$$\mathbb{V}_{f^*,\sigma} = \{f \in F : f(X_i) = f^*(X_i) \text{ for all } 1 \le i \le N\}.$$

Thus, functions in $\mathbb{V}_{f^*,\sigma}$ are indistinguishable from f^* on the given sample; therefore, if, for a given sample σ the L_2 diameter of $\mathbb{V}_{f^*,\sigma}$ is at least r, there is no hope to of being able to separate f^* and *every* $f \in F$ that satisfies $\|f - f^*\|_{L_2} \ge r/2$ using only the data $(X_i, f^*(X_i))_{i=1}^N$.

Example 5.19. *Let us return to the case in which $T \subset \mathbb{R}^d$ and $F_T = \{\langle t, \cdot\rangle : t \in T\}$ is the class of linear functionals defined by T. Let X be a centred random vector in $\mathbb{R}^d$, set $t^* \in T$ and put $Y = \langle t^*, \cdot\rangle$. Thus, the triplet (F_T, X, Y) defines a noise-free problem. Given the sample $(X_i, Y_i)_{i=1}^N = (X_i, \langle t^*, X_i\rangle)_{i=1}^N$, let $\Gamma = \sum_{i=1}^N \langle X_i, \cdot\rangle e_i$ be the random matrix whose rows are $X_1, ..., X_N$, and note that*

$$\begin{aligned}\mathbb{V}_{t^*,\sigma} &= \{\langle t, \cdot\rangle : t \in T,\ \langle t - t^*, X_i\rangle = 0 \text{ for every } 1 \le i \le N\} \\ &= \{\langle t, \cdot\rangle : t \in T,\ \Gamma(t - t^*) = 0 \text{ for every } 1 \le i \le N\}.\end{aligned}$$

Thus,

$$\mathbb{V}_{t^*,\sigma} \subset \langle t^*, \cdot\rangle + \{\langle u, \cdot\rangle : u \in \ker(\Gamma) \cap (T - T)\}.$$

If T is convex and centrally-symmetric, then $T - T = 2T$. Hence, for every possible $t^* \in T$,

$$\mathbb{V}_{t^*,\sigma} \subset \langle t^*, \cdot\rangle + \{\langle u, \cdot\rangle : u \in \ker(\Gamma) \cap 2T\}.$$

Finally, if X happens to be isotropic, implying that $\|\langle t, \cdot\rangle\|_{L_2} = \|t\|_2$ for every $t \in \mathbb{R}^d$, the maximal L_2 diameter of a version space in T is at most the ℓ_2 diameter of $\ker(\Gamma) \cap 2T$. *This parameter appears frequently in the study of the Gelfand widths of a convex body [**?**], and we will encounter it again in what follows.*

Intuitively, a large version space appears when F is 'rich' around f^*, since distinguishing when $f \ne f^*$ becomes easier the larger $\|f - f^*\|_{L_2}$ is. Indeed, under mild assumptions, the differences between f and f^* are easier to expose using random sampling the further f and f^*

are from each other. Therefore, one should expect a balancing act between the radius r and the 'size' of the set $\{f \in F : \|f - f^*\|_{L_2} = r\}$: the larger r is, and unless $\{f \in F : \|f - f^*\|_{L_2} = r\}$ is 'very rich', it is likely that set does not intersect $\mathbb{V}_{f^*,\sigma}$. Having said that, identifying the right notion of size and obtaining quantitative estimates on the 'critical level' r is a highly nontrivial task. As we explain in what follows, that notion is based on a fixed point, denoted by $r_{\mathbb{Q}}$, which finely balances the radius r and the 'size' of the set $\{f \in F : \|f - f^*\|_{L_2} \leq r\}$. We show that the fixed point is determined by the geometry structure of a typical coordinate projection $P_\sigma F$ (see Section 5.4 for more details).

The role of $r_{\mathbb{Q}}$ is to upper bound the smallest distance between an arbitrary f and f^* that still guarantees one can distinguish between the functions using the given data. More accurately, we show that there is a data-dependent procedure Φ, such that, with high probability, for any $f \in F$ that satisfies $\|f - f^*\|_{L_2} \geq cr_{\mathbb{Q}}$, one has

$$\Phi((X_i)_{i=1}^N, f, f^*) \gtrsim \|f - f^*\|_{L_2}^2.$$

Thus, not only is the L_2 diameter of a typical version space $\mathbb{V}(f^*, \sigma)$ smaller than cr, but, in fact, one has much more: for larger distances one can derive a uniform one-sided estimate of $\|f - f^*\|_{L_2}$ using the given data.

Remark 5.20. *These observation, in themselves, do not imply that the typical diameter of the version space is a lower bound on $\mathcal{E}_e$ in noise-free problems. In fact, there are known examples where that is not the case, though in general the situation is not fully understood. We refer the reader to the appendix of [?] for more details.*

Once the target Y 'moves away' from F, there are other reasons why statistical errors are likely to occur when trying to determine the signs of $\mathbb{E}(f(X) - Y)^2 - \mathbb{E}(h(X) - Y)^2$ using the sample $(X_i, Y_i)_{i=1}^N$. Again, when considering only 'interesting pairs' where $h = f^*$, one has that for every $f \in F$,

$$(*) = \mathbb{E}(f(X) - Y)^2 - \mathbb{E}(f^*(X) - Y)^2 = \mathbb{E}(f - f^*)^2(X) + 2\mathbb{E}(f^*(X) - Y) \cdot (f - f^*)(X).$$

In other words, exhibiting that the difference is positive can be achieved by obtaining an accurate data-dependent estimate on the two terms: the *quadratic term* $\mathbb{E}(f - f^*)^2(X)$ and the *multiplier term* $2\mathbb{E}(f^*(X) - Y) \cdot (f - f^*)(X)$.

One can easily be convinced that the reason why statistical estimates of $(*)$ are correct (i.e. positive) is the domination of (estimates of) the quadratic term over (estimates of) the multiplier one. Indeed, the latter could easily be 0, for example, when $Y = f^*(X) + W$ for a centred W that is independent of X. In such a case, when the problem consists of independent additive noise, statistical estimates of the multiplier term are likely to fluctuate around 0 and in particular can be negative with constant probability. Thus, the only reason why one can "guess" the signs of $(*)$ correctly is because the positive contribution of the quadratic term is dominant.

The quadratic term is the same as in the noise-free problem, and the intuition behind $r_{\mathbb{Q}}$ is that one can accurately "guess" $\mathbb{E}(f - f^*)^2(X)$ using the given sample when the $l_2(\mu)$ distance between f and f^* is at least $\sim r_{\mathbb{Q}}$. With that in mind, satisfactory control of the multiplier term means that for a well chosen $r_{\mathbb{M}}$, if $\|f - f^*\|_{L_2} \geq r_{\mathbb{M}}$, there is a way of "guessing" $\mathbb{E}(f^*(X) - Y) \cdot (f - f^*)(X)$ sufficiently accurately. To that end one has to capture the way the "noise" $\xi = f^*(X) - Y$ interacts with the functions $(f - f^*)(X)$: a strong interaction can cause fluctuations that prevent any procedure from accurately guessing the

means $\mathbb{E}(f^*(X)-Y)\cdot(f-f^*)(X)$. And, one has to keep in mind that "guess" means *uniformly* in $f \in F$.

Clearly, the noise level, captured by $\|\xi\|_{L_2}$ should have an important effect on the interaction between ξ and functions in F; and one should expect the signal-to-noise ratio to come into play as well.

We show that just as in the noise-free case, $r_{\mathbb{M}}$ is reflected through the geometry of (subsets of) $P_\sigma F$.

A problem is considered "low-noise" if $r_{\mathbb{Q}}$ dominates $r_{\mathbb{M}}$; it belongs to the "high -noise" regime if the reverse inequality holds. And, at least at the intuitive level, once the noise level $\|\xi\|_{L_2}$ is small enough, that is, once Y is close enough to F, the error should tend to the noise-free solution. Specifically, when exact recovery in the noise-free case is possible (i.e., when it is possible to have $\mathcal{E}_e = 0$), the error should tend to 0 with $\|\xi\|_{L_2}$.

The formal definition of $r_{\mathbb{Q}}$ and $r_{\mathbb{M}}$ is presented in the next section, as are some insights regarding a geometric interpretation of the two. The statistical motivation behind their definition is explained in Section 5.5.

It will be a long journey before one is able to justify in full why these parameters are the right choices, and discuss whether they lead to a sharp characterization of the accuracy/confidence tradeoff in learning problems.

5.4 Complexity terms and fixed points

The question "what is the right way of measuring the size of a set?" does not have a single answer. As one might expect, an answer depends on the context in which the question is asked. Before we defined $r_{\mathbb{Q}}$ and $r_{\mathbb{M}}$ formally, let us describe a natural way of measuring the size of a subset of $\mathbb{R}^N$. The notion in question has originated from asymptotic geometric analysis and is based on the idea of a *mean width*.

Definition 5.21. *Let $T \subset \mathbb{R}^N$ and set $\mathcal{Z}$ to be a centred random vector. The mean-width of T with respect to $\mathcal{Z}$ is*

$$\mathbb{E} \sup_{t\in T} \left|\langle \mathcal{Z}, t\rangle\right|. \tag{5.14}$$

The reason for the name "mean-width" is simple: for any $z \in S^{N-1}$, $2\sup_{t\in T}|\langle z,t\rangle|$ is the width of T in the direction defined by z. If one would give equal weight to every direction on the Euclidean unit sphere and average the width over all directions, the natural notion of mean-width is

$$\int_{S^{N-1}} \sup_{t\in T} \left|\langle z,t\rangle\right| d\sigma(z), \tag{5.15}$$

with integration performed with respect to the Haar measure on the Euclidean unit sphere.

There is no reason to restrict oneself to directions, nor to a uniform distribution on the sphere. Instead, (5.14) is natural analogue when considering a general random vector $\mathcal{Z}$ instead of the uniform distribution on the Euclidean unit sphere.

Example 5.22. *A natural random vector that closely resembles the uniform measure on the sphere (up to normalization) is the standard gaussian vector $G = (g_i)_{i=1}^N$. By rotation invariance and the uniqueness of the Haar measure, $G/\|G\|_2$ is uniformly distributed on S^{N-1}, and one may verify that*

$$\int_{S^{N-1}} \sup_{t\in T} \left|\langle z, t\rangle\right| d\sigma(z) \sim \frac{1}{\sqrt{N}} \mathbb{E}\left|\langle G, t\rangle\right|.$$

Exercise 11. *Complete the details of Example 5.22.*

The most important random vector from the perspective of statistical learning theory is the Bernoulli random vector, $\mathcal{E} = (\varepsilon_i)_{i=1}^N$, that has independent, symmetric $\{-1,1\}$-valued random variables as coordinates (i.e., it is uniformly distributed in $\{-1,1\}^N$. The random vector $\mathcal{E}$ is isotropic and its Euclidean norm is $\|\mathcal{E}\|_2 = \sqrt{N}$. One should note that the mean-width of T with respect to G and the mean width with respect to $\mathcal{E}$ need not be equivalent. For one, the Bernoulli mean-width is obtained by averaging the width with respect vectors that are the vertices of the combinatorial cube $\{-1,1\}^N$, implying that it is far from being rotationally invariant.

Exercise 12. *Give an example of a set $T \subset \mathbb{R}^N$ for which $\mathbb{E}\sup_{t\in T}\left|\langle \mathcal{E}, t\rangle\right|$ and $\mathbb{E}\sup_{t\in T}\left|\langle G, t\rangle\right|$ are not equivalent.*

Remark 5.23. *For a set $T \subset \mathbb{R}^N$ set*

$$T^\circ = \left\{ u \in \mathbb{R}^N : \sup_{t\in T} |\langle t, u\rangle| \le 1 \right\}.$$

It is straightforward to verify that if T is bounded and the symmetric convex hull of T has a nonempty interior in R^N then that convex hull is the unit ball B_X of a normed space $X = (\mathbb{R}^N, \|\ \|)$. In that case, T° is the unit ball of the dual space X^*, and in particular, $\mathbb{E}\sup_{t\in T}\left|\langle \mathcal{Z}, t\rangle\right| = \mathbb{E}\|Z\|_{X^*}$. At times we denote the latter by $\mathbb{E}\|Z\|_{T^\circ}$.

The sets T that are of interest in statistical learning theory are not arbitrary subsets of $\mathbb{R}^N$; they are the random coordinate projections of localizations of F around f^*, and since they are the result of iid sampling, they have a rather special structure. To define those sets, let

$$\mathrm{star}(F, h) = \{\lambda f + (1-\lambda)h : \ f \in F, \text{ and } 0 \le \lambda \le 1\},$$

which is the star-shaped hull of F and h; i.e., the set $\mathrm{star}(F,h)$ consists of all the intervals whose end-point are h and any point in F.

Taking the star-shaped hull of a set around a designated point might appear strange at a first glance. It is actually a technical trick used to increase the regularity of F around f^* while at the same time not increasing the size of F by too much. At an intuitive level, one may think of the star-shaped hull as adding one extra dimension to F, and as a result, other standard notions of size (e.g. the covering numbers) do not increase by much. At the same time, the added regularity is very useful: taking the star-shaped hull makes the set relatively richer closer to its centre. Indeed, if $\rho \ge r$, any point in $\mathrm{star}(F,h)$ whose distance from h is ρ, has a "scaled-down copy" whose distance from h is r (this is very easy to see when $h = 0$:

if $u \in \text{star}(F,0)$ and $\|u\|_{L_2} = \rho$, then since $0 < r/\rho < 1$, $(r/\rho)u \in \text{star}(F,0)$ and its norm is r). In other words, the subsets of the sphere

$$\{u/\|u\|_{L_2} : u \in \text{star}(F-h,0) \cap \rho S(L_2)\}$$

become bigger as ρ decreases.

This regularity plays an important role in many of the results we present. A useful observation to keep in mind is as follows:

> If the goal is to show that a certain property is satisfied by any function in F whose distance from f^* is at least r, and if that property is homogeneous in $f - f^*$ in some appropriate sense, it is enough to show that the property holds for any function in $\text{star}(F - f^*, 0) \cap rS(L_2)$.

Indeed, consider $u = f - f^*$ such that $\|u\|_{L_2} = \|f - f^*\|_{L_2} = \rho \geq r$. Then $v = (r/\rho)u \in \text{star}(F - f^*, 0) \cap rS(L_2)$; therefore, v satisfies the wanted property. If the property is preserved by multiplying v by $\alpha > 1$, then u must satisfy the property as well.

After one makes the sets more regular by taking the star-shaped hull around a designated centre (f^*), the next step is to shift that centre to 0 and localize the resulting set by intersecting it with balls in $L_2(\mu)$. For the next definition recall that D denotes the unit ball in $L_2(\mu)$.

Definition 5.24. *For $f^* \in F$ set*

$$F_{f^*,r} = \text{star}(F - f^*, 0) \cap rD = \{h = \lambda(f - f^*) : \ 0 \leq \lambda \leq 1, \ \|h\|_{L_2} \leq r\}.$$

Let $\mathcal{E} = (\varepsilon_i)_{i=1}^N$ be the standard Bernoulli vector, and assume that it is independent of $X_1, ..., X_N$. Consider the functions

$$\phi_{\mathbb{Q}}(r, f^*, N) = \mathbb{E} \sup_{u \in F_{f^*,r}} \left| \frac{1}{N} \sum_{i=1}^N \varepsilon_i u(X_i) \right|, \tag{5.16}$$

and for $\xi_i = f^*(X_i) - Y_i, \quad 1 \leq i \leq N$, put

$$\phi_{\mathbb{M}}(r, f^*, N) = \mathbb{E} \sup_{u \in F_{f^*,r}} \left| \frac{1}{N} \sum_{i=1}^N \varepsilon_i \xi_i u(X_i) \right|. \tag{5.17}$$

Thus, $\phi_{\mathbb{Q}}$ is the expectation of the generalized mean-width—relative to the Bernoulli vector—, of the random coordinate projection $P_\sigma F_{f^*,r}$ of the localization $F_{f^*,r}$. Indeed,

$$\phi_{\mathbb{Q}}(r, f^*, N) = \frac{1}{N} \mathbb{E}_X \left(\mathbb{E}_\varepsilon \sup_{v \in P_\sigma F_{f^*,r}} \left| \sum_{i=1}^N \varepsilon_i v_i \right| \right).$$

At the same time, $\phi_{\mathbb{M}}(r, f^*, N)$ is the generalized mean-width of a weighted version of $P_\sigma F_{f^*,r}$, with the weights $(\xi_i)_{i=1}^N$ calibrating the effect of the noise.

As we explain in the next section, the critical levels $r_{\mathbb{Q}}$ and $r_{\mathbb{M}}$ are given by the fixed point conditions $\phi_{\mathbb{Q}}(r, f^*, N) \sim r$ and $\phi_{\mathbb{M}}(r, f^*, N) \sim r^2$, respectively. Before diving into

more technicalities, let us give a geometric interpretation of those conditions. Although at this point the interpretation will seems like hand-waving, the reader is invited to take a leap of faith with us; this interpretation will prove to be very useful and quite accurate.

The idea of structure preservation loosely implies that at the right scale, what one "sees" in the real-world of $L_2(\mu)$ is exhibited by a typical random coordinate projection, and

$$\|f-h\|_{L_2^\sigma} = \left(\frac{1}{N}\sum_{i=1}^N (f-h)^2(X_i)\right)^{1/2} \sim \|f-h\|_{L_2}.$$

Armed with that belief, and because $F_{f^*,r} \subset rD$, the "complexity" of $N^{-1/2}P_\sigma F_{f^*,r}$ should not exceed that of a Euclidean ball of radius proportional to r. If the complexity happens to be larger, that would hint that the sample σ, which is used to define the mapping from $L_2(\mu)$ to L_2^σ, distorts the structure of $F_{f^*,r} \subset rD$ by too much to be of any use. Since the Bernoulli mean-width of rB_2^N is $\sim r\sqrt{N}$, the intuitive condition on $r_{\mathbb{Q}}$ is that it is the smallest r for which

$$\mathbb{E}_X\left(\mathbb{E}_\varepsilon \sup_{v\in P_\sigma F_{f^*,r}}\left|\sum_{i=1}^N \varepsilon_i \frac{v_i}{\sqrt{N}}\right|\right) \lesssim \sqrt{N}r. \tag{5.18}$$

And that is precisely the fixed point condition that is used to define $r_{\mathbb{Q}}$:

> The fixed point $r_{\mathbb{Q}}$ is the smallest radius r such that, for a typical sample σ, $N^{-1/2}P_\sigma F_{f^*,r}$ is not "richer" than the Euclidean ball in $\mathbb{R}^N$ of radius cr. Moreover, by the star-shape property, if $r > r_{\mathbb{Q}}$ then $N^{-1/2}P_\sigma F_{f^*,r}$ will be "less rich" than $\sim rB_2^N$, and if $r < r_{\mathbb{Q}}$, it will be at least as rich as $\sim rB_2^N$.

Naturally, one may ask if the fact that at the level $r_{\mathbb{Q}}$, $N^{-1/2}P_\sigma F_{f^*,r}$ becomes as rich as crB_2^N has anything to do with geometry; specifically, if $N^{-1/2}P_\sigma F_{f^*,r}$ has the same Bernoulli mean-width as crB_2^N implies that one can find a Euclidean ball of radius r "living inside" $N^{-1/2}P_\sigma F_{f^*,r}$. The remarkable answer to that is *yes*[4]—up to the right understanding of what "living inside" means.

The situation regarding $r_{\mathbb{M}}$ is a little more subtle, even at the level of hand-waving, because of the normalization of r^2 in the fixed point condition. Again, let us write that condition as

$$\mathbb{E}_X\left(\mathbb{E}_\varepsilon \sup_{v\in P_\sigma F_{f^*,r}}\left|\sum_{i=1}^N \varepsilon_i \frac{\xi_i}{\|\xi\|_{L_2}}\frac{v_i}{\sqrt{N}}\right|\right) \sim \frac{r}{\|\xi\|_{L_2}}\sqrt{N}r$$

and consider the differences between it and (5.18). The Bernoulli mean-width is of the set

$$W = \left\{\left(\frac{\xi_i}{\|\xi\|_2}\frac{v_i}{\sqrt{N}}\right)_{i=1}^N : v \in p_\sigma F_{f^*,r}\right\};$$

the weights $(\xi_i/\|\xi\|_{L_2})_{i=1}^N$ distort the geometry of $N^{-1/2}P_\sigma F_{f^*,r}$ in coordinate directions when either $\xi_i/\|\xi\|_{L_2}$ is very large or when it is very close to 0. Having said that, one may show that

[4]The "yes" is valid under some additional assumptions on F and X; and a slightly weaker statement is true in full generality—though we conjecture that the weaker statement could be improved.

fundamentally, that distortion is not devastating (although it could still have a meaningful impact). A more significant issue is the extra multiplicative factor of $r/\|\xi\|_{L_2}$ appearing in the definition of the fixed point. It captures the worst *signal-to-noise ratio* one encounters when $\|f - f^*\|_{L_2} \geq r$. As a result, the fixed point is not defined by comparing the Bernoulli mean width of W with that of a Euclidean ball of radius $\sim r$, but rather with a Euclidean ball whose radius is multiplied by the worst signal-to-noise ratio.

In the next section we will explain the statistical motivation behind the definitions of the fixed points $r_{\mathbb{Q}}$ and $r_{\mathbb{M}}$, and formulate the main result we shall present in these notes.

5.5 The satisfactory sample complexity

In this section we give further insights to the reasons behind the definitions of the fixed points, and following that, formulate the main result of these notes. The key notion here is the *satisfactory sample complexity*.

The satisfactory sample complexity is defined using a list of obstructions, which are, in some sense, trivial. Overcoming each one of those obstacles is something one would expect of any reasonable learning procedure. On the other hand, overcoming each obstruction comes at a price:

- A certain geometric obstruction on the class F forces one to consider procedures that need not take values in F.
- Overcoming some trivial statistical obstructions requires a minimal number of sample points.

> At an intuitive level, the main result of [**?**] is that the sample complexity needed to overcome the trivial statistical obstructions actually suffices (up to some absolute multiplicative constant) for the solution of an (almost) arbitrary learning problem. And because of the geometric obstruction, the solution is carried out using a procedure that is allowed to take values outside the given class.

Here, for the sake of simplicity, we only explore the case in which F is convex, which, as we show in what follows, implies that no geometric obstructions exist.

Let us begin by describing the trivial obstructions we have in mind.

A geometric obstruction

In the standard (*proper*) learning model the procedure is only allowed to take values in the given class F. At a first glance this restriction seems to be completely reasonable; after all, the learner's goal is to find a function that mimics the behaviour of the best function in F, and there is no apparent reason to look for such a function outside F. However, a more careful consideration shows that this restriction comes at a high cost:

Example 5.25. *Let $F = \{f_1, f_2\}$ be defined on a probability space (Ω, μ), and as always, X is distributed according to μ. Fix an integer N and set Y to be a 'noisy' $\sim 1/\sqrt{N}$-perturbation*

of the midpoint $(f_1(X)+f_2(X))/2$*, that is slightly closer in* L_2 *to* $f_1(X)$ *than to* $f_2(X)$*. Then, given samples* $(X_i, Y_i)_{i=1}^N$*, any proper procedure* Φ *will necessarily make the wrong choice with probability* $1/10$*; that is, with probability at least* $1/10$*,* $\Phi((X_i, Y_i)_{i=1}^N) = f_2$ *and on that event, the excess risk is* $\mathcal{E}_p \sim 1/\sqrt{N}$*. Thus, the best sample complexity estimate one can hope for when using a proper learning procedure is* $\sim 1/\varepsilon^2$*.*

A proof of this standard fact may be found, for example, in [**?**, **?**].

Example 5.25 serves as a strong indication of a general phenomenon: there are seemingly innocent-looking problems, including ones involving classes with a finite number of functions (in this example, only two functions...), in which the sample complexity is significantly higher than what one would expect given the size of the class. The reason for such *slow rates*, or equivalently, for large sample complexities, is that the 'location' of the target relative to the class is not favourable. As will be clarified in Section 5.6, the notion of a favourable location is characterized by a convexity condition on the excess risk functional:

Definition 5.26. *A triplet* (F, X, Y) *satisfies a convexity condition with constant* $B > 0$ *with respect to the squared loss, if for every* $f \in F$*,*

$$\|f - f^*\|_{L_2}^2 \leq B\left(\mathbb{E}(f(X) - Y)^2 - \mathbb{E}(f^*(X) - Y)^2\right).$$

Note that the convexity condition implies that the minimizer in F of the risk functional $f \to \mathbb{E}(f(X) - Y)^2$ is unique. And, one may also verify that under a convexity condition, the set of functions in F that 'almost minimize' the risk consists only of perturbations of f^*. For more information on the convexity condition, see [**?**] and Section 5.6.

If F happens to be convex then for any target Y, the triplet (F, X, Y) satisfies the convexity condition with $B = 1$. In particular, there is no geometric obstruction when dealing with a convex class F, which makes the analysis of learning problems involving convex classes significantly simpler than in the general case.

The following exercise requires some knowledge of Functional Analysis:

Exercise 13. *Show that if* $F \subset L_2(\mu)$ *is closed and convex and* $Y \in L_2$*, then:*

(1) $f^* = \operatorname{argmin}_{f \in F} \mathbb{E}(f(X) - Y)^2$ *exists and is unique.*

(2) *The triplet* (F, X, Y) *satisfies the convexity condition with constant* $B = 1$*.*

(3) *What features of the* L_2 *norm are used in* (1) *and* (2)*? Can you deduce a similar result with respect to other loss functions?*

Remark 5.27. *The focus of [**?**] is on general triplets* (F, X, Y)*, making it is impossible to guarantee that the unknown target* Y *is in a favourable location, and therefore, there is no convexity condition at one's disposal. As a result, to have any hope of addressing the geometric obstruction one must remove the restriction that the procedure is proper.*

Statistical obstructions

A natural way of finding generic statistical obstructions is by identifying reasons why a statistical procedure may make mistakes. As was explained in Section 5.3.1, there are two simple sources of error:

- *Intrinsic errors and the low-noise regime:* When F is 'rich' close to the true minimizer f^*, it is difficult to 'separate' class members with the limited data the learner has, even if $Y = f^*(X)$. The problem is compounded with the introduction of low-level noise, e.g., by setting $Y = f^*(X) + W$ where W is a centered random variable that is independent of X and has a small variance relative to the wanted accuracy. One way of addressing learning problems with low-level noise is by ensuring that in the noise-free (realizable) problem, the version space has a small diameter in L_2.

- *External errors and the high-noise regime:* When the noise level becomes significant and $\|f^*(X) - Y\|_{L_2}^2$ is larger than the wanted accuracy ε^2, interactions between Y and class members can cause distortions. These interactions make functions that are close to f^* indistinguishable and forces the learning procedure to make mistakes in the choices it makes.

The "statistical obstructions" we refer to are the intrinsic and external errors caused by two specific collections of triplets involving F and X (keeping in mind that the learner does not know the distribution of X), and with targets Y that belong to $\mathcal{Y}_{\min}$, i.e., either:

(1) *Realizable targets*; that is, targets of the form $Y = f_0(X)$ where $f_0 \in F$.

(2) *Additive, independent gaussian noise*; that is, targets of the form $Y = f_0(X) + W$, where $f_0 \in F$ and W is a centred gaussian random variable, independent of X and with variance σ^2.

The idea is that a satisfactory statistical procedure should be able to deal with targets in $\mathcal{Y}_{\min}$. Moreover, the sample complexity needed to overcome the intrinsic errors caused by targets in (1) by ensuring that each version space has a small diameter in L_2, and the external errors caused by targets in (2) is a rather minimal price. One should be willing to pay such a price if the goal is addressing general estimation and prediction problems that are likely to be far more complex than these trivial targets.

The satisfactory sample complexity is a relatively sharp upper bound on the sample size needed to overcome the trivial obstructions with constant confidence.

Definition 5.28. *For a triplet $T = (F, X, Y)$ let*

$$N_{\text{int}}(T, \varepsilon, \kappa) = \min\left\{N : \phi_{\mathbb{Q}}(\sqrt{\varepsilon}, f^*, N) \leq \kappa r\right\}, \tag{5.19}$$

and

$$N_{\text{ext}}(T, \varepsilon, \kappa) = \min\left\{N : \phi_{\mathbb{M}}(\sqrt{\varepsilon}, f^*, N) \leq \kappa r^2\right\}. \tag{5.20}$$

Theorem 5.29. *Let $T = (F, X, Y)$ be a triplet that satisfies the convexity condition with constant B. Under minimal conditions on T, there are constants c_1 and c_2 that depend on B (and on the minimal conditions), and a procedure that, given a sample of cardinality at least*

$$N = N_{\text{int}}(T, \varepsilon, c_1) + N_{\text{ext}}(T, \varepsilon, c_2)$$

such that, with probability at least 0.5*,*

$$\|\hat{f} - f^*\|_{L_2} \leq \sqrt{\varepsilon},$$

(and 0.5 *can be modified to any fixed constant).*

Theorem 5.29 is harder than what might suspect and the procedure in question is *empirical risk minimization* (ERM) which is studied in detail in Section 6.1. The proof when F is convex was established in [**?**]. We present highlights of the argument at a later point in these notes, after the necessary machinery has been developed. Also, we shall not present a proof of the fact that a sample size of $N = N_{\text{int}}(T, \varepsilon, c_1) + N_{\text{ext}}(T, \varepsilon, c_2)$ is not far from the lower bound on the necessary sample size for overcoming the trivial obstacles. For more information on that we refer the reader to the appendix in [**?**].

A sample complexity of $N \geq N_{\text{int}}(T, \sqrt{\varepsilon}, c_1) + N_{\text{ext}}(T, \sqrt{\varepsilon}, c_2)$ might have been a reasonable candidate for the satisfactory sample complexity had one been interested in the constant confidence level. But since the interest is in higher confidence levels as well, there is a need for an additional term whose origin is the following lower bound:

Theorem 5.30. *[**?**] There are absolute constants c_1 and c_2 for which the following holds. Let W be a centred gaussian random variable that is independent of X and with variance σ^2, and assume that F contains an interval $[f_1, f_2]$ such that $\|f_1 - f_2\|_{L_2} \geq c_1\sigma$. Then for any learning procedure Φ there is some $f_0 \in F$ such that the sample complexity Φ needs in order to perform with accuracy ε and confidence $1-\delta$ when faced with the triplet $(F, X, f_0(X)+W)$ is at least*

$$c_2 \frac{\sigma^2}{\varepsilon} \cdot \log\left(\frac{2}{\delta}\right).$$

Remark 5.31. *Note that in Theorem* (5.30)*, $\sigma = \|f^*(X) - Y\|_{L_2}$ and that the triplet (F, X, Y) satisfies the convexity condition with constant $B = 1$.*

With all these ingredients in place, let us define the satisfactory sample complexity:

Definition 5.32. *An unrestricted procedure performs with a satisfactory sample complexity for a collection of triplets $\mathcal{T}$ if there are constants c_1 and c_2 such that for every triplet $T = (F, X, Y) \in \mathcal{T}$, the procedure performs with accuracy ε and confidence $1 - \delta$ with sample complexity*

$$N = N_{\text{int}}(T, \sqrt{\varepsilon}, c_1) + N_{\text{ext}}(T, \sqrt{\varepsilon}, c_1) + c_2 \left(\frac{\|f^*(X) - Y\|_{L_2}^2}{\varepsilon} + 1\right) \log\left(\frac{2}{\delta}\right). \tag{5.21}$$

At a first glance, the satisfactory sample complexity seems a very optimistic notion. Not only is its motivation a collection of trivial obstructions, $N_{\text{int}} + N_{\text{ext}}$ is known to be valid upper estimates only at the constant confidence level and for estimation problems that involve classes that satisfy a convexity condition. The only term in (5.21) that depends on the wanted confidence level is based on a lower bound that is essentially one dimensional.

A more tangible indication that achieving the satisfactory sample complexity may prove to be difficult (if at all possible) is the problem of linear regression in $\mathbb{R}^d$:

Example 5.33. *Let $F = \{\langle t, \cdot \rangle : t \in \mathbb{R}^d\}$ be the class of linear functionals on $\mathbb{R}^d$. Assume that X is an isotropic random vector in $\mathbb{R}^d$, (i.e. for any $t \in \mathbb{R}^d$, $\mathbb{E}\langle X, t\rangle^2 = \|t\|_2^2$). Let $Y = \langle t_0, X\rangle + W$, where W is centred, independent of X and satisfies that $\mathbb{E}W^2 = \sigma^2$. It is*

straightforward to verify that $f^* = \langle t_0, \cdot \rangle$ *and that for every* $r > 0$,

$$F_{f^*,r} = \left\{ \langle t, \cdot \rangle : t \in rB_2^d \right\},$$

where B_2^d *is the Euclidean unit ball. Moreover, one can show that*

$$\mathbb{E} \sup_{u \in F_{f^*,r}} \left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i u(X_i) \right| \leq r\sqrt{\frac{d}{N}}, \tag{5.22}$$

and

$$\mathbb{E} \sup_{u \in F_{f^*,r}} \left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i (f^*(X_i) - Y_i) u(X_i) \right| \leq \sigma r\sqrt{\frac{d}{N}}. \tag{5.23}$$

Hence, the satisfactory sample complexity in this case satisfies that

$$\lesssim \max\left\{ d,\ \sigma^2 \frac{d}{\varepsilon},\ \sigma^2 \frac{\log(2/\delta)}{\varepsilon} \right\}$$

which is the minimax estimate for linear regression in $\mathbb{R}^d$. *However, finding a procedure that performs with that sample complexity when* X *and* W *are heavy-tailed has been a long-standing open problem; it was resolved only recently, in [**?**].*

Exercise 14. *Prove* (5.22) *and* (5.23).

The main contribution of [**?**] is identifying a procedure that performs with the satisfactory sample complexity in the sense of Definition 5.32, with one modification: for a problem involving the triplet (F, X, Y), one requires a sample size of

$$\sim \sup_{T'} \left(N_{\text{int}}(T', \sqrt{\varepsilon}, c_1) + N_{\text{ext}}(T', \sqrt{\varepsilon}, c_1) \right) + c_2 \left(\frac{\|f^*(X) - Y\|_{L_2}^2}{\varepsilon} + 1 \right) \log\left(\frac{2}{\delta}\right) \tag{5.24}$$

where the supremum is taken with respect to all triplets $T' = (H, X, Y)$, for classes H that satisfy

$$H \subset \frac{F + F}{2} = \left\{ \frac{f + h}{2} : f, h \in F \right\}, \quad \text{and} \quad f^* \in H.$$

However, here we focus on the simpler problem, in which the class F is convex, and in which case the following holds:

If F is convex, then under minimal assumptions on the triplet (F, X, Y) there is a procedure that attains the satisfactory sample complexity: a sample size of

$$\sim N_{\text{int}}(T', \sqrt{\varepsilon}, c_1) + N_{\text{ext}}(T', \sqrt{\varepsilon}, c_1) + \left(\frac{\|f^*(X) - Y\|_{L_2}^2}{\varepsilon} + 1 \right) \log\left(\frac{2}{\delta}\right)$$

suffices to produce $\hat{f}$ such that with probability at least $1 - \delta$,

$$\|\hat{f} - f^*\|_{L_2} \leq \sqrt{\varepsilon}, \quad \text{and} \quad \mathbb{E}\left((\hat{f}(X) - Y)^2 | (X_i, Y_i)_{i=1}^N \right) \leq \mathbb{E}(f^*(X) - Y)^2 + \varepsilon.$$

An alternative formulation of this estimate is given in terms of the accuracy/confidence tradeoff:

If F is convex, then under minimal assumptions on the triplet (F, X, Y) there is a procedure that satisfies the following: given a sample size N, if $r \geq \max\{r_{\mathbb{Q}}, r_{\mathbb{M}}\}$, then with probability at least

$$1 - 2\exp\left(-c\min\left\{\frac{r^2}{\sigma^2}, 1\right\}\right),$$

one has that

$$\mathbb{E}\left((\hat{f}(X) - Y)^2 | (X_i, Y_i)_{i=1}^N\right) \leq \mathbb{E}(f^*(X) - Y)^2 + r^2.$$

Here, $\sigma = \|f^*(X) - Y\|_{L_2}$ and the constants used in the definition of $r_{\mathbb{Q}}$ and $r_{\mathbb{M}}$ depend only on the "minimal assumptions" on the triplet.

At the heart of the results from [**?**] and [**?**] is that the sample complexity/tradeoff estimates hold in heavy-tailed situations, in which there is no hope of having satisfactory concentration of empirical means around the true ones. In particular, ERM does not have a chance of performing well given only the satisfactory sample complexity.

Heavy-tailed problems were considered totally out of reach until very recently. Up to that point, an overwhelming majority of the work on statistical learning theory focused on situations in which empirical means did concentrate well around their true means. This is not to say that solving problems where there is enough concentration is easy; far from it: such questions have been left unanswered for ~ 40 years and their solutions require sophisticated mathematical machinery. Although that *bounded framework* is not the main focus of these notes, it is still highly instructive to see how estimation and prediction problems can be resolved when there is sufficient concentration of empirical means. We present two results that are based on concentration in the next chapter.

5.6 The convexity condition

Let $F \subset L_2$ be compact, denote by $N(F)$ the set of functions in L_2 that have more than a unique best approximation in F, let $Y \notin \overline{N(F)}$ and set f^* to be the best approximation of Y in F. Put

$$r = d(Y, N(F)) \quad \text{and set} \quad R = \|Y - f^*\|_{L_2}.$$

Theorem 5.34. *There exists an absolute constant C for which the following holds. If $\mathcal{L}_f = (f(X) - Y)^2 - (f^*(X) - Y)^2$ is the excess squared loss functional, then*

$$\|f - f^*\|_{L_2}^2 \leq B\mathbb{E}\mathcal{L}_f$$

for $B = C(1 + R/r)$.

Remark 5.35. *The estimate one actually obtains is a little better—see Corollary 5.40.*

The proof of Theorem 5.34 is based on a general estimate in a uniformly convex Banach space, which is of independent interest and which we describe now.

Recall that a Banach space X is smooth if the norm is Gâteaux differentiable in any $x \neq 0$. In other words, for every x on the unit sphere of X there is a unique functional $x^* \in X^*$

such that $\|x^*\| = 1$ and $x^*(x) = 1$ (thus, there is a unique hyperplane $\{x : x^*(x) = 1\}$ that supports the unit ball of X, B_X, in x).

Definition 5.36. *A Banach space X is strictly convex if for every $x, y \in B_X$ one has $\|x+y\| < 2$. The space X is uniformly convex if there is a positive function $\delta_X(\varepsilon)$ such that for every $0 < \varepsilon < 2$ and every $x, y \in B_X$ for which $\|x - y\| \geq \varepsilon$, one has $\|x + y\| \leq 2 - 2\delta(\varepsilon)$. In other words,*

$$\delta_X(\varepsilon) = \inf \left\{ 1 - \frac{1}{2}\|x + y\| : \|x\|, \|y\| \leq 1, \ \|x - y\| \geq \varepsilon \right\}$$

is positive for $\varepsilon > 0$.

The function $\delta_X(\varepsilon)$ is called the modulus of convexity of X.

For basic properties of $\delta_X(\varepsilon)$ we refer the reader to [**?**], Chapter 8. Important facts that are used in what follows are that $\delta_X(\varepsilon)$ is an increasing function of ε, and that if $0 < \varepsilon_1 \leq \varepsilon_2 \leq 2$, then

$$\frac{\delta(\varepsilon_1)}{\varepsilon_1} \leq \frac{\delta(\varepsilon_2)}{\varepsilon_2}.$$

The proof of Theorem 5.34 is based on properties of the nearest point map (also known as the metric projection).

Definition 5.37. *Let X be a Banach space and put $F \subset X$. The nearest-point map is a set valued function $P_F : X \to 2^F$, defined by assigning to each $x \in X$ the set of nearest points to x in F with respect to the norm.*

For basic facts regarding the nearest point map, see, for example, [**?**] Chapter 2.2 and references therein.

Note that if F is compact then $P_F x$ is a nonempty set for every $x \in X$. It is also standard to verify that if X is strictly convex and F is convex then $P_F x$ consists of at most a single element.

Exercise 15. *Prove these facts.*

Of course, if F is not convex then some points in X may have more than a unique nearest point. In fact, under various assumptions on the norm, if F is compact, then $|P_F x| = 1$ for every x if and only if F is convex. In its full generality, when removing the compactness assumption, this equivalence is the famous problem of the convexity of Chebyshev sets. We refer the reader to [**?**] for a survey on this topic.

Recall that $N(F)$ is the set of points in X that have more than a unique best approximation in F, and we abuse notation by also setting $P_F : X \backslash N(F) \to F$.

Remark 5.38. *It is well known that the sets $N(F)$ are small; indeed, if F is compact and X is strictly convex then $N(F)$ is a G_δ set of the first category [**?**]. Thus, "most" points in X have a unique nearest element in F. For other results in this direction see also the survey [**?**]. .*

Throughout we use the following notation: for every $x, y \in X$ let

$$[x, y] = \{tx + (1 - t)y : 0 \leq t \leq 1\},$$

and set $\overrightarrow{x, y}$ to be the ray originating in x and passing through y.

The crucial geometric estimate needed in the proof is as follows. The triangle inequality shows that if $x \notin N(F)$, i.e., if x has a unique best approximation $P_F x \in F$, then for every $0 \leq t \leq 1$, $tx + (1 - t)P_F x$ has a unique best approximation in F—namely, $P_F x$.

Exercise 16. *Prove that indeed, for $0 < t < 1$, $tx+(1-t)P_Fx$ has a unique best approximation in F, and that is P_Fx.*

The heart of the argument is to show that if x is "far away" from $N(F)$, any point on the ray $\overrightarrow{P_Fx, x}$ of the form $x(t) = tx + (1-t)P_Fx$ for $1 < t < \lambda(x)$ is still projected (uniquely) to P_Fx—up to $\lambda(x) > 1$ that depends on the distance between x and $N(F)$.

This observation is connected with the so-called "solar" property of a set, introduced by Efimov and Stechkin [**?**, **?**, **?**] in the study of the metric projection onto convex sets. The proof of this "local" solar property is based on a fixed point argument similar to the one used by Vlasov in [**?**] (and at the same time, though with a slightly different flavor, by Klee in [**?**]).

Lemma 5.39. *Let F be a compact subset of a Banach space X. Consider $x \in X\backslash(\overline{N(F)} \cup F)$, set $r = d(x, N(F))$, $R = d(x, F)$ and for every $t \geq 0$ let $x(t) = tx + (1-t)P_Fx$. Then, for any $0 \leq t < 1 + r/R$, $P_Fx(t) = P_Fx$.*

The claim for $0 \leq t \leq 1$ follows immediately from Exercise 16. The nontrivial part of Lemma 5.39 is that if x is "far away" from $N(F)$, one can move further up the ray $\overrightarrow{P_Fx, x}$—beyond x and still be (uniquely) projected to P_Fx.

Exercise 17. *Show that if X is strictly convex and F is compact and convex, then for every $t \geq 0$, $P_Fx(t) = P_Fx$*

Exercise 17 is the so-called *solar property* of a convex set. The assertion of Theorem 5.39 is that F has some local solar property despite the fact it is not convex (and without any assumptions on the norm).

Proof of Lemma 5.39. Let $0 < \delta < r$ and set $B = B(x, r-\delta)$ to be the closed ball around x and of radius $r - \delta$. Put

$$t_0 = \sup\{t : \ x(t) \in B, \ P_Fx(t) = P_F(x)\} \tag{5.25}$$

and let $x_0 = x(t_0)$. By the uniqueness of the nearest point in F for points in B, the supremum in (5.25) is attained and $P_F(x_0) = P_Fx$. Clearly, $t_0 \geq 1$ and assume that $t < 1 + (r-\delta)/R$, i.e., that x_0 belongs to the interior of B.

Consider the function $\phi : F \to B$, given by

$$\phi(f) = \left(1 + \frac{r-\delta}{\|x-f\|}\right)x - \frac{r-\delta}{\|x-f\|}f,$$

and observe that ϕ maps each $f \in F$ to the unique point in $\overrightarrow{f, x} \cap \partial B$ for which $x \in [\phi(f), f]$.

Since $x \notin F$ then ϕ is continuous. Also, since $B \cap N(F) = \emptyset$ and F is compact then P_F is continuous on B. Thus, $\psi : B \to B$ given by $\psi(z) = \phi(P_Fz)$ is also continuous, and using the compactness of F once again, $\overline{\psi(B)}$ is compact and is contained in ∂B. By the Schauder-Tikhonov Fixed Point Theorem (see, e.g. [**?**], pg. 61) and the fact that $\psi(B) \subset \partial B$, there is some $z \in \partial B$ for which $\psi(z) = z$. Note that $z \notin \overrightarrow{P_Fx, x}$. Indeed, any fixed point on that ray must be projected onto P_Fx; on the other hand, the only candidate for a fixed point on $\overrightarrow{P_Fx, x}$ is on ∂B and "beyond" $x(t_0)$, which, by our assumption, is not projected onto P_Fx. Moreover, since $z \notin \overrightarrow{P_Fx, x}$ then $P_Fz \neq P_Fx$, because a fixed point projected onto P_Fx must be on the ray $\overrightarrow{P_Fx, x}$.

Thus, from the definition of ϕ and since z is a fixed point, $x \in [z, P_Fz]$. On the other hand, by Exercise 16, $P_Fz = P_Fx$, which is a contradiction.

The claim follows by taking $\delta \to 0$. ∎

Observe that what matters is actually not $d(x, N(F))$, but rather "how much further" up the ray $\overrightarrow{P_F x, x}$ can one go an still be outside $N(F)$; such points will be projected to $P_F x$ as well. To formulate that fact, let

$$\lambda^*(x) = \sup\left\{\lambda \geq 1 : \lambda x + (1-\lambda)P_F x \notin \overline{N(F)}\right\}. \tag{5.26}$$

Corollary 5.40. *Let X be a Banach space and let $F \subset X$ be a compact set. If $x \notin (F \cup \overline{N(F)})$ then for every $1 \leq t < \lambda^*(x)$, $P_F x(t) = P_F x$.*

Proof. Consider the functions $r(y) = d(y, N(F))$, $R(y) = d(y, F)$ and define a sequence $(x_n)_{n=1}^{\infty} \subset \overrightarrow{P_F x, x}$ as follows. Let $x_1 = x$ and set $x_2 = tx_1 + (1-t)P_F x_1$ for $t = 1 + r(x_1)/2R(x_1)$. By Lemma 5.39, $x_2 \in \overrightarrow{P_F x, x}$ and its unique nearest point in F is $P_F x$. Also,

$$R(x_2) = \left(1 + \frac{r(x_1)}{2R(x_1)}\right) R(x_1)$$

and

$$\begin{aligned} r(x_2) &\geq \left\| x\left(1 + \frac{r(x_1)}{2R(x_1)}\right) - x\left(1 + \frac{r(x_1)}{R(x_1)}\right)\right\| = \frac{r(x_1)}{2R(x_1)}\|x - P_F x\| \\ &= \frac{r(x_1)}{2} > 0. \end{aligned}$$

Since $x_2 \notin \overline{N(F)}$, one can repeat this argument for $x = x_2$ and so on. It is clear that $(x_n)_{n=1}^{\infty} \subset \overrightarrow{P_F x, x} \cap (X \backslash \overline{N(F)})$, that $P_F x_n = P_F x$ for every integer n, and that

$$R(x_{n+1}) = \left(1 + \frac{r(x_n)}{2R(x_n)}\right) R(x_n).$$

Moreover, relative to the natural order on $\overrightarrow{P_F x, x}$, $(x_n)_{n=1}^{\infty}$ is increasing, and set $x' = \sup_n x_n$ (where x' might be infinite). By the construction of $(x_n)_{i=1}^{n}$ it is evident that for every $z \in [x, x')$, $P_F z = P_F x$ and that $r(z) > 0$. Thus, there are two possibilities: if $x' = \infty$ then every point in $\overrightarrow{P_F x, x}$ is uniquely projected to $P_F x$ and the claim is trivially true. Otherwise, $(R(x_n))_{n=1}^{\infty}$ converges to a finite, positive limit, implying that $r(x') = \lim_{n\to\infty} r(x_n) = 0$. This observation combined with the fact that $r > 0$ on $[x, x')$ implies that $x' = x(t)$ for $t = \lambda^*(x)$. ∎

Next, we turn to the second component needed for the proof of the convexity condition.

Lemma 5.41. *Let X be a uniformly convex, smooth space and consider $w, y \in X$ and $\rho \in \mathbb{R}_+$ such that $\|y - w\| = \rho$. Let $0 < \theta < 1$ and set $w_\theta = (1-\theta)w + \theta y$. If z satisfies that $\|z - w\| \geq \rho$ then*

$$\|w_\theta - z\| - (1-\theta)\rho \geq 2\theta\|z - w\|\delta_X\left(\frac{\|z - y\|}{\|z - w\|}\right).$$

Note that the bound improves the larger θ is, i.e., the further w_θ is from w.

Proof. Without loss of generality one can assume that $w = 0$. Fix $z \neq y$ and by the assumption, $\|z\| \geq \|y\|$. Define the function

$$H(\theta) = \frac{\|w_\theta - z\|}{\|z\|} = \frac{\|\theta y - z\|}{\|z\|}$$

and observe that H is a convex function and $H(0) = 1$. Also, since X is smooth, H is differentiable in $\theta = 0$. Thus, $H(\theta) - H(0) = H(\theta) - 1 \geq H'(0)\theta$, implying that

$$H(\theta) - (1 - \theta) \geq \left(H'(0) + 1\right)\theta,$$

and to complete the proof one has to bound $H'(0)$ from below.

Applying the chain rule, $H'(0) = u^* \left(\frac{y}{\|z\|}\right)$, where u^* is the unique functional of norm one supporting the unit sphere in $-z/\|z\| \equiv u$. Let $v = y/\|z\|$ and since $\|u^*\| = 1$ then

$$u^*(u - v) \leq \|u - v\| \leq 2 - 2\delta_X(\|u + v\|) = 2 - 2\delta_X\left(\frac{\|y - z\|}{\|z\|}\right).$$

Clearly, $u^*(u) = 1$ and thus $-u^*(v) \leq 1 - 2\delta_X\left(\frac{\|y-z\|}{\|z\|}\right)$, implying that

$$H(\theta) - (1 - \theta) \geq \left(H'(0) + 1\right)\theta \geq 2\theta\delta_X\left(\frac{\|y - z\|}{\|z\|}\right).$$

Therefore

$$\|w_\theta - z\| - (1 - \theta)\rho \geq \|w_\theta - z\| - (1 - \theta)\|z\| \geq 2\theta\|z\|\delta_X\left(\frac{\|y - z\|}{\|z\|}\right).$$

■

Corollary 5.42. *Let X be a uniformly convex and smooth Banach space and assume that $F \subset X$ is compact. Let $x \in X\backslash(F \cup N(F))$, set $r = d(x, N(F))$ and $R = \|x - P_F x\|$. Then, for every $f \in F$,*

$$\|x - f\| - \|x - P_F x\| \geq 2\left(\|x - f\| + r\right)\left(\frac{r}{R + r}\right)\delta_X\left(\frac{\|f - P_F x\|}{\|x - f\| + r}\right).$$

Proof. Fix $0 < \delta < r$, x and f as above. Using the notation of Lemma 5.39 and of Lemma 5.41, let

$$\begin{aligned} w = x_0 &= \left(1 + \frac{r - \delta}{R}\right)x - \frac{r - \delta}{R}P_F x, \\ w_\theta &= (1 - \theta)w + \theta P_F x, \end{aligned}$$

where θ is chosen to ensure that $w_\theta = x$. Since $P_F w_\theta = P_F x$ and

$$\|w - w_\theta\| = \|w - P_F x\| - \|w_\theta - P_F x\| = r - \delta$$

then by Lemma 5.39, $P_F w = P_F x$. Thus, setting $\rho = \|w - P_F x\|$, it is evident that $B(w, \rho) \cap F = \{P_F x\}$. In particular, if $f \in F$ then $\|f - w\| \geq \rho$ and the assumptions of Lemma 5.41 are satisfied. A straightforward calculation shows that

$$\theta = (r - \delta)/(R + r - \delta), \quad \rho = R + r - \delta, \quad \text{and} \quad (1 - \theta)\rho = \|x - P_F x\|.$$

Thus, by Lemma 5.41,

$$\|x - f\| - \|x - P_F x\| \geq 2\|x_0 - f\|\frac{r - \delta}{R + r - \delta}\delta_X\left(\frac{\|f - P_F x\|}{\|x_0 - f\|}\right).$$

Finally, by the triangle inequality, $\|x_0 - f\| \leq \|x - f\| + (r - \delta) \equiv \Delta$, and recall that for $0 < \varepsilon_1 \leq \varepsilon_2 \leq 2$ one has $\varepsilon_1^{-1}\delta_X(\varepsilon_1) \leq \varepsilon_2^{-1}\delta_X(\varepsilon_2)$. Therefore, if $\|x_0 - f\| \leq \Delta$ and setting $\varepsilon_1 = \|f - P_F x\|/\Delta$ and $\varepsilon_2 = \|f - P_F x\|/\|x_0 - f\|$, one has that

$$\|x - f\| - \|x - P_F x\| \geq 2\Delta \left(\frac{r - \delta}{R + r}\right) \delta_X \left(\frac{\|f - P_F x\|}{\Delta}\right);$$

the claim follows by the monotonicity of δ_X and taking $\delta \to 0$. ∎

Proof of Theorem 5.34. Let $X = L_2$, recall that Y is the unknown target and that $Y \notin F \cup N(F)$. Setting $r = d(Y, N(F))$ and $R = \|Y - f^*\|_{L_2}$ it is evident from Corollary 5.42 and the bound of the modulus of convexity of a Hilbert space, that

$$\|Y - f\|_{L_2} - \|Y - f^*\|_{L_2} \geq 2(\|Y - f\|_{L_2} + r) \left(\frac{r}{R + r}\right) \cdot c \frac{\|f - f^*\|_{L_2}^2}{\|Y - f\|_{L_2}^2}.$$

Therefore, using that $a^2 - b^2 = (a - b)(a + b)$ and that $\|Y - f\|_{L_2} \geq \|Y - f^*\|_{L_2}$,

$$\|Y - f\|_{L_2}^2 - \|Y - f^*\|_{L_2}^2 \geq (\|Y - f\|_{L_2} - \|Y - f^*\|_{L_2}) \cdot \|Y - f\|_{L_2} \geq c \frac{r}{r + R} \|f - f^*\|_{L_2}^2,$$

as claimed. ∎

Chapter 6

When things concentrate

In this chapter we present two generic examples in which the triplets satisfy a convexity condition and the random variables involved are well behaved—in the sense that certain empirical means exhibit enough concentration around the true means and that concentration is uniform. Thanks to that strong concentration one can address the estimation and prediction problems using the most natural of procedures—*empirical risk minimization*.

6.1 Empirical risk minimization

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

$$\hat{f} = \operatorname{argmin}_{f \in F} \frac{1}{N} \sum_{i=1}^{N} \ell\left(f(X_i) - Y_i\right).$$

In the case of the squared loss, that candidate is a minimizer in F of the empirical risk

$$f \to \frac{1}{N} \sum_{i=1}^{N} \left(f(X_i) - Y_i\right)^2, \tag{6.1}$$

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

$$\hat{f} \in \operatorname{argmin}_{f \in F} \frac{1}{N} \sum_{i=1}^{N} \left(f(X_i) - Y_i\right)^2. \tag{6.2}$$

This procedure is called *Empirical Risk Minimization* (ERM). It is of central importance in statistical learning theory and was studied extensively over the last 50 years.

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

Definition 6.1. *Let ℓ be a loss function and set f^* to be the minimizer in F of the risk functional $f \to \mathbb{E}\ell(f(X) - Y)$. 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 not true for the excess loss $\mathcal{L}_f(X,Y)$: the identity of f^* 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 it. With this disclaimer out of the way, observe 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, if ERM produces $\hat{f}$ then

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

It follows that

$$\frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{\hat{f}}(X_i,Y_i) \leq 0 \tag{6.3}$$

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

6.2 The bounded framework

In this section we study learning problems in extremely favourable condition: the loss satisfies a Lipschitz condition; the class consists of uniformly bounded functions; and the geometry of the problem is well-behaved, in a sense that the triplet satisfies a suitable version of the convexity condition. To put this in perspective, the classical statistical prediction problem of linear regression in $\mathbb{R}^d$ with respect to the squared loss and additive gaussian noise does not fall within the scope of the bounded framework. Indeed:

Example 6.2. *Let $T \subset \mathbb{R}^d$ be bounded and set $F_T = \{\langle t, \cdot \rangle : t \in T\}$. Let X be the standard gaussian vector in $\mathbb{R}^d$, set $Y = \langle z_0, X \rangle + W$ where $z_0 \in \mathbb{R}^d$ and W is a standard gaussian variable; and put $\ell(t) = t^2$—the squared loss. For the triplet (F,X,Y) all the conditions of the bounded framework are false: linear functionals are not bounded, the loss in not a Lipschitz function, and, without more information on the location of z_0 relative to T, the convexity condition need not be true.*

At the same time, there are countless interesting problems that do fall within the scope of the bounded framework (see, e.g., the books [**?**] and references therein as a starting point).

Formally,

Assumption 6.1. *The classical approach is based on the following assumptions:*

1. *The class F consists of functions that are bounded by almost surely by b and so is the target Y.*
2. *The loss ℓ is a Lipschitz function with constant L in $[-2b, 2b]$.*
3. *The excess loss $\mathcal{L}$ satisfies a convexity condition: there is a constant $B \geq 1$ such that for every $f \in F$,*
$$\|f - f^*\|_{L_2}^2 \leq B\mathbb{E}\mathcal{L}_f.$$

Out of the three assumptions, it is straightforward to relax (1), by assuming that the class F has a well behaved envelope function $H(x) = \sup_{f\in F} |f(x)|$ that belongs to L_p or to L_{ψ_α}. Having said that, an assumption on the envelope does not really go beyond the bounded case. An envelope condition restricts the 'peaky' part of all the functions in the class to a fixed area on Ω (exactly where the envelope is large), and so those peaky parts may be controlled uniformly by studying a single function. Thus, by applying a simple truncation argument, one reverts to the bounded case.

As for assumption (3), one may show that it holds if $F \subset L_2$ is a convex set and the loss ℓ is strongly convex. For simplicity, if ℓ is also smooth, then strong convexity means that for any $u, v \in \mathbb{R}$
$$\ell(u) \geq \ell(v) + \ell'(v) \cdot (u - v) + \frac{C}{2}|u - v|^2,$$

Exercise 18. (1) *Show that if ℓ is smooth and strongly convex, $F \subset L_2(\mu)$ is closed and convex and $Y \in L_2$, then the triplet (F, X, Y) satisfies the convexity condition* (3)*, with B that depends only on the strong convexity constant of ℓ.*

(2) *Show that the same holds without the smoothness assumption on ℓ.*

Combining (1) and (2) it follows that for every $f \in F$ and every (X, Y),
$$|\mathcal{L}_f(X, Y)| = |\ell(f(X) - Y) - \ell(f^*(X) - Y)| \leq L|f(X) - f^*(X)| \leq 2Lb, \tag{6.4}$$
implying that $\{\mathcal{L}_f : f \in F\}$ is a class of uniformly bounded functions.

Also, combining (2) and (3) one has for every $f \in F$,
$$\mathbb{E}\mathcal{L}_f^2 = \mathbb{E}\left(\ell(f(X) - Y) - \ell(f^*(X) - Y)\right)^2 \leq L^2\mathbb{E}|f - f^*|^2(X) \leq BL^2\mathbb{E}\mathcal{L}_f, \tag{6.5}$$
which is the so-called Bernstein condition (and the key part of (6.5) is, as one could expect, the convexity condition).

To formulate the accuracy/confidence tradeoff and the sample complexity estimate for prediction and estimation under Assumption 6.1, recall that $F_{f^*,r} = \mathrm{star}(F - f^*, 0) \cap rD$ and set
$$\phi_N(r, \sigma) = \frac{1}{\sqrt{N}} \sup_{u\in F_{f^*,r}} \left|\sum_{i=1}^N \varepsilon_i u(X_i)\right|, \tag{6.6}$$
and
$$\bar{k}_N(\gamma) = \inf\left\{r > 0 : \mathbb{E}\phi_N(r, \sigma) \leq \gamma r^2\sqrt{N}\right\}, \tag{6.7}$$
where the expectation is taken with respect to both $(\varepsilon_i)_{i=1}^N$ and $(X_i)_{i=1}^N$.

Theorem 6.3. *Under Assumption 6.1 there exists a constant c that depends only on B and a constant c_1 that depends only on L, b and B for which the following holds. Let $\gamma = c/L$ and set $r = 2\bar{k}_N(\gamma)$. Then, with probability at least $1 - \exp(-c_1 N r^2)$,*

$$\sup_{\{f \in F : \mathbb{E}\mathcal{L}_f \geq r\}} \left| \frac{\frac{1}{N}\sum_{i=1}^N \mathcal{L}_f(X_i, Y_i)}{\mathbb{E}\mathcal{L}_f} - 1 \right| \leq \frac{1}{2}.$$

In particular, on that event, the excess risk of the empirical minimizer satisfies

$$\mathbb{E}\mathcal{L}_{\hat{f}} \leq 4\bar{k}_N^2(\gamma),$$

and

$$\|\hat{f} - f^*\|_{L_2} \leq 2\sqrt{B}\bar{k}_N(\gamma).$$

An alternative formulation of Theorem 6.3 is in terms of the sample complexity—and its proof, once Theorem 6.3 is established, is straightforward.

Theorem 6.4. *Under the same conditions as in Theorem 6.3, let ε and δ be the wanted accuracy and confidence levels. Set*

$$N_0 = \min\{N : \mathbb{E}\phi_N(\sqrt{\varepsilon}, \sigma) \leq \gamma\varepsilon\sqrt{N}\} + c\frac{\varepsilon \log(2/\delta)}{N},$$

and let $N \geq N_0$. Then, ERM satisfies that, with probability at least $1 - \delta$ with respect to $(X_i, Y_i)_{i=1}^N$,

$$\|\hat{f} - f^*\|_{L_2} \leq \sqrt{\varepsilon} \quad \text{and} \quad \mathbb{E}\left(\left(\hat{f}(X) - Y \right)^2 | (X_i, Y_i)_{i=1}^N \right) \leq \mathbb{E}(f^*(X) - Y)^2 + \varepsilon.$$

Observe that the fixed point $\bar{k}_N(\gamma)$ (and therefore, the error term in Theorem 6.3) does not improve when the noise level of the problem, $\|f^*(X) - Y\|_{L_2}$, decreases.

Before turning to the proof of Theorem 6.3, let us compare the fixed point $r = \bar{k}_N(\gamma)$ to $r_{\mathbb{Q}}$ and $r_{\mathbb{M}}$. First of all, note that $r \geq r_{\mathbb{M}}$, because r is a legal value in the fixed-point condition (up to the right choice of constant used to define $r_{\mathbb{M}}$). Indeed, by the contraction inequality for Bernoulli processes and recalling that $\|\xi\|_{L_\infty} = \|f^*(X) - Y\|_{L_\infty} \leq 2b$

$$\mathbb{E}_X \mathbb{E}_\varepsilon \sup_{u \in F_{f^*, r}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i \xi_i u(X_i) \right| \leq 2b \mathbb{E}_X \mathbb{E}_\varepsilon \sup_{u \in F_{f^*, r}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i u(X_i) \right| \leq 2b\gamma r^2 \sqrt{N}$$

where the last inequality holds because r satisfies (6.7).

As for $r_{\mathbb{Q}}$, observe first that $r \lesssim \sqrt{b/\gamma}$. Indeed, $P_\sigma F \subset bB_\infty^N$, and thus, for every ρ, $\mathbb{E}_\varepsilon \phi_N(\rho, \sigma) \leq \sqrt{N} b$. Hence,

$$\mathbb{E} \sup_{u \in F_{f^*, r}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i u(X_i) \right| \leq \gamma\sqrt{N} r^2 \leq \sqrt{b}\sqrt{N} r,$$

and up to the dependence on γ and b, $r = \bar{k}_N(\gamma)$ is larger than $\max\{r_{\mathbb{Q}}, r_{\mathbb{M}}\}$.

The first observation required for the proof of Theorem 6.3 has to do with the nature of the fixed point $\bar{k}_N(\gamma)$. A similar phenomenon holds for $r_{\mathbb{Q}}$ and $r_{\mathbb{M}}$, and all of them are based on the fact that the indexing set consists of localizations of a set that is star-shaped around 0—$\mathrm{star}(F - f^*, 0)$.

Lemma 6.5. *If $r > \bar{k}_N(\gamma)$ then $\mathbb{E}\phi_N(r,\sigma) \le \gamma r^2\sqrt{N}$, and if $r < \bar{k}_N(\gamma)$, the reverse inequality holds.*

Proof. Fix $\rho_1 > 0$ for which

$$\mathbb{E}\phi_N(\rho_1,\sigma) = \mathbb{E}\sup_{u\in F_{f^*,\rho_1}}\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N \varepsilon_i u(X_i)\right| \le \gamma\rho_1^2\sqrt{N},$$

and note that if $\rho_2 > \rho_1$ and $u \in F_{f^*,\rho_2}$ then by the star-shape property, $(\rho_1/\rho_2)h \in F_{f^*,\rho_1}$. Given $(\varepsilon_i)_{i=1}^N$ and $(X_i)_{i=1}^N$, assume that $\sup_{h\in F_{f^*,\rho_2}}\left|\sum_{i=1}^N \varepsilon_i u(X_i)\right|$ is attained in h and that $\rho_1 \le \|h\|_{L_2} \le \rho_2$. Therefore,

$$\sup_{u\in F_{f^*,\rho_2}}\left|\sum_{i=1}^N \varepsilon_i u(X_i)\right| = \frac{\|h\|_{L_2}}{\rho_1}\left|\sum_{i=1}^N \varepsilon_i \frac{\rho_1}{\|h\|_{L_2}}h(X_i)\right| \le \frac{\rho_2}{\rho_1}\sup_{u\in F_{f^*,\rho_1}}\left|\sum_{i=1}^N \varepsilon_i u(X_i)\right|.$$

Taking expectations on both sides,

$$\mathbb{E}\sup_{u\in F_{f^*,\rho_2}}\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N \varepsilon_i u(X_i)\right| \le \frac{\rho_2}{\rho_1}\mathbb{E}\sup_{u\in F_{f^*,\rho_1}}\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N \varepsilon_i u(X_i)\right| \le \gamma\rho_2\rho_1\sqrt{N} \le \gamma\rho_2^2\sqrt{N}.$$

The proof of the second part follows an identical path and is omitted. ∎

The proof of Theorem 6.3 relies heavily on Talagrand's concentration inequality for bounded empirical processes, a version of which is formulated below (see also [**?**]). The impact that Talagrand's inequality has on statistical learning theory cannot be overstated. Almost every major contribution to the area is based, to some extent, on that inequality, and the fact that no similar inequality holds for empirical processes involving more heavy-tailed random variables caused a major delay in the study of prediction and estimation problems.

Talagrand's inequality will be used many times throughout these notes. The reader is strongly encouraged to study its proof—as well as the entire beautiful machinery developed by Talagrand for the analysis of the concentration of measure phenomenon in product spaces.

Theorem 6.6. *There exist an absolute constant C for which the following holds. Let H be a class of functions and set $\sigma_H^2 = \sup_{h\in H}\operatorname{var}(h)$ and $b = \sup_{h\in H}\|h\|_{L_\infty}$. For every $x > 0$, with probability at least $1 - 2\exp(-x)$,*

$$\sup_{h\in H}\left|\frac{1}{N}\sum_{i=1}^N h(X_i) - \mathbb{E}h\right| \le C\left(\mathbb{E}\sup_{h\in H}\left|\frac{1}{N}\sum_{i=1}^N \varepsilon_i h(X_i)\right| + \sigma_H\sqrt{\frac{x}{N}} + b\frac{x}{N}\right).$$

The second preliminary result the proof of Theorem 6.3 requires is based on is the following version of the contraction theorem for Bernoulli processes [**?**]:

Theorem 6.7. *Let $T \subset \mathbb{R}^N$ and let $\phi_i : \mathbb{R} \to \mathbb{R}$ satisfy that $\phi_i(0) = 0$ and that $\|\phi_i\|_{\text{lip}} \le L$. Then*

$$\mathbb{E}\sup_{t\in T}\left|\sum_{i=1}^N \varepsilon_i\phi_i(t_i)\right| \le 2L\mathbb{E}\sup_{t\in T}\left|\sum_{i=1}^N \varepsilon_i t_i\right|.$$

Proof of Theorem 6.3. The classes we are interested in are level sets of F, scaled according to their excess risk: let $r = 2\bar{k}_N(\gamma)$ for $\gamma = c/L$, where the absolute constant c is to be specified in what follows. Set

$$F_j = \{f \in F : 2^{j-1}r^2 \leq \mathbb{E}\mathcal{L}_f < 2^j r^2\}, \quad \text{and} \quad F_0 = \{f \in F : \mathbb{E}\mathcal{L}_f \leq r^2\},$$

and consider the classes $\mathcal{L}_{F_j} = \{\mathcal{L}_f : f \in F_j\}$. The key to the proof is an estimate on the probability with which

$$\sup_{f \in F_j} \left| \frac{\frac{1}{N}\sup_{i=1}^N \mathcal{L}_f(X_i, Y_i)}{\mathbb{E}\mathcal{L}_f} - 1 \right| \leq \frac{1}{2}.$$

To that end, because every $f \in F_j$ satisfies that $2^{j-1}r^2 \leq \mathbb{E}\mathcal{L}_f \leq 2^j r^2$, it suffices to estimate the probability with which

$$\sup_{f \in F_j} \left| \frac{1}{N}\sum_{i=1}^N \mathcal{L}_f(X_i, Y_i) - \mathbb{E}\mathcal{L}_f \right| \leq \frac{1}{2} \cdot 2^{j-1}r^2. \tag{6.8}$$

One may invoke Talagrand's concentration inequality to each one of the class $\mathcal{L}_{F_j}$. Note that by (6.4)

$$\sup_{f \in F_j} \|\mathcal{L}_f\|_\infty \leq 2Lb,$$

and by (6.5), for any $f \in F_j$,

$$\mathbb{E}\mathcal{L}_f^2 \leq BL^2\mathbb{E}\mathcal{L}_f \leq BL^2 \cdot (2^j r^2);$$

thus,

$$\sigma_{\mathcal{L}_{F_j}} \leq \sqrt{B}L \cdot 2^{j/2}r \quad \text{and} \quad \{f - f^* : f \in F_j\} \subset F_{f^*,\rho}$$

for $\rho = \sqrt{B}2^{j/2}r$, because $\|f - f^*\|_{L_2}^2 \leq B\mathbb{E}\mathcal{L}_f$.

Finally, one has to estimate

$$\mathbb{E}\sup_{f \in F_j} \left| \frac{1}{N}\sum_{i=1}^N \varepsilon_i \mathcal{L}_f(X_i, Y_i) \right| = \mathbb{E}_{X,Y}\left(\mathbb{E}_\varepsilon \sup_{f \in F_j} \left| \frac{1}{N}\sum_{i=1}^N \varepsilon_i \mathcal{L}_f(X_i, Y_i) \right| \right),$$

and to do so we invoke a contraction argument. Fix $(X_i, Y_i)_{i=1}^N$, set $\xi_i = f^*(X_i) - Y_i$ and put

$$\phi_i(z) = \ell(z - \xi_i) - \ell(\xi_i).$$

Clearly, each ϕ_i is a Lipschitz function with constant L and satisfies that $\phi_i(0) = 0$. Also,

$$\mathcal{L}_f(X_i, Y_i) = \ell\left((f - f^*)(X_i) + \xi_i\right) - \ell\left(\xi_i\right) = \phi_i\left((f - f^*)(X_i)\right).$$

By the contraction inequality for Bernoulli processes (Theorem 6.7), for every fixed $(X_i, Y_i)_{i=1}^N$,

$$\begin{aligned}\mathbb{E}_\varepsilon \sup_{f \in F_j} \left| \frac{1}{N}\sum_{i=1}^N \varepsilon_i \mathcal{L}_f(X_i, Y_i) \right| =& \mathbb{E}_\varepsilon \sup_{f \in F_j} \left| \frac{1}{N}\sum_{i=1}^N \varepsilon_i \left(\phi_i(f - f^*)(X_i)\right) \right| \\ \leq& 2L\mathbb{E}_\varepsilon \sup_{f \in F_j} \left| \frac{1}{N}\sum_{i=1}^N \varepsilon_i (f - f^*)(X_i) \right|. \end{aligned} \tag{6.9}$$

Therefore, recalling that $\{f - f^* : f \in F_j\} \subset F_{f^*,\rho}$ for $\rho = \sqrt{B}2^{j/2}r$,

$$\begin{aligned}\mathbb{E}\sup_{f\in F_j}\left|\frac{1}{N}\sum_{i=1}^N \varepsilon_i \mathcal{L}_f(X_i,Y_i)\right| \leq 2L\mathbb{E}\sup_{f\in F_j}\left|\frac{1}{N}\sum_{i=1}^N \varepsilon_i (f-f^*)(X_i)\right| \\ \leq 2L\mathbb{E}\sup_{u\in F_{f^*,\rho}}\left|\frac{1}{N}\sum_{i=1}^N \varepsilon_i u(X_i)\right| \leq 2L\cdot\gamma(\sqrt{B}2^{j/2}r)^2,\end{aligned}$$

because $\sqrt{B}2^{j/2}r > r$ and applying Lemma 6.5.

Combining these observations, it follows that with probability at least $1-\exp(-x)$,

$$\begin{aligned}\sup_{f\in F_j}\left|\frac{1}{N}\sum_{i=1}^N \mathcal{L}_f(X_i,Y_i) - \mathbb{E}\mathcal{L}_f\right| \leq & C\left(2L\cdot\gamma B 2^j r^2 + \sqrt{\frac{x}{N}}\sqrt{B}L\cdot 2^{j/2}r + 2Lb\frac{x}{N}\right) \\ \leq & \frac{1}{2}\cdot 2^{j-1}r^2,\end{aligned}$$

provided that

$$\gamma \leq \frac{c_1}{LB} \tag{6.10}$$

and setting

$$x = c_2(L,b,B)N2^j r^2; \tag{6.11}$$

(one may take $c_1 = 1/24C$ and $c_2 = \min\{1/(144BL^2C^2), 1/(24CLb)\}$—which clearly are not the optimal choices).

Hence, with probability at least $1 - 2\exp(-c_2(L,b,B)N2^j r^2)$, for any $f \in F_j$,

$$\frac{1}{2}\cdot 2^{j-1}r^2 \leq \frac{1}{2}\mathbb{E}\mathcal{L}_f \leq \frac{1}{N}\sum_{i=1}^N \mathcal{L}_f(X_i,Y_i) \leq \frac{3}{2}\mathbb{E}\mathcal{L}_f,$$

and on top of having the required ratio estimate, it is evident that on the same event, $\hat{f} \notin F_j$. Applying the union bound for $j \geq 1$, one has that with probability at least

$$1 - \sum_{j\geq 1}\exp(-c_2(L,b,B)N2^j r^2) \geq 1 - \exp(-c_3(L,b,B)Nr^2),$$

the ratio estimate holds for any $f \in F\backslash F_0$ and the empirical minimizer satisfies

$$\hat{f} \notin \bigcup_{j\geq 1} F_j = F\backslash F_0.$$

Thus, $\hat{f} \in F_0$, and as a result, the risk of the empirical minimizer satisfies

$$\mathbb{E}\mathcal{L}_{\hat{f}} \leq r^2 = 4\bar{k}_N^2(\gamma).$$

The final part of the proof is immediate, recalling that for every $f \in F$, $\|f - f^*\|_{L_2}^2 \leq B\mathbb{E}\mathcal{L}_f$. ■

Remark 6.8. *Note that the probability estimate is not trivial only when $r^2 \gtrsim 1/N$. Thus, this method of proof cannot give an error bound that is better than $\sim 1/N$, and specifically, cannot be used to ensure exact recovery even in noise-free problems.*

The crucial point to notice here is the damaging effect of the application of a contraction argument. The outcome of Theorem 6.3 depends on the Lipschitz constant of ℓ rather than on the 'magnitude' of the noise (captured, for example, by $\|Y - f^*(X)\|_{L_2}$ rather than by a function of $\|Y - f^*(X)\|_{L_\infty}$). The estimate does not improve even when the problem becomes 'more realizable'—that is, when Y is closer to F, and therefore it is, at times loose.

It turns out that ERM performs with the best possible accuracy and the optimal confidence when F is a subgaussian class (in which case $\gamma = c/L$ can be replaced by the correct quantity $\gamma \sim (\|Y - f^*(X)\|_{L_2})^{-1}$). Unfortunately, in more heavy tailed problems ERM is no longer a reasonable procedure and one has to come up with better alternatives.

Exercise 19. *Give an example of estimation/prediction problems involving a convex class in which ERM does no perform with the sample complexity of, say,* (5.24)*. (Hint: it is possible to find a one-dimensional example).*

6.3 Subgaussian learning

Let us turn to another example in which concentration of empirical means is still strong enough. As is the case throughout these notes, the focus is on a convex class F (and as a result, $F_{f^*,r} = (F - f^*) \cap rD$).

The assumption that leads to sufficient concentration is that the class is L-subgaussian and that $\xi = f^*(X) - Y$ has a finite ψ_2 norm, where the meaning of the ψ_2 and L_2 norm equivalence is that

$$\|f - h\|_{\psi_2} \leq L\|f - h\|_{L_2} \text{ for every } f, h \in F \cup \{0\}.$$

To formulate the result, given $H \subset L_2$ let $\{G_h : h \in H\}$ be the canonical gaussian process indexed by H; that is, each G_h is a centred gaussian random variable and for any $f, h \in H$, $\mathbb{E}G_hG_f = \langle f, h\rangle = \int f(x)h(x)d\mu(x)$. For an extensive survey no gaussian processes we refer the reader to [**?**].

Let

$$\mathbb{E}\sup_{h\in h} G_h = \sup\{\mathbb{E}\sup_{h\in H'} G_h,\ H' \subset H \text{ is finite}\}.$$

Theorem 6.9. *There are constants c_0, c_1, c_2 that depend only on L for which the following holds. Let r satisfy that*

$$\mathbb{E}\sup_{h\in F_{f^*,r}} G_h \leq c_0\sqrt{N} \cdot \min\left\{r, \frac{r^2}{\|\xi\|_{\psi_2}}\right\}.$$

Then with probability at least

$$1 - 2\exp\left(-c_1 N \min\left\{\frac{r^2}{\|\xi\|_{\psi_2}^2}, 1\right\}\right),$$

ERM performed in F using a sample $(X_i, Y_i)_{i=1}^N$, returns $\hat{f}$ that satisfies

$$\|\hat{f} - f^*\|_{L_2} \leq c_2 r \quad \text{and} \quad \mathbb{E}\left(\left(\hat{f}(X) - Y\right)^2 \big| (X_i, Y_i)_{i=1}^N\right) \leq c_2 r^2.$$

The rest of this section is devoted to the proof of Theorem 6.9.

Remark 6.10. *Note that the estimate in Theorem 6.9 relies heavily on $\|\xi\|_{\psi_2}$ rather than on $\|\xi\|_{L_2}$, the latter being what the optimal estimate should be based on. As a result, if $\|\xi\|_{\psi_2}$ is significantly larger than $\|\xi\|_{L_2}$ the estimate from Theorem 6.9 is suboptimal—though for the optimal performance one would have to use a different procedure than ERM.*

Equation (6.3) presents an opportunity to analyze ERM in more detail. For every fixed sample, 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

$$\{f \in F : \|f - f^*\|_{L_2} \geq r\} \subset \left\{ f \in F : \sum_{i=1}^N \mathcal{L}_f(X_i, Y_i) > 0 \right\}, \tag{6.12}$$

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

$$\{f \in F : R(f) - R(f^*) \geq r^2\} \subset \left\{ f \in F : \sum_{i=1}^N \mathcal{L}_f(X_i, Y_i) > 0 \right\} \tag{6.13}$$

then $R(\hat{f}) - R(f^*) \leq r^2$, leading to an estimate on $\mathcal{E}_p$.

Let us examine (6.12) for the squared loss. To obtain a useful bound one has to show that with high probability, for any function that is sufficiently far away from (the unknown) f^*, the excess empirical risk is positive.

Recall the decomposition of the excess loss functional to its quadratic and multiplier components:

$$(f(X) - Y)^2 - (f^*(X) - Y)^2 = (f(X) - f^*(X))^2 + 2\,(f(X) - f^*(X))\,(f^*(X) - Y)\,,$$

and denote by P_N the empirical mean functional, i.e.,

$$\begin{aligned} P_N \mathcal{L}_f =& P_N\,(f(X) - f^*(X))^2 + 2P_N\,(f(X) - f^*(X))\,(f^*(X) - Y) = \\ =& \frac{1}{N}\sum_{i=1}^N (f(X_i) - f^*(X_i))^2 + \frac{2}{N}\sum_{i=1}^N (f(X_i) - f^*(X_i))\,(f^*(X_i) - Y)\,. \end{aligned}$$

Note that $P_N \mathcal{L}_f$ has some homogeneity in $f - f^*$: suppose that $P_N \mathcal{L}_f > 0$, and consider $u \in F$ that 'lives' on the ray originating from f^* 'beyond' f; thus $u - f^* = \theta(f - f^*)$ for some $\theta > 1$ and

$$P_N \mathcal{L}_u = \theta^2 P_N (f - f^*) + 2\theta P_N\,(f(X) - f^*(X))\,(f^*(X) - Y) \geq \theta P_N \mathcal{L}_f > 0.$$

This, and the fact that F is star-shaped around f^* (a convex set is star-shaped around any of its points) implies the following:

Let F be a convex class. To prove that $\mathbb{E}\mathcal{L}_f > 0$ for any $f \in F$ which satisfies that $\|f - f^*\|_{L_2} \geq r$, it suffice to show that $\mathbb{E}\mathcal{L}_f > 0$ for any $f \in F$ that satisfies $\|f - f^*\|_{L_2} = r$.

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

$$\mathcal{Q}_{f-f^*} = \frac{1}{N}\sum_{i=1}^{N}(f-f^*)^2(X_i), \tag{6.14}$$

and

$$\mathcal{M}_{f-f^*} = \frac{2}{N}\sum_{i=1}^{N}\xi_i(f-f^*)(X_i), \tag{6.15}$$

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

$$P_N\mathcal{L}_f = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_f(X_i,Y_i) = \mathcal{Q}_{f-f^*} + \mathcal{M}_{f-f^*}. \tag{6.16}$$

For obvious reasons we call $\mathcal{Q}_{f-f^*}$ the *quadratic component* of the empirical excess risk functional and $\mathcal{M}_{f-f^*}$ is the multiplier component.

Let us first find a high probability upper bound on $\mathcal{E}_e$, by showing that if $\|f-f^*\|_{L_2} \geq r$ then $P_N\mathcal{L}_f > 0$, implying that ERM selects $\hat{f}$ whose distance from f^* is at most r. Once again, by the star-shape property, it suffices to show that with high probability,

$$P_N\mathcal{L}_f > 0 \quad \text{when } \|f-f^*\|_{L_2} = r.$$

Clearly, $\mathbb{E}\mathcal{L}_f = \mathbb{E}\mathcal{Q}_{f-f^*} + \mathbb{E}\mathcal{M}_{f-f^*}$, and therefore, by the convexity condition,

$$\begin{aligned} P_N\mathcal{L}_f &= (\mathcal{Q}_{f-f^*} - \mathbb{E}\mathcal{Q}_{f-f^*}) + (\mathcal{M}_{f-f^*} - \mathbb{E}\mathcal{M}_{f-f^*}) + \mathbb{E}\mathcal{L}_f \\ &\geq -|\mathcal{Q}_{f-f^*} - \mathbb{E}\mathcal{Q}_{f-f^*}| - |\mathcal{M}_{f-f^*} - \mathbb{E}\mathcal{M}_{f-f^*}| + \frac{1}{2}\left(\mathbb{E}\mathcal{L}_f + \frac{1}{B}\|f-f^*\|_{L_2}^2\right). \end{aligned} \tag{6.17}$$

The question becomes estimating the oscillations

$$|\mathcal{Q}_{f-f^*} - \mathbb{E}\mathcal{Q}_{f-f^*}| \quad \text{and} \quad |\mathcal{M}_{f-f^*} - \mathbb{E}\mathcal{M}_{f-f^*}|$$

on the star-shaped set $(F-f^*)\cap rD$, and showing that if $\|f-f^*\|_{L_2} = r$ these oscillations are dominated by $\mathbb{E}\mathcal{L}_f + r^2/B$.

Obtaining a useful bound of these oscillations is a nontrivial task. This has been studied extensively over the years under various conditions on the class F (see, for example, [**?**]). The following estimate is from [**?**], and to formulate it one requires the next definition.

Definition 6.11. *For an integer s,*

$$\Lambda_s(H) = \mathbb{E}\sup_{h\in H} G_h + 2^{s/2}\mathrm{diam}(H, L_2).$$

Theorem 6.12. *For every $L > 1$ there exist constants c_1 and c_2 that depend only on L and for which the following holds. Let H be an L-subgaussian class of functions. If $\xi \in L_{\psi_2}$ then for every $u, w \geq 8$ and $s \geq 0$, with probability at least $1 - 2\exp(-c_1u^2 2^s) - 2\exp(-c_1 N\min\{w^2, w^4\})$,*

$$\sup_{h\in H}\left|\sum_{i=1}^{N}(\xi_i h(X_i) - \mathbb{E}\xi h)\right| \leq c_2 uw\sqrt{N}\|\xi\|_{\psi_2}\Lambda_s(H). \tag{6.18}$$

Moreover, with probability at least $1 - 2\exp(-c_1 u^2 2^s)$,

$$\sup_{h\in H}\left|\sum_{i=1}^N (h^2(X_i) - \mathbb{E}h^2)\right| \le c_2\left(u^2\Lambda_s^2(H) + u\sqrt{N}\sup_{h\in H}\|h\|_{L_2}\Lambda_s(H)\right). \tag{6.19}$$

In the case that is of interest here, $H = F_{f^*,r} = (F - f^*)\cap rD$ and $\xi = (f^*(X) - Y)$. To ensure that both oscillation terms are bounded by at most $(\theta/2)r^2$ for $0 < \theta < 1$ it suffices that $u, w \sim 1$ and

$$\Lambda_s(H) \sim \theta\sqrt{N}\cdot\min\left\{\frac{r^2}{\|\xi\|_{\psi_2}}, r\right\} = (*).$$

Hence, if r is the smallest such that

$$\mathbb{E}\sup_{h\in F_{f^*,r}} G_h \le c\frac{\theta}{\sqrt{N}}\cdot\min\left\{\frac{r^2}{\|\xi\|_{\psi_2}}, r\right\}$$

for a suitable constant c, and $2^s = ((*)/r)^2$, one has that on an event $\mathcal{A}$ of probability at least

$$1 - 2\exp\left(-c_1(L)\theta^2 N\min\left\{\frac{r^2}{\|\psi_2\|_{L_2}^2}, 1\right\}\right),$$

$$\sup_{\{f\in F,\ \|f-f^*\|_{L_2}\le r\}} |\mathcal{Q}_{f-f^*} - \mathbb{E}\mathcal{Q}_{f-f^*}| \le \frac{\theta}{2}r^2,$$

and

$$\sup_{\{f\in F,\ \|f-f^*\|_{L_2}\le r\}} |\mathcal{M}_{f-f^*} - \mathbb{E}\mathcal{M}_{f-f^*}| \le \frac{\theta}{2}r^2.$$

Recalling (6.17), on the event $\mathcal{A}$, one has that for every $f \in F$ that satisfies $\|f - f^*\|_{L_2} = r$,

$$P_N\mathcal{L}_f \ge -\theta r^2 + \frac{1}{2}\left(\mathbb{E}\mathcal{L}_f + \frac{1}{B}\|f - f^*\|_{L_2}^2\right) \ge r^2\left(-\theta + \frac{1}{2B}\right)$$

where the last inequality follows because $\mathbb{E}\mathcal{L}_f \ge 0$. Thus, if $\theta < 1/2B$ then $P_N\mathcal{L}_f$ is positive for any $f \in F$ that satisfies $\|f - f^*\|_{L_2} = r$. By the star-shape property, the same is true when $\|f - f^*\|_{L_2} \ge r$, showing that $\mathcal{E}_e \le r$ with the wanted probability estimate.

To establish that $\mathcal{E}_p \lesssim r^2$ on the same event, assume that it is not—and the conditional expectation $\mathbb{E}\mathcal{L}_{\hat{f}} \ge Cr^2$. But since $\|\hat{f} - f^*\|_{L_2} \le r$, that means that there is some $f \in F$ for which $\|f - f^*\|_{L_2} \le r$, $\mathbb{E}\mathcal{L}_f \ge Cr^2$ and $P_N\mathcal{L}_f \le 0$. That is impossible on the event A by (6.17): for any $f \in F$ such that $\|f - f^*\|_{L_2} \le r$,

$$P_N\mathcal{L}_f \ge -\theta r^2 + \frac{1}{2}\left(\mathbb{E}\mathcal{L}_f + \frac{1}{B}\|f - f^*\|_{L_2}^2\right) \ge -\theta r^2 + \frac{1}{4}\mathbb{E}\mathcal{L}_f > 0$$

if $C \ge \max\{4\theta, 2/B\}$. ∎

Remark 6.13. *It is instructive to see how the choice of r is connected with $r_{\mathbb{Q}}$ and $r_{\mathbb{M}}$. For the former, note that if H is an L-subgaussian class of functions then by Talagrand's majorizing measures theorem,*

$$\mathbb{E}\sup_{h\in H}\left|\frac{1}{\sqrt{N}}\sum_{i=1}^N \varepsilon_i h(X_i)\right| \le cL\mathbb{E}\sup_{h\in H} G_h,$$

which means that the condition $\mathbb{E}\sup_{h\in F_{f^*,r}} G_h \le c(L)\sqrt{N}r$ *implies that* $r_{\mathbb{Q}} \le r$ *(of course, assuming that the constants in the definitions are chosen properly).*

Also, an immediate application of the first part of Theorem 6.12 (which is also based on the majorizing measures theorem) shows that

$$\mathbb{E}\sup_{h\in H}\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\varepsilon_i\xi_i h(X_i)\right| \le cL^2\|\xi\|_{\psi_2}\mathbb{E}\sup_{h\in H} G_h,$$

which, by the same argument, implies that $r_{\mathbb{M}} \le r$.

Therefore, the estimate in Theorem 6.9 is weaker than the main result of these notes—even at the level of fixed points. Also, as noted previously, the estimate depend on $\|\xi\|_{\psi_2}$ *rather than on* $\|\xi\|$ *for a smaller norm (ideally—*$\|\xi\|_{L_2}$*).*

Part III

Heavy tailed problems

Chapter 7

The small-ball method

The fact that the sum of iid random variables concentrates around the mean is an important and useful fact. However, it is not without its problems. At least the bounds that were presented in the first part of these notes require the given random variable to be 'well-behaved': for example, it should have a finite L_∞ or ψ_α norm, and even in such cases, the quality of the estimate often depends on norm equivalence rather than of an upper bound with respect to some norm. But is concentration of the empirical mean around the true one really a single phenomenon? Concentration is a two-sided inequality, and unlikely as it sounds at first, it is possible that the two inequalities are of a completely different nature.

Consider the following case. Let Z be a mean-zero variance 1 random variable, and set $Z_1, ..., Z_N$ to be independent copies of Z. For some fixed constants c and C, consider the two events

$$\mathcal{A} = \left\{ \frac{1}{N} \sum_{i=1}^N Z_i^2 \geq c\mathbb{E}Z^2 \right\} \text{ and } \mathcal{B} = \left\{ \frac{1}{N} \sum_{i=1}^N Z_i^2 \leq C\mathbb{E}Z^2 \right\}.$$

A two-sided estimate of $N^{-1}\sum_{i=1}^N Z_i^2$ holds on $\mathcal{A} \cap \mathcal{B}$, and therefore, the worse estimate of the two will determine the way the empirical means deviate from $\mathbb{E}Z^2$.

There is a clear difference between the two event: a single very large Z_j^2 might result in $(Z_i)_{i=1}^N \notin \mathcal{B}$. On the other hand, even if a proportional number of the Z_i are 0, it is still possible that $(Z_i)_{i=1}^N \in \mathcal{A}$. This lack of symmetry is very significant. it hints that the lower bound may be true with a much higher probability estimate than the upper one.

Example 7.1. *Define a random variable Z by*

$$Pr(Z = N) = \frac{1}{N^2}, \quad \text{and} \quad Pr(Z = 1) = Pr(Z = 0) = \frac{1}{2}\left(1 - \frac{1}{N^2}\right).$$

Therefore, $\mathbb{E}Z^2 = 3/2 - 1/2N^2$. Let $Z_1, ..., Z_N$ be independent copies of Z, and observe that

$$Pr(\exists\ 1 \leq i \leq N\ :\ Z_i = N) = 1 - \left(1 - \frac{1}{N^2}\right)^N \geq \frac{1}{2N},$$

and that on this event,

$$\frac{1}{N}\sum_{i=1}^N Z_i^2 \geq N.$$

Hence, even if C is ridiculously large, the best upper estimate one can have in this case holds only with probability $1-1/(2N)$. On the other hand, with probability at least $1-2\exp(-c_1 N)$,

$$\frac{1}{N}\sum_{i=1}^{N} Z_i^2 \geq \frac{1}{8}\mathbb{E}Z^2.$$

It follows from Example 7.1 that there are situations in which what is a weak upper estimate holds only with probability $(1-\mathrm{poly}(1/N))$, but a much stronger lower estimate is true with very high probability.

The aim of this chapter is to study this phenomenon and see when high-probability 'isomorphic' or 'almost isometric' lower bounds are true. The main message is that such lower bounds are almost universally true, and are certainly are less restrictive than the upper estimates. Moreover, under minimal conditions the lower bounds are *stable*: they remain true even if someone maliciously removes a large part of the sample.

The importance of the universality of the lower bound and its stability is extreme: both facts play a crucial role in what follows, especially in the study of heavy tailed learning problems.

In this presentation the focus is on lower bounds on $\frac{1}{N}\sum_{i=1}^{N} Z_i^2$, though the universality of lower bounds is true in far more general situations, involving the sum of nonnegative random variables.

7.1 The small-ball condition

The small-ball condition quantifies the fact that the random variable Z does not assign too much 'weight' close to 0. It has nothing to do with the rate of decay of Z; that calibrates the L_p or ψ_α norms of Z.

Definition 7.2. *A random variable $Z \in L_2$ satisfies a small-ball condition with constants κ and δ if*

$$Pr(|Z| \geq \kappa \|Z\|_{L_2}) \geq \delta.$$

Example 7.3. *Let us give two extreme examples of random variables that satisfy a small-ball condition.*

- *Let g be the standard gaussian random variable. It is straightforward to verify that for $0<\varepsilon<1$, $Pr(|g|<\varepsilon) \leq c_0\varepsilon$, where c_0 is a suitable absolute constant. Therefore, given any $0<\delta<1$, $Pr(|g| \geq c_1\delta\|g\|_{L_2}) \geq 1-\delta$.*

- *Let $h(t) = c_2/1+|t|^4$ and set Z to be a random variable whose density is h. Then $\mathbb{E}Z^2 \sim 1$, but the tail decay of Z is slow. In fact, $Z \notin L_3$. On the other hand, Z satisfies the same small-ball condition as the gaussian: close to 0 its density is of the order of a constant. Hence, for $0<\delta<1$, $\int_{-\delta}^{\delta} h(t)dt \sim \delta$, implying that $Pr(|Z| \geq c_2\delta\|Z\|_{L_2}) \geq 1-\delta$.*

The reason why a small-ball condition is so useful is the following almost trivial observation:

Lemma 7.4. *There exists an absolute constant c for which the following holds. If $Z \in L_2$ is a random variable that satisfies a small-ball condition with constants κ and δ and $Z_1, ..., Z_N$ are independent copies of Z then with probability at least $1 - 2\exp(-c\delta N)$,*

$$|\{i : |Z_i| \geq \kappa \|Z\|_{L_2}\}| \geq \frac{\delta N}{2}.$$

In particular, on that event,

$$\frac{1}{N}\sum_{i=1}^{N} Z_i^2 \geq \frac{1}{2}\kappa^2 \delta \|Z\|_{L_2}^2.$$

Proof. Let $\eta_i = \mathbb{1}_{\{|Z_i| \geq \kappa \|Z\|_{L_2}\}}$. Thus, $(\eta_i)_{i=1}^N$ are independent selectors with mean at least δ. By the concentration of iid selectors (see (3.14)), the first claim follows. The second claim is an immediate outcome of the first. ∎

Remark 7.5. *Note that the lower bound on $\frac{1}{N}\sum_{i=1}^{N} Z_i^2$ is stable, because it the result of having a proportional number ($\delta N/2$) of the Z_is are of the order of $\|Z\|_{L_2}$. Even if a malicious adversary is allowed to change any subset of $\{Z_1, ..., Z_N\}$ of cardinality at most $\delta N/4$, the lower bound would still hold—with a constant of $\kappa^2\delta/4$ instead of $\kappa^2\delta/2$. This is our first encounter with a recurring theme, of a* stable lower bound.

Because of its importance, let us turn to a more detailed study of the small-ball condition, starting with the all-important question of ways in which a small-ball condition can be verified.

The first example is rather obvious, as it was featured in Example 7.3:

Lemma 7.6. *Let Z be a random variable that has a density bounded by L. Then for every $\varepsilon > 0$,*

$$Pr(|Z| < \varepsilon \|Z\|_{L_2}) \leq \varepsilon \cdot 2L\|Z\|_{L_2}.$$

The proof is obvious: if h denotes the density of Z then

$$Pr(Z \in [a, b]) = \int_a^b h(t)dt \leq L(b - a),$$

and setting $a = -\varepsilon\|Z\|_{L_2}$ and $b = \varepsilon\|Z\|_{L_2}$ the claim follows. ∎

7.1.1 Norm equivalence

The natural hierarchy of the L_p spaces implies that $\|Z\|_{L_p} \leq \|Z\|_{L_q}$ when $1 \leq p \leq q \leq \infty$. But suppose that one is in a situation where the inequality is reversed, and there is some constant L such that $\|Z\|_{L_q} \leq L\|Z\|_{L_p}$. This norm equivalence forces Z to have nontrivial structure: a significant part of Z's weight "lives" in the interval $[\alpha\|Z\|_{L_p}, \beta\|Z\|_{L_p}]$ for constants α and β that depend only L, p and q. In some sense, this is a combination of a small-ball condition (plenty of weight in $[\alpha\|Z\|_{L_p}, \infty)$ and a relatively nice tail decay, i.e., not too much weight in $[\beta\|Z\|_{L_p}, \infty)$. While both facts are useful, the one that is more significant for this discussion is being far from 0, and that is an outcome of the Paley-Zygmund inequality.

Theorem 7.7. *Let $1 \leq p < q \leq \infty$. If $\|Z\|_{L_q} \leq L\|Z\|_{L_p}$ then for $0 < \theta < 1$,*

$$Pr\left(|Z| \geq \theta\|Z\|_{L_p}\right) \geq \left(\frac{1 - \theta^p}{L^p}\right)^{\frac{q}{q-p}}.$$

Proof. By Hölder's inequality for $r = q/p > 1$ and its conjugate index $r' = q/(q-p)$,

$$\begin{aligned}\|Z\|_{L_p}^p =& \mathbb{E}|Z|^p \mathbb{1}_{\{|f|\geq t\}} + \mathbb{E}|Z|^p \mathbb{1}_{\{|Z|<t\}} \leq (\mathbb{E}|Z|^{p\cdot q/p})^{p/q} \cdot Pr^{1-p/q}(|Z|\geq t) + t^p \\ \leq& \|Z\|_{L_q}^p Pr^{1-p/q}(|Z|\geq t) + t^p.\end{aligned}$$

Therefore,

$$Pr(|Z|\geq t) \geq \left(\frac{\|Z\|_{L_p}^p - t^p}{\|Z\|_{L_q}^p}\right)^{\frac{q}{q-p}}. \tag{7.1}$$

If $\|Z\|_{L_q} \leq L\|Z\|_{L_p}$, $0<\theta<1$ and $t = \theta\|Z\|_{L_p}$, one has

$$Pr(|Z|\geq \theta\|Z\|_{L_p}) \geq \left(\frac{1-\theta^p}{L^p}\right)^{\frac{q}{q-p}}.$$

■

Exercise 20. *Show that for α, β and δ that depend only on p, q and L, one has*

$$Pr\left(Z \in \left[\alpha\|Z\|_{L_p}, \beta\|Z\|_{L_p}\right]\right) \geq \delta.$$

7.1.2 Small-ball and linear functionals

An important set of functions that is of considerable interest consists of linear functionals on $\mathbb{R}^N$. Let us examine when linear functionals satisfy a small-ball condition with an important twist: the goal is to have a uniform estimate, namely, if X is a random vector in $\mathbb{R}^N$, then there are constants κ and δ such that for every $t \in S^{N-1}$,

$$Pr(|\langle X, t\rangle| \geq \kappa\|\langle X, t\rangle\|_{L_2}) \geq \delta$$

(obviously, the implicit assumption is that X has a nontrivial covariance, and in particular, all linear functionals are square-integrable).

By the Paley-Zygmund inequality, if the class of linear functionals satisfies any sort of norm equivalence, for example, that there is some $p > 2$ such that for every $t \in S^{d-1}$ $\|\langle X, t\rangle\|_{L_p} \leq L\|\langle X, t\rangle\|_{L_2}$, then the small-ball condition holds with constants κ and δ that depend only on L. However, proving such a norm equivalence is often a nontrivial task.

Exercise 21. *Let X be an isotropic random vector in $\mathbb{R}^N$.*

(1) *Give some examples of random vectors that satisfy $\left(\mathbb{E}\|X\|_2^4\right)^{1/4} \leq L\sqrt{N}$ for an absolute constant L.*

(2) *Show that given such a random vector, "most" directions $\theta \in S^{N-1}$ satisfy a small ball condition with constants κ and δ that depend only on L.*

Consider a special choice of a random vector: $\mathcal{Z} = (Z_1, ..., Z_N)$—a random vector whose coordinates are independent copies of a symmetric random variable Z. As it happens, the random vector $\mathcal{Z}$ 'inherits' the small-ball behaviour of random variable Z without resorting to norm equivalence.

Theorem 7.8. *There exist absolute constant c_1 and c_2 for which the following holds. Let Z be a symmetric random variable that satisfies a small-ball condition with constants κ and δ. If $Z_1, ..., Z_N$ are independent copies of Z then, for every $(a_1, ..., a_N) \in \mathbb{R}^N$,*

$$Pr\left(\left|\sum_{i=1}^N a_i Z_i\right| \geq c_1\sqrt{\delta}\kappa\right) \geq c_2\delta.$$

Proof. Since Z is symmetric, $\sum_{i=1}^N a_i X_i$ has the same distribution as $\sum_{i=1}^N \varepsilon_i |a_i| \cdot |X_i|$. Observe that for every $(b_i)_{i=1}^N \in \mathbb{R}^N$, if $Y = \sum_{i=1}^N \varepsilon_i b_i$ then

$$Pr_\varepsilon\left(\left|\sum_{i=1}^N \varepsilon_i b_i\right| \geq c_1\|b\|_2\right) \geq c_2 \tag{7.2}$$

for absolute constants c_1 and c_2. Indeed, because a Bernoulli vector is c_0-subgaussian for an absolute constant c_0, one has

$$\left\|\sum_{i=1}^N \varepsilon_i b_i\right\|_{L_2} = \|b\|_2 \quad \text{and} \quad \left\|\sum_{i=1}^N \varepsilon_i b_i\right\|_{L_4} \lesssim \|b\|_2.$$

Now (7.2) follows from the Paley-Zygmund Theorem.

Next, let $b_i = |a_i| \cdot |Z_i|$ and apply (7.2) for a realization of $Z_1, ..., Z_N$. Thus,

$$Pr_\varepsilon\left(\left|\sum_{i=1}^N \varepsilon_i |a_i| \cdot |Z_i|\right| \geq c_1\left(\sum_{i=1}^N a_i^2 Z_i^2\right)^{1/2}\right) \geq c_2, \tag{7.3}$$

and it remains to obtain a lower bound on $(\sum_{i=1}^N a_i^2 Z_i^2)^{1/2}$ that holds with a high enough probability. To that end, let $\eta_i = \mathbb{1}_{\{|Z_i|>\kappa\}}$; thus $(\eta_i)_{i=1}^N$ are independent selectors with mean at least δ, and pointwise,

$$\sum_{i=1}^N a_i^2 Z_i^2 \geq \kappa \sum_{i=1}^N a_i^2 \eta_i.$$

Let $W = \sum_{i=1}^N a_i^2 \eta_i$ and observe that $\mathbb{E}W = \delta\|a\|_2^2$. Moreover,

$$\mathbb{E}W^2 = \delta^2\left(\sum_{i=1}^N a_i^2\right)^2 - \delta^2 \sum_{i=1}^N a_i^4 + \delta \sum_{i=1}^N a_i^4,$$

and

$$\|W\|_{L_2} \leq \delta\left(\sum_{i=1}^N a_i^2\right) + \sqrt{\delta}\left(\sum_{i=1}^N a_i^4\right)^{1/2} \leq 2\sqrt{\delta}\left(\sum_{i=1}^N a_i^2\right).$$

Applying the Paley-Zygmund Theorem again, using that $\|W\|_{L_2}/\|W\|_{L_1} \leq 2/\sqrt{\delta}$ one has

$$Pr\left(W \geq \frac{\delta}{2}\|a\|_2^2\right) \geq \frac{\delta}{8},$$

and therefore,

$$Pr\left(\sum_{i=1}^N a_i^2 Z_i^2 \geq \kappa^2\delta \sum_{i=1}^N a_i^2\right) \geq \frac{\delta}{8}. \tag{7.4}$$

The claim follows by combining (7.3) and (7.4). ∎

7.2 Uniform lower bounds

A crucial part in the study of lower bounds is the ability to obtain such bounds that hold simultaneously for a class of functions. The *small-ball method* is a way of deriving such uniform estimates, and the standard fist step is a uniform version of the small-ball condition.

Definition 7.9. *The class $H \subset L_2(\mu)$ satisfies a small-ball condition with constants κ and δ if for every $f, h \in H \cup \{0\}$,*

$$Pr\left(|f-h|(X) \geq \kappa \|f-h\|_{L_2}\right) \geq \delta.$$

Let us formulate a version of the key estimate that leads to a uniform lower bound.

Theorem 7.10. *There exist absolute constants c_1 and c_2 for which the following holds. Let $H \subset rS(L_2)$ be a class that satisfies a small-ball condition with constants κ and $\delta \geq 1/N$. If*

$$\mathbb{E} \sup_{h \in H} \left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i h(X_i) \right| \leq c_1 (\kappa\delta)^2 r,$$

then with probability at least $1 - 2\exp(-c_2\delta^2 N)$, for every $h \in H$,

$$\left| \left\{ i : |h(X_i)| \geq \frac{\kappa}{2} r \right\} \right| \geq \frac{\delta N}{4}.$$

Remark 7.11. *There are various cases in which the probability estimate can be improved to $1 - 2\exp(-c\delta N)$. We shall return to the point after the proof of Theorem 7.10.*

The idea behind the proof of Theorem 7.10 is simple, and versions of the same idea can be used in many other cases. To make the presentation as general as possible, let us consider the following problem. Suppose that there is a certain property **P** that one would like to show is satisfied by every $v \in P_\sigma H$ (in the context of Theorem 7.10, that property is that each $v \in P_\sigma H$ has a proportional number of coordinates that are at least of the order of r.

(1) The first step in deriving a uniform estimate is to show that for every fixed $h \in H$, $P_\sigma h$ satisfies **P** with very high probability. That is usually achieved thanks to the underlying assumption on H, and the independence of $X_1, ..., X_N$. In the context of Theorem 7.10, the small-ball condition implies that with probability at least $1 - 2\exp(-c\delta N)$, the cardinality of the set $I_h = \{i : |h(X_i)| \geq \kappa \|h\|_{L_2}\}$ is at least $\delta N/2$.

(2) The high probability with which each $P_\sigma h$ satisfies **P** implies that one can have such a control uniformly for a large number of class members. Indeed, by the union bound, it is possible to have $|I_h| \sim \delta N$ for every $h \in H'$ as long as $|H'| \leq \exp(c_1 \delta N)$ for a suitable absolute constant c_1.

The natural choice of H' is an appropriate approximating subset of H, and in the context of Theorem 7.10 is will be a maximal separated set with respect to the L_2 norm (though obviously one can think of other alternatives that make sense in various examples). Denote the 'mesh width' of the separated set by ρ.

(3) The crucial part of the argument is the uniform estimate: that if h and f are 'close enough', and **P'** is a slightly weaker version of **P** then the fact that f satisfies **P** implies that h satisfies **P'**. In other words, a perturbation does not 'spoil' **P** by too much.

(4) By combining (2) and (3) one has that with high probability, every $h \in H$ satisfies (the slightly weaker) **P** which is the estimate one is looking for.

In the context of Theorem 7.10, f satisfies **P'** if

$$\left|\left\{i : |f(X_i)| \geq \frac{\kappa}{2} r\right\}\right| \geq \frac{\delta N}{4},$$

and to establish (3) one has to show that with high probability

$$\sup_{\{f,h\in H,\ \|f-h\|_{L_2}\leq\rho\}} |\{i : |f-h|(X_i) \geq \kappa r/2\}| \leq \frac{\delta N}{4}.$$

Indeed, recall that $H \subset rS(L_2)$, that $H' \subset H$ is a maximal ρ-separated subset of H and therefore, if $h \in H$ and πh denotes its best approximation in H' then $\|h - \pi h\|_{L_2} \leq \rho$. Hence, it follows that on the intersection of the events from (2) and (3) that:

- by (2), for every $h' \in H'$, $|\{i : |h(X_i)| \geq \kappa r\}| \geq \delta N/2$ (i.e., each point in the net satisfies **P**).
- by (3), for every $h \in H$, $|\{i : |(h-f)(X_i)| \geq (\kappa/2)r\}| \leq \delta N/4$ (i.e., perturbations do not spoil **P** by 'too much').

Therefore, for every $h \in H$ there is $I_h \subset \{1, ..., N\}$ such that $|I_h| \geq \delta N/4$ and for every $i \in I_h$,

$$|h(X_i)| \geq |\pi h(X_i)| - |(h - \pi h)(X_i)| \geq \kappa r - \frac{\kappa}{2} r \geq \frac{\kappa}{2} r.$$

On that event, every function in H satisfies the slightly weaker $\mathbf{P}'$, as required.

Proof of Theorem 7.10. Part (1) is simply Lemma 7.4. Now, let us turn to (2): let $H' \subset H$ be a maximal ρ-separated subset of H for ρ which satisfies that

$$\log \mathcal{M}(H, \rho D) \leq \frac{c_0}{2} \delta N. \tag{7.5}$$

Invoking the individual estimates, one has that with probability at least $1 - 2\exp(-(c_0/2)\delta N)$, for every $h' \in H'$, there is $I_{h'} \subset \{1, ..., N\}$ such that $|I_{h'}| \geq \delta N/2$ and for every $i \in I_{h'}$,

$$|h'(X_i)| \geq \kappa \|h'\|_{L_2} = \kappa r. \tag{7.6}$$

Next, let us turn to the "main event"—establishing (3). One has to show that with high probability,

$$\sup_{\{h,f\in H,\ \|h-f\|_{L_2}\leq\rho\}} \left|\{i : |f-h|(X_i) \geq \frac{\kappa}{2} r\}\right| \leq \frac{\delta N}{4}.$$

To that end, set $U = (H - H) \cap \rho D$ and let us estimate the probability with which

$$(*) = \sup_{u\in U} \sum_{i=1}^{N} \mathbb{1}_{\{|u|(X_i)\geq(\kappa/2)r\}}. \tag{7.7}$$

Observe that $(*)$ is the supremum of a sum of independent, binary-valued random variable. By Talagrand's concentration inequality for bounded empirical processes, one has that with probability at least $1 - 2\exp(-x)$,

$$(*) \leq C\left(\mathbb{E}(*) + \sigma\sqrt{xN} + xN\right),$$

where

$$\sigma^2 = \sup_{u\in U} \mathbb{E}\mathbb{1}_{\{|u|(X)\geq(\kappa/2)r\}} = Pr(|u|(X) \geq (\kappa/2)r) \leq \frac{4\|u\|_{L_2}^2}{\kappa^2 r^2} \leq \frac{4\rho^2}{\kappa^2 r^2}.$$

Thus, to have a chance that $(*) \leq \delta N/4$ one must ensure that

$$x \lesssim \min N \left\{ \delta, \delta^2 \left(\frac{\kappa r}{\rho}\right)^2 \right\}. \tag{7.8}$$

All that is left is to estimate $\mathbb{E}(*)$, which based on standard methods from empirical processes theory. Clearly,

$$\begin{aligned}\mathbb{E}(*) \leq & \frac{2}{\kappa r}\mathbb{E}\sup_{u\in U}\sum_{i=1}^N |u(X_i)| \leq \frac{2}{\kappa r}\left(\mathbb{E}\sup_{u\in U}\left|\sum_{i=1}^N (|u(X_i)| - \mathbb{E}|u|)\right| + N\sup_{u\in U}\mathbb{E}|u|\right) \\ \leq & \frac{4}{\kappa r}\left(\mathbb{E}\sup_{u\in U}\left|\sum_{i=1}^N \varepsilon_i u(X_i)\right| + N\rho\right) \leq \frac{8}{\kappa r}\left(\mathbb{E}\sup_{h\in H}\left|\sum_{i=1}^N \varepsilon_i h(X_i)\right| + N\rho\right),\end{aligned}$$

where we have used the fact that $\|u\|_{L_1} \leq \|u\|_{L_2} \leq \rho$; the Giné-Zinn symmetrization inequality; the contraction inequality for Bernoulli processes; and the triangle inequality. As a result $\mathbb{E}(*) \leq \delta N/16$ provided that

$$\frac{\rho}{\kappa r} \lesssim \delta \quad \text{and} \quad \mathbb{E}\sup_{h\in H}\left|\frac{1}{N}\sum_{i=1}^N \varepsilon_i h(X_i)\right| \lesssim \kappa\delta r. \tag{7.9}$$

Taking into account the first part of (7.9) and (7.5), it suffices to show that for a suitable absolute constant c_2,

$$\log \mathcal{M}(H, c_2\kappa\delta r) \leq \frac{c_0}{2}\delta N. \tag{7.10}$$

The proof of that fact is based on Sudakov's inequality for Bernoulli processes, which is presented at a later stage of these notes.

Theorem 7.12. *There exists an absolute constant c for which the following holds. Let $p \geq 1$. Consider $T \subset \mathbb{R}^N$, a set of cardinality $\exp(p)$ which is η-separated with respect to the $L_p(\mathcal{E})$ norm (that is, for every $t_1, t_2 \in T$, $\|\langle t_1 - t_2, \mathcal{E}\rangle\|_{L_p} \geq \eta$). Then*

$$\mathbb{E}\sup_{t\in T}\sum_{i=1}^N \varepsilon_i t_i \geq c\eta.$$

Assume that there is a set $H_1 \subset H$ consisting of $\exp(k)$ points in H that are βr separated with respect to the L_2 norm for $k \leq (c_0/4)\delta N$. Just as in the proof of (1), by the small-ball condition and the union bound, one has that with probability at least $1 - 2\exp(-(c_0/2)\delta N)$, for every $f, h \in H_1$,

$$|\{i : |(f-h)(X_i)| \geq \kappa\beta r\}| \geq \frac{1}{2}\delta N. \tag{7.11}$$

Conditioned on this event set $W = \{(h(X_i))_{i=1}^N : h \in H_1\}$. Observe that for every $u, v \in W$ and $p \leq \delta N/4$

$$\|\langle u - v, \mathcal{E}\rangle\|_{L_p} \gtrsim \sqrt{p}\kappa\beta r\sqrt{\delta N};$$

indeed, that is the case because thanks to (7.11) each vector $u - v$ has at least $\delta N/2 \geq 2p$ "large coordinates". Recalling that

$$\left\|\langle u - v, \mathcal{E}\rangle\right\|_{L_p} \gtrsim \sqrt{p}\left(\sum_{i\geq p}[(u_i - v_i)^*]^2\right)^{1/2},$$

Theorem 7.12 implies that

$$\mathbb{E}_\varepsilon \sup_{w\in W}\left|\sum_{i=1}^N \varepsilon_i w_i\right| \gtrsim \sqrt{p}\kappa\beta r\sqrt{\delta N} \gtrsim \kappa\beta r\delta N,$$

where the last inequality holds by setting $k = p \sim \delta N$. Since the event we have conditioned on has probability at least $1/2$ one has

$$\mathbb{E}\sup_{h\in H}\left|\sum_{i=1}^N \varepsilon_i h(X_i)\right| \geq c_3\kappa\beta r\delta N.$$

In other words, setting $\beta = c_2\kappa\delta$, if the reverse inequality holds, i.e., if

$$\mathbb{E}\sup_{h\in H}\left|\frac{1}{N}\sum_{i=1}^N \varepsilon_i h(X_i)\right| \lesssim (\kappa\delta)^2 r, \tag{7.12}$$

then $\log \mathcal{M}(H, \beta r) \leq (c_0/2)\delta N$, as required. ∎

It should be stressed that the proof of Theorem 7.10 reveals the claim can potentially be loose, and there could be instances in which the estimate may be improved. The sources of possible looseness are:

- Using $(2/\kappa r)\mathbb{E}\sup_{u\in U}\left|\sum_{i=1}^N \varepsilon_i u(X_i)\right|$ to control $\mathbb{E}\sup_{u\in U}\sum_{i=1}^N \mathbb{1}_{\{|u|(X_i)\geq(\kappa/2)r\}}$ is frequently very convenient, but at the same time it is far from sharp. There are some important examples in which the latter can be bounded directly, leading to a much better outcome.

- Replacing $U = (H - H)\cap \rho D$ with H can be very costly, especially if ρ, which is the mesh-width of a net of log-cardinality $\sim \delta N$, is very small. Again, there are examples where that reduction is unwise.

- Estimating ρ using Sudakov's inequality for Bernoulli processes is, again, convenient but suboptimal. That leads to a fixed-point condition on the Bernoulli process that scales like δ^2 rather than like δ, and to a probability estimate of the order of $1 - 2\exp(-c\delta^2 N)$ rather than better estimate of $1 - 2\exp(-c\delta N)$.

Since there are important examples in which each one of these steps is loose, let us formulate two relatively 'raw' versions of Theorem 7.10. The two are not the most convenient formulations, and can be made more 'clean' just as in the proof of Theorem 7.10, but that will become handy in some examples, including the proof of the main result of these notes—sample complexity bounds on estimation and prediction.

A part of one raw version is a modification of the small-ball condition: instead of considering $H \subset rS(L_2)$ and assuming that $Pr(|h(X)| \geq \kappa r) \geq \delta$ it suffices to assume that $Pr(|h(X)| \geq \kappa) \geq \delta$. While the outcome might slightly change, the path the proofs take remain the same.

Theorem 7.13. *Let $H \subset L_2(\mu)$ and assume that for every $h \in H$, $Pr(|h(X)| \geq \kappa) \geq \delta$. Assume further that*

(1) $\log \mathcal{M}(H, \rho D) \leq c_1 \delta N \cdot \log(2/(1-\delta))$*;*

(2) *if $U = (H-H) \cap \rho D$ then $\mathbb{E} \sup_{u \in U} \sum_{i=1}^N \mathbb{1}_{\{|u|(X_i) \geq (\kappa/2)\}} \leq \delta N/16$.*

Then with probability at least

$$1 - 2\exp(-c_2 N \min\left\{\delta, \delta^2 \left(\frac{\kappa}{\rho}\right)^2\right\},$$

for every $h \in H$,

$$|\{i : |h(X_i)| \geq (\kappa/2)\}| \geq \frac{1}{4}\delta N.$$

Remark 7.14. *Note that the factor of $\log(2/(1-\delta))$ in (1); it is meaningful when δ is very close to 1, or, in other words, when the probability of "failure" for each random variable is close to 0. As a result, if $\eta_i = \mathbb{1}_{\{|h(X_i)| \geq \kappa\}}$ and $\mathbb{E}\eta_i$ is close to 1, then the probability that $\sum_{i=1}^N \eta_i \geq \delta N/2$ is very high—the tail estimate is in the Poisson range of Bennett's inequality. This simple observation plays a key role in some of the applications presented in what follows.*

Also, just as in the proof of Theorem 7.10 one may show that

$$\mathbb{E} \sup_{u \in U} \sum_{i=1}^N \mathbb{1}_{\{|u|(X_i) \geq (\kappa/2)\}} \leq \frac{c}{\kappa} \mathbb{E} \sup_{u \in (H-H) \cap \rho D} \left| \frac{1}{N} \sum_{i=1}^N \varepsilon_i u(X_i) \right|,$$

leading to a slightly relaxed version of condition (2) in Theorem 7.13.

In a similar fashion, one may prove an analogous bound, in which the "likely events" satisfy the reverse inequality, and for reasons that will become clearer in what follows, one should consider δ as being relatively close to 1.

Theorem 7.15. *Let $H \subset L_2(\mu)$ and assume that for every $h \in H$, $Pr(|h(X)| \leq \kappa) \geq \delta \geq 0.99$. Assume further that, for suitable absolute constants c_1 and c_2,*

(1) $\log \mathcal{M}(H, \rho D) \leq c_1 \delta N \cdot \log(2/(1-\delta))$*, and*

(2) *if $U = (H-H) \cap \rho D$ then $\mathbb{E} \sup_{u \in (H-H) \cap \rho D} \left| \frac{1}{N} \sum_{i=1}^N \varepsilon_i u(X_i) \right| \leq c_2 \kappa$.*

Then with probability at least

$$1 - 2\exp\left(-c_3 N \min\left\{1, \left(\frac{\kappa}{\rho}\right)^2\right\}\right),$$

for every $h \in H$,

$$|\{i : |h(X_i)| \leq (3\kappa/2)\}| \geq 0.9N.$$

Exercise 22. *Prove Theorem 7.13 and Theorem 7.15.*

Chapter 8

Simple outcomes of Theorem 7.10

In what follows we implicitly assume that all the functions involved satisfy a small-ball condition with constants κ and δ (or the reverse condition if needed). It suffices that the right condition holds in $\mathrm{span}(F)$, but often much less is needed. For the sake of simplicity we make no attempt of pin-pointing the minimal assumptions needed for the proofs to work. Instead, the discussion is restricted to cases in which the required condition holds on $\mathrm{span}(F)$.

8.1 Proof of Theorem 5.29

At this point there is enough machinery set in place to prove Theorem 5.29, thus obtaining an constant confidence estimate on the estimation error when F is convex and satisfies a small-ball condition. Instead of presenting a complete proof, let us highlight the components that are needed in the proof.

The first observation is that by the quadratic-multiplier decomposition, and recalling that $\xi = f^*(X) - Y$,

$$P_N \mathcal{L}_f = \frac{1}{N}\sum_{i=1}^N (f-f^*)^2(X_i) + 2\left(\frac{1}{N}\sum_{i=1}^N \xi_i (f-f^*)(X_i) - \mathbb{E}\xi(f-f^*)\right) + 2\mathbb{E}\xi(f-f^*).$$

And, because F is convex, the characterization of the nearest point map onto a convex subset of a Hilbert space implies that for any $f \in F$, $\mathbb{E}\xi(f-f^*)(X) \geq 0$.

Invoking Theorem 7.10, it follows that with high probability, if $\|f-f^*\|_{L_2} = r \geq r_{\mathbb{Q}}$ then

$$P_N (f-f^*)^2 \geq c_0 \|f-f^*\|_{L_2}^2,$$

where c_0 depends on the constants in the small-ball condition. By homogeneity and the star-shape property, the same lower bound is true if $\|f-f^*\|_{L_2} \geq r$.

Next, Chebyshev's inequality implies that if $r \geq r_{\mathbb{M}}$ (for the right choice of constants) then with constant probability, for every $f \in F$ such that $\|f-f^*\|_{L_2} = r$,

$$\left|\frac{1}{N}\sum_{i=1}^N \xi_i (f-f^*)(X_i) - \mathbb{E}\xi(f-f^*)\right| \leq \frac{c_0}{4} r^2 = \frac{c_0}{4}\|f-f^*\|_{L_2}^2. \tag{8.1}$$

Again, by the star-shape property and on the same event, (8.1) holds whenever $\|f-f^*\|_{L_2} \geq r$.

Setting $r \geq \max\{r_{\mathbb{Q}}, r_{\mathbb{M}}\}$, one has that with constant probability, if $\|f-f^*\|_{L_2} \geq r$ then $P_N \mathcal{L}_f > 0$, implying that on that event, $\|f-f^*\|_{L_2} \leq r$. ∎

It is instructive to see what feature of this argument must be improved to obtain Theorem 5.29 in its full generality: an estimate of the prediction error and for classes that satisfy a convexity condition rather than being convex. The key point is that Theorem 7.10 only provides an *isomorphic* lower bound on $P_N(f-f^*)^2$, namely, that $P_N(f-f^*)^2 \geq c\|f-f^*\|_{L_2}^2$, but the constant c need not be close to 1. The proof of Theorem 5.29 calls for an *almost isometric* estimate: that $P_N(f-f^*)^2 \geq (1-\zeta)\|f-f^*\|_{L_2}^2$, where ζ can be as close to 0 as we like. Unfortunately, such an estimate is beyond the scope of these notes; the interested reader can find the result in [**?**], with the critical level being $r_{\mathbb{Q}}$ for a constant that depends on ζ.

Let us show how such an almost isometric estimate can be used in the study of the prediction error of ERM. Observe that by the quadratic-multiplier decomposition and the convexity condition,

$$\begin{aligned} P_N\mathcal{L}_f \geq & P_N\mathcal{Q}_{f-f^*} - \mathbb{E}\mathcal{Q}_{f-f^*} - |P_N\mathcal{M}_{f-f^*} - \mathbb{E}\mathcal{M}_{f-f^*}| + \mathbb{E}\mathcal{L}_f \\ \geq & P_N\mathcal{Q}_{f-f^*} - \mathbb{E}\mathcal{Q}_{f-f^*} - |P_N\mathcal{M}_{f-f^*} - \mathbb{E}\mathcal{M}_{f-f^*}| + \frac{1}{2B}\|f-f^*\|_{L_2}^2 + \frac{1}{2}\mathbb{E}\mathcal{L}_f. \end{aligned}$$

Invoking the almost isometric lower estimate on the quadratic component for $\zeta = 1/(4B)$, it follows that if $r \geq r_{\mathbb{Q}}$ and $\|f-f^*\|_{L_2} \geq r$, one has

$$P_N\mathcal{Q}_{f-f^*} - \mathbb{E}\mathcal{Q}_{f-f^*} = P_N\mathcal{Q}_{f-f^*} - \|f-f^*\|_{L_2}^2 \geq -\zeta\|f-f^*\|_{L_2}^2.$$

Therefore, on that event,

$$\begin{aligned} P_N\mathcal{L}_f \geq & \left(\frac{1}{2B} - \zeta\right)\|f-f^*\|_{L_2}^2 - |P_N\mathcal{M}_{f-f^*} - \mathbb{E}\mathcal{M}_{f-f^*}| + \frac{1}{2}\mathbb{E}\mathcal{L}_f \\ \geq & -|P_N\mathcal{M}_{f-f^*} - \mathbb{E}\mathcal{M}_{f-f^*}| + \frac{1}{4B}\|f-f^*\|_{L_2}^2 + \frac{1}{2}\mathbb{E}\mathcal{L}_f. \end{aligned}$$

From here the proof continues along the lines of Theorem 6.9, where a similar functional was analyzed to control the prediction error. ■

8.2 Dealing with malicious noise

Let $F \subset L_2(\mu)$ be a class of functions and set $f^* \in F$. Let $H_r = \mathrm{star}(F-f^*,0) \cap rS$ where r satisfies that

$$\mathbb{E}\sup_{h\in H_r}\left|\frac{1}{N}\sum_{i=1}^N \varepsilon_i h(X_i)\right| \leq c_1(\kappa,\delta)r,$$

and in particular, Theorem 7.10 holds for H_r: with probability at least $1-2\exp(-c_2(\kappa,\delta)N)$, for every $h \in H$ one has

$$|\{i : |h|(X_i) \geq (\kappa/2)\|h\|_{L_2}\}| \geq \frac{1}{4}\delta N. \tag{8.2}$$

Observe that functions if H are of a particular form: if $\|f-f^*\|_{L_2} \geq r$, there is $0 < \lambda \leq 1$ such that $\lambda(f-f^*) \in H_r$. And, since (8.2) is positive homogeneous, the same claim holds for any $f-f^*$ as long as $\|f-f^*\|_{L_2} \geq r$; i.e.,

With probability at least $1 - 2\exp(-c_2\delta N)$, for any $f \in F$ such that $\|f - f^*\| \geq r$ one has

$$|\{i : |f - f^*|(X_i) \geq (\kappa/2)\|f - f^*\|_{L_2}\}| \geq \frac{1}{4}\delta N. \tag{8.3}$$

Now, consider the following recovery problem in F: assume that one wishes to identify $f^* \in F$ given only a random sample of the form $f^*(X_1), ..., f^*(X_N)$, where $X_1, ..., X_N$ are independent, distributed as the (unknown) X. In a cruel twist of events, the Devil, (which from here on will be called "Toby"), can change ηN of the sample points in any way that he see fit. Therefore, instead of seeing the true values $f^*(X_1), ..., f^*(X_N)$ the given data is the corrupted sample $Z_1, ..., Z_N$, where $|\{i : f^*(X_i) \neq Z_i\}| \leq \eta N$. The question is how well (that is, with what accuracy and confidence) can f^* be approximated given the corrupted data?

Thanks to (8.3) one possible method of recovery is as follows. Assume that $\eta \leq \delta/16$, which means that Toby can corrupt only a small fraction of the points—and "small" is relative to the constant in the small-ball condition). Select any $f \in F$ such that $|\{i : f(X_i) \neq Z_i\}| \leq \delta N/16$.

First of all, observe that such a function exists: $f^*(X_i) = Z_i$ for at least $(1-\eta)N$ indices, and $(1-\eta)N \geq (1-\delta)N/16 > \delta N/16$ provided that $\delta < 1/2$, which one can clearly assume without loss of generality. Second, on the event on which (8.3) holds, if $\|f - f^*\|_{L_2} \geq r$ then for at least $\delta N/4$ of the indices one has that $f(X_i) \neq f^*(X_i)$. Therefore, even after Toby's malicious interference there are still at least $\delta N/4 - \eta N \geq (3/16)\delta N$ indices such that $f(X_i) \neq Z_i$. As a result, if $\|f - f^*\|_{L_2} > r$ there will be too much disagreement between the values $f(X_i)$ and Z_i, and such a function will not be a candidate. Hence,

With probability at least $1-2\exp(-c_0\delta N)$ any function selected by the procedure satisfies that $\|f - f^*\|_{L_2} \leq r$. The choice of r is the smallest one such that

$$\mathbb{E} \sup_{u \in F_{f^*,r}} \left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i u(X_i) \right| \leq c_1 r,$$

and c_0, c_1 are constants that depend only on the small-ball condition satisfied in $\mathrm{span}(F)$.

8.3 Geometric applications I

An immediate outcome of (8.3) is as follows. Let $H = \mathrm{star}(F, 0) \cap rS$, where r satisfies that

$$\mathbb{E} \sup_{u \in \mathrm{star}(F,0) \cap rS} \left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i u(X_i) \right| \leq c_1 r.$$

It follows that with probability at least $1 - 2\exp(-c_2 N)$, if $\|f\|_{L_2} \geq r$ then

$$|\{i : |f(X_i)| \geq (\kappa/2)\|f\|_{L_2}\}| \geq \frac{1}{4}\delta N,$$

and in particular,

$$\frac{1}{N} \sum_{i=1}^{N} f^2(X_i) \geq c_3(\kappa, \delta)\|f\|_{L_2}^2. \tag{8.4}$$

Let us invoke (8.4) to address two well-studied geometric problems. In what follows, X is an isotropic random vector in $\mathbb{R}^d$ that satisfies a small-ball condition in the following sense:

Definition 8.1. *The random vector $X \in \mathbb{R}^d$ satisfies a weak small-ball condition with constants κ and δ if for any $t \in \mathbb{R}^d$,*

$$Pr\left(|\langle X, t\rangle| \geq \kappa \|\langle X, t\rangle\|_{L_2}\right) \geq \delta.$$

In other words, a random vector X satisfies a weak small-ball condition with constants κ and δ if every one dimensional marginal $\langle X, t\rangle$ is a random variable that satisfies a small-ball condition with the same constants. This is the same as saying that the class of linear functionals on $\mathbb{R}^d$ satisfies a small-ball condition. The reason we call this a "weak small-ball condition" is because there is a stronger version in which one has control on a certain small-ball property for every marginal of X, not just one-dimensional marginals.

8.3.1 The smallest singular value of a random matrix

A question that has been studied extensively in recent years has to do with the smallest singular value of a random matrix generated by X: let $X_1, ..., X_N$ be independent copies of X and define the random matrix $\Gamma = N^{-1/2}\sum_{i=1}^N \langle X_i, \cdot\rangle e_i$, which is a matrix whose rows are $X_1, ..., X_N$. Note that the smallest singular value of Γ is given by

$$\lambda_{\min}(\Gamma) = \inf_{t\in S^{d-1}} \|\Gamma t\|_2^2 = \inf_{t\in S^{d-1}} \frac{1}{N}\sum_{i=1}^N \langle X_i, t\rangle^2,$$

and the estimate from (8.4) when $F = \left\{\langle t, \cdot\rangle : t \in S^{d-1}\right\}$ is precisely what is needed to bound $\lambda_{\min}$ from below.

Observe that $\mathrm{star}(F, 0) = \{\langle t, \cdot\rangle : t \in B_2^d\}$; therefore, if one can show that for $r = 1$,

$$\mathbb{E}\sup_{t\in B_2^d}\left|\frac{1}{N}\sum_{i=1}^N \varepsilon_i\langle X_i, t\rangle\right| \leq c_1,$$

then by (8.4), with probability at least $1 - 2\exp(-c_2 N)$, $\lambda_{\min} \geq c_3$. To that end, note that

$$\mathbb{E}\sup_{t\in B_2^d}\left|\frac{1}{N}\sum_{i=1}^N \varepsilon_i\langle X_i, t\rangle\right| = \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^N \varepsilon_i X_i\right\|_2 \leq \frac{(\mathbb{E}\|X\|_2^2)^{1/2}}{\sqrt{N}} = (*).$$

And, since X is isotropic, $\mathbb{E}\|X\|_2^2 = \mathbb{E}\sum_{i=1}^d \langle X, e_i\rangle^2 = d$, implying that

$$(*) \leq \sqrt{\frac{d}{N}} \leq c_1$$

provided that $N \geq c_1^2 d$.

Remark 8.2. *One can show that a sharper estimate on $\lambda_{\min}$ is possible if one assumes that the linear forms $\left\{\langle t, \cdot\rangle : t \in S^{d-1}\right\}$ satisfy an $L_q - L_2$ norm equivalence for some $q > 2$. For example If $q > 4$, one recovers the* Bai-Yin asymptotics *that $\lambda_{\min} \geq 1 - c\sqrt{d/N}$ (see [**?**]). Obtaining an upper estimate on $\lambda_{\max}$ requires additional assumptions, and the key component is the ability to control*

$$\max_{1\leq i\leq N} \frac{\|X\|_2^2}{d}.$$

*The best estimates in this direction can be found in [**?**].*

8.3.2 Gelfand widths of convex bodies

The following question arises naturally in asymptotic geometric analysis.

Question 8.3. *Let K be convex body in $\mathbb{R}^d$ (that is, K is a bounded, convex, centrally-symmetric set with a nonempty interior). Can one find a central section of K of small co-dimension and has a relatively small diameter?*

Estimates on the smallest possible diameter of a section of K for a fixed co-dimension is a well-known parameter from approximation theory (the connection can be seen by identifying K with a unit ball of a normed space). Formally:

Definition 8.4. *Given an integer $N \leq d$, the* Gelfand *N-width of K is defined as*

$$\inf \operatorname{diam}(K \cap E),$$

where the diameter is with respect to the Euclidean norm and E is a subspace of $\mathbb{R}^d$ of co-dimension N.

One way of generating an N co-dimensional subspace of $\mathbb{R}^d$ is as the kernel of a matrix $\Gamma : \mathbb{R}^d \to \mathbb{R}^N$. Historically, in the context of Question 8.3, the matrix Γ that was considered was a gaussian matrix (g_{ij}), with independent, standard gaussian random variables as entries. In [**?**] it was shown that a typical kernel of such a matrix satisfies that

$$\operatorname{diam}(K \cap \ker(\Gamma)) \leq c \frac{\mathbb{E}\|G\|_{K^\circ}}{\sqrt{N}}, \tag{8.5}$$

where G is the standard gaussian random vector in $\mathbb{R}^d$, and $\|z\|_{K^\circ} = \sup_{x \in K} \langle x, z \rangle$.

A very useful observation if the following, linking the way a linear operator acts on certain subset of the sphere endowed by the convex body K and the Gelfand widths of K:

> Let $\Gamma : \mathbb{R}^d \to \mathbb{R}^N$. If one can find $r > 0$ such that
>
> $$\inf_{x \in K \cap r S^{d-1}} \|\Gamma x\|_2^2 > 0,$$
>
> then
>
> $$\operatorname{diam}(K \cap \ker(\Gamma)) \leq 2r.$$

Indeed, K is convex and therefore it is star-shaped around 0. In particular, if $y \in K$ and $\|y\|_2 > r$ then by homogeneity $\|\Gamma y\|_2 > 0$: set $\lambda = r/\|y\|_2 < 1$; thus $\lambda y \in K \cap r S^{d-1}$ and $\|\Gamma y\|_2 = \lambda^{-1} \|\Gamma(\lambda y)\|_2 > 0$. Therefore, $K \cap \ker(\Gamma) \subset r B_2^d$, and hence the diameter of that set is at most $2r$.

Identifying when Γ "acts well" on such subsets of the sphere is not a task that can be performed in complete generality. However, when it comes to random matrices, the situation is much better. To that end, let X be a random vector and consider the matrix $\Gamma = N^{-1/2} \sum_{i=1}^N \langle X_i, \cdot \rangle e_i$, where $X_1, ..., X_N$ are independent copies of a random vector X. If

X happens to be isotropic and satisfies a small-ball condition with constants κ and δ, then finding the wanted value of r is the outcome of (8.4): set

$$H = \left\{ \langle t, \cdot \rangle : t \in K,\ \mathbb{E}\langle X, t\rangle^2 = r^2 \right\} = \left\{ \langle t, \cdot \rangle : t \in K \cap rS^{d-1} \right\}.$$

The condition by Theorem 7.10 is that

$$\mathbb{E} \sup_{t \in K \cap rS^{d-1}} \left| \frac{1}{N} \sum_{i=1}^N \varepsilon_i \langle X_i, t \rangle \right| \leq c_1 r,$$

hence, one is looking for the smallest $r > 0$ such that

$$\mathbb{E} \left\| \sum_{i=1}^N \varepsilon_i X_i \right\|_{(K \cap rS^{d-1})^\circ} \leq c_1 r.$$

For that choice of r, with probability at least $1 - 2\exp(-c_2 N)$, $\inf_{x \in K \cap rS^{d-1}} \|\Gamma x\|_2^2$, implying that $\operatorname{diam}(K \cap \ker(\Gamma)) \leq 2r$.

■

Remark 8.5. *This improves the estimate from [?].*

To see the connection with (8.5), note that the random vector $Z = N^{-1/2} \sum_{i=1}^N \varepsilon_i X_i$ is isotropic, and for $r > 0$,

$$\mathbb{E} \left\| \sum_{i=1}^N \varepsilon_i X_i \right\|_{(K \cap rS^{d-1})^\circ} \leq \frac{1}{\sqrt{N}} \mathbb{E} \left\| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i X_i \right\|_{K^\circ} = \frac{\mathbb{E}\|Z\|_{K^\circ}}{\sqrt{N}};$$

Hence, one may always select $r \sim \mathbb{E}\|Z\|_{K^\circ}/\sqrt{N}$. Moreover, if X is L-subgaussian then Z is also L-subgaussian. One can show that in such a case, for any norm on $\mathbb{R}^d$,

$$\mathbb{E}\|Z\| \leq cL\mathbb{E}\|G\|.$$

In particular, if X is L-subgaussian then

$$\operatorname{diam}(K \cap \ker(\Gamma)) \leq cL \frac{\mathbb{E}\|G\|_{K^\circ}}{\sqrt{N}}$$

with probability at least $1 - 2\exp(-c'(L)N)$.

8.4 Random Polytopes

Information on the connection the following estimate has with the geometry of random polytopes can be found in [?].

Lemma 8.6. *Let $T \subset B_2^d$, set $0 < \alpha < 1$, assume that $N \gtrsim_\alpha d$ and that for every $z \in T$,*

$$Pr\left(|\langle z, X \rangle| \geq \kappa \right) \geq 4 \left(\frac{d}{N} \right)^\alpha. \tag{8.6}$$

With probability at least $1 - 2\exp(-c_0 N^{1-\alpha} d^\alpha)$, for every $z \subset T$, one has that

$$\left| \left\{ i : |\langle z, X_i \rangle| \geq \kappa/2 \right\} \right| \geq c_1 N^{1-\alpha} d^\alpha > 0.$$

To establish that fact one may use Theorem 7.13 for the class $H = \{\langle z, \cdot \rangle : z \in T\}$.

Proof. First, observe that (8.6) plays the analog of the small-ball condition for $\delta = 4(d/N)^\alpha$, implying that $\delta N = 4N^{1-\alpha}d^\alpha$. Moreover, $T \subset B_2^d$, and since X is isotropic, $\mathcal{M}(H, \rho D) = \mathcal{M}(T, \rho B_2^)$. By a volumetric covering estimate for $\rho = 5\exp(-c_0(N/d)^{1-\alpha})$, one has

$$\mathcal{M}(T, \rho B_2^d) \leq \left(\frac{5}{\rho}\right)^d \leq \exp(c_1 \delta N)$$

as required in (1) of Theorem 7.13. Turning to Condition (2) in that theorem, one may use the relaxed version from Remark 7.14, and show that

$$\mathbb{E} \sup_{u \in (H-H) \cap \rho D} \left| \sum_{i=1}^N \varepsilon_i u(X_i) \right| \lesssim \kappa \delta N.$$

To that end, observe that each $u \in (H-H) \cap \rho D$ is of the form $\langle z, \cdot \rangle$ for $z \in (T-T) \cap \rho B_2^d \subset \rho B_2^d$, and therefore it suffices to show that

$$\mathbb{E} \sup_{z \in \rho B_2^d} \left| \sum_{i=1}^N \varepsilon_i \langle X_i, z \rangle \right| \lesssim \kappa \delta N \sim \kappa d^\alpha N^{1-\alpha}.$$

But

$$\mathbb{E} \sup_{z \in \rho B_2^d} \left| \sum_{i=1}^N \varepsilon_i \langle X_i, z \rangle \right| = \rho \mathbb{E} \left\| \sum_{i=1}^N \varepsilon_i X_i \right\|_2 \leq \rho \sqrt{Nd},$$

and all that is left is to verify that $\rho\sqrt{Nd} \lesssim \kappa N(d/N)^\alpha$, i.e. that $\exp(-c(N/d)^{1-\alpha}) \lesssim \kappa(N/d)^{-\alpha+1/2}$, which is always the case if $N \geq c(\kappa, \alpha)d$.

Therefore, by Theorem 7.13, with probability at least $1 - 2\exp(-cN^{1-\alpha}d^\alpha$, for every $z \in T$,

$$|\{i : |\langle X_i, z \rangle| \geq \kappa/2\}| \gtrsim \delta N \sim N^{1-\alpha}d^\alpha,$$

as claimed. ∎

Chapter 9

Mean estimation

Estimating the mean of a random variable or a random vector is a very important question in modern high-dimensional statistics. In fact, the goal of these notes—identifying the optimal accuracy/confidence tradeoff, is a problem of that flavour: one has to estimate from data the value $\mathbb{E}(f(X)-Y)^2-\mathbb{E}(h(X)-Y)^2$ for every pair (h,f) where one of which is the best in the class. As a result, successful mean estimation has to be carried out uniformly over all such pairs.

As we explain in this chapter, the solution of the simplest problem—mean estimation of a random variable—leads to the right path for the solution of that general problem. But before heading-off in that direction, let us digress and describe a notion from probability in Banach spaces that will prove to be instructive.

9.1 Strong-Weak inequalities and mean estimation

Let Z be a centred random vector in $\mathbb{R}^d$, and one would like to study $(\mathbb{E}\|Z\|^p)^{1/p}$ where $\|\ \|$ is a fixed, arbitrary norm on $\mathbb{R}^d$. Clearly, $(\mathbb{E}\|Z\|^p)^{1/p}$ is bounded from below by two terms:

- Since the L_p norm dominates the L_1 norm, it is evident that

$$(\mathbb{E}\|Z\|^p)^{1/p} \geq \mathbb{E}\|Z\|,$$

  and the latter is called the "strong norm" of Z.

- If B_{X^*} is the unit ball in the dual space to $(\mathbb{R}^d,\|\ \|)$, then for any $x\in\mathbb{R}^d$, $\|x\| = \sup_{x^*\in B_{X^*}}|x^*(x)|$. Therefore,

$$\mathbb{E}\|Z\|^p = \mathbb{E}\sup_{x^*\in B_{X^*}}|x^*(x)|^p \geq \sup_{x^*\in B_{X^*}}\mathbb{E}|x^*(x)|^p,$$

  implying that

$$(\mathbb{E}\|Z\|^p)^{1/p} \geq \sup_{x^*\in B_{X^*}}\|x^*(Z)\|_{L_p}.$$

  The term $\sup_{x^*\in B_{X^*}}\|x^*(Z)\|_{L_p}$ is called the weak L_p norm of Z.

As a result, it follows that for any random vector and any norm, one has

$$(\mathbb{E}\|Z\|^p)^{1/p} \gtrsim \mathbb{E}\|Z\| + \sup_{x^*\in B_{X^*}}\|x^*(Z)\|_{L_p}.$$

The notion of a strong-weak inequality we use here is the reverse inequality: i.e., that there is an absolute constant C such that for any norm on $\mathbb{R}^d$ and every $p \geq 1$,

$$(\mathbb{E}\|Z\|^p)^{1/p} \leq C \left(\mathbb{E}\|Z\| + \sup_{x^* \in B_{X^*}} \|x^*(Z)\|_{L_p} \right). \tag{9.1}$$

Remark 9.1. *It should be stressed that there are similar, more subtle notions of a strong-weak inequalities which arise from questions regarding the concentration of $\|Z\|$ around its mean (see, e.g., [?]). One important aspect is to explore when the constant in front of $\mathbb{E}\|Z\|$ is* 1.

An inequality like (9.1) is a very strong feature of a random vector, because it implies that the moment growth of $\|Z\|$, for any norm, depends only on the moment growth of the 'worst' one dimensional marginal of Z, rather than on other, "global" properties of Z (e.g., the behaviour of high dimensional marginals).

Among the examples of random vectors that satisfy a strong-weak inequality are gaussian vectors, which follows from the gaussian concentration theorem.

Theorem 9.2. *There exists an absolute constant C for which the following holds. Let Z be a centred gaussian random vector in $\mathbb{R}^d$. Then, for any norm $\| \ \|$ and any $1 \leq p < \infty$ one has*

$$(\mathbb{E}\|Z\|^p)^{1/p} \leq \mathbb{E}\|Z\| + C \sup_{x^* \in B_{X^*}} \left(\mathbb{E}\|x^*(Z)\|_{L_2}^p \right)^{1/p}.$$

Exercise 23. *Prove Theorem 9.2.*

Let us return to the mean estimation problem, in which X is an unknown random vector in $\mathbb{R}^d$ that has a finite mean and covariance. One would like to identify $\mu = \mathbb{E}X$, and to do so when the given data consists of N independent copies of X, $X_1, ..., X_N$. In other words, given an arbitrary norm $\| \ \|$ on $\mathbb{R}^d$, one has to find a procedure $\hat{\mu}$, such that, with high probability, $\|\hat{\mu} - \mu\|$ is as small as possible.

Question 9.3. *What is the best possible accuracy/confidence tradeoff in a mean-estimation problem with respect to a general norm, and what procedure attains that tradeoff?*

To get a feeling of what is possible, let us consider a gaussian random vector X. Set $\bar{X} = X - \mu$, and consider the simplest possible procedure: choose $\hat{\mu}$ to be the empirical mean

$$\frac{1}{N} \sum_{i=1}^{N} X_i.$$

Set $Z_N = \frac{1}{N} \sum_{i=1}^{N} X_i$ and observe that by the strong-weak inequality for gaussian vectors,

$$(\mathbb{E}\|Z_N - \mu\|^p)^{1/p} \leq C \left(\mathbb{E}\|Z - \mu\| + \sup_{x^* \in B_X^*} (\mathbb{E}|x^*(Z_N - \mu)|^p)^{1/p} \right).$$

Recall that a centred gaussian random vector W satisfies that for any linear functional x^*,

$$(\mathbb{E}|x^*(W)|^p)^{1/p} \sim \sqrt{p}(\mathbb{E}|x^*(W)|^2)^{1/2}.$$

Hence,

$$(\mathbb{E}\|Z_N - \mu\|^p)^{1/p} \leq C\left(\mathbb{E}\|Z-\mu\| + \sqrt{p}\sup_{x^* \in B_X^*}\left(\mathbb{E}|x^*(Z_N-\mu)|^2\right)^{1/2}\right),$$

and clearly,

$$\left(\mathbb{E}|x^*(Z_N-\mu)|^2\right)^{1/2} = \sqrt{\frac{\mathbb{E}|x^*(\bar{X})|^2}{N}}.$$

Moreover, by a symmetrization argument,

$$\mathbb{E}\|Z-\mu\| = \mathbb{E}\sup_{x^* \in B_{X^*}}\left|\frac{1}{N}\sum_{i=1}^N x^*(X_i) - \mathbb{E}x^*(X)\right| \lesssim \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^N \varepsilon_i \bar{X}_i\right\|,$$

and selecting $p \sim \log(1/\delta)$, by Chebyshev's inequality, with probability at least $1-\delta$,

$$\|\hat{\mu}-\mu\| \leq C\left(\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^N \varepsilon_i \bar{X}_i\right\| + \sup_{x^* \in B_{X^*}}(\mathbb{E}|x^*(\bar{X})|^2)^{1/2}\cdot\sqrt{\frac{\log(1/\delta)}{N}}\right). \tag{9.2}$$

Remark 9.4. *Since X is gaussian, one can simplify* (9.2) *further, because $\sum_{i=1}^N \varepsilon_i \bar{X}_i$ has the same distribution as $\bar{X}/\sqrt{N}$. However, unlike all the other properties that have been use to establish* (9.2) *(i.e., strong-weak inequality and subgaussian tail estimates for one-dimensional marginals), this property is rather special and it makes no sense to invoke it when trying to 'guess' the right estimate that should hold in the general case.*

Taking (9.2) as a guide, one might risk making a wild conjecture:

> Given $X_1, ..., X_N$ which are iid copies of a random vector X that has a finite mean and covariance, and for an arbitrary norm $\|\ \|$ on $\mathbb{R}^d$, there is a procedure $\hat{\mu}$ such that with probability $1-\delta$
>
> $$\|\hat{\mu}-\mu\| \leq C\left(\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^N \varepsilon_i \bar{X}_i\right\| + \sup_{x^* \in B_{X^*}}(\mathbb{E}|x^*(\bar{X})|^2)^{1/2}\cdot\sqrt{\frac{\log(1/\delta)}{N}}\right), \tag{9.3}$$
>
> and where C is an absolute constant.

To clarify why this is indeed a wild conjecture, note that general random vectors do not satisfy a strong-weak inequality and do not exhibit a subgaussian tail decay of marginals. Moreover, it is quite clear that selecting the empirical mean is a bad choice, as it fails even in one-dimensional problems (see more details below). Hence, one must come up with an alternative to the empirical mean if there is to be any hope for the conjecture to be true.

In what follows we show that the conjecture is indeed true for the ℓ_2 norm, and explain why it is almost true (with one minor gap) for a general norm. The first step is to understand how to address the mean estimation problem in dimension one—for real valued random variables, as that turns out to be useful when dealing with the "weak term" in (9.3). We then study the general case, and explain how the global structures of X interacts with the norm.

As one may expect, the proofs will be based on the small-ball method, specifically, on Theorem 7.13.

9.2 real-valued mean estimation

Let us recall the optimal mean estimation estimate one is looking for in the real-valued case: if $X_1, ..., X_N$ are independent copies of an unknown random variable whose mean is μ, the goal is to find $\hat{\mu}$ such that with probability at least $1 - \delta$ satisfies that

$$|\hat{\mu} - \mu| \leq C\sigma\sqrt{\frac{\log 1/\delta}{N}},$$

which is the best estimate one can hope for, and which, for a gaussian random variable, is attained by the empirical mean.

Exercise 24. *Show that indeed this is the best one can hope for, and that for the gaussian random variable it is attained by the empirical mean.*

The starting point in the design of an optimal mean estimation procedure is the basic argument behind the small-ball method: that $Pr(\{Z \text{ satisfies } P\})$ is reflected, with very high probability over samples, by $|\{j : \ Z_j \text{ satisfies } P\}|$. With that realization, all that remains is to identify the right choice of Z and P. Here, P is rather obvious: let

$$Z = \frac{1}{m}\sum_{i=1}^{m} X_i$$

for the right choice of m that is specified in what follows. One has that

$$|Z - \mu| \leq c_0\frac{\sigma_X}{\sqrt{m}} \tag{9.4}$$

with probability at least 0.99. If $n = N/m$ and $Z_1, ..., Z_n$ are independent copies of Z, then with probability at least $1 - 2\exp(-c_1 n)$, there at least $0.98n$ of the $Z_1, ..., Z_n$ that satisfy (9.4), implying that the same holds for a median $\hat{\mu}$ of the values $\{Z_1, ..., Z_n\}$; i.e., if $\hat{\mu}$ is any such median then

$$|\hat{\mu} - \mu| \leq c_0\frac{\sigma_X}{\sqrt{m}} \quad \text{with probability at least } 1 - 2\exp(-c_1 n).$$

Now one has to specify the values of n and m: since the goal is to have an estimator that performs with confidence $1 - \delta$; one should choose $n \sim \log(2/\delta)$; and since $m = N/n$,

$$|\hat{\mu} - \mu| \leq c_0\frac{\sigma_X}{\sqrt{m}} = c_0\sigma_X\sqrt{\frac{n}{N}} \sim \sigma_X\sqrt{\frac{\log(2/\delta)}{N}},$$

as required.

This idea of choosing $\hat{\mu}$, which is a *median-of-means* of $X_1, ..., X_N$ with a specific choice of block size, is used extensively in what follows. Intuitively, taking an empirical mean has a smoothing effect on X, in the sense that with constant probability, the empirical mean is close to the true one. The 'majority vote' is used to increase the probability of success from constant to exponential in the number of blocks; thus, the wanted probability dictates the number of blocks, and therefore the size of each block.

Exercise 25. *Give an example showing that the best performance of the empirical mean is one can hope for is* $\sigma\text{poly}(1/\delta)/\sqrt{N}$*, and that* $\text{poly}(1/\delta)$ *can be arbitrarily bad.*

9.3 Vector mean estimation

Now that the one-dimensional case is resolved, let us turn to the much harder problem of estimating the mean of a random vector. Let $(\mathbb{R}^d, \|\ \|)$ be a normed space and set $\mathcal{B}^\circ$ to be the unit ball of its dual space. As a starting point, suppose that one can perform mean estimation uniformly—for all the one dimensional marginals $\{x^*(X) : x^* \in \mathcal{B}^\circ\}$. Here, uniformly means two things: firstly, a high probability event on which the estimate holds for all the marginals; and secondly, that all the individual means are generated by a single point in $\mathbb{R}^d$. Formally, one is looking for some $z \in \mathbb{R}^d$ such that, with high probability,

$$|x^*(z) - \mathbb{E}x^*(X)| \leq A \quad \text{for every } x^* \in \mathcal{B}^\circ. \tag{9.5}$$

In that case, by the linearity of the expectation, $\sup_{x^* \in \mathcal{B}^\circ} |x^*(z - \mathbb{E}X)| \leq A$, implying that $\|z - \mathbb{E}X\| \leq A$, as required.

The first part of the question will be resolved using the small-ball method, and the second part is the real obstacle: how to generate a point z as in (9.5) using information on one-dimensional marginals. To that end, note that if, for a fixed $x^* \in \mathcal{B}^\circ$, one finds a number α_{x^*} such that $|\mathbb{E}x^*(X) - \alpha_{x^*}| \leq A$, that defined a slab in $\mathbb{R}^d$:

$$\mathcal{S}_{x^*} = \{v \in \mathbb{R}^d : |x^*(v) - \alpha_{x^*}| \leq A\},$$

and if α_{x^*} is indeed a good mean estimator for the marginal $x^*(X)$, then the true mean $\mu \in \mathcal{S}_{x^*}$ belongs to the slab $\mathcal{S}_{x^*}$, implying that $\mathcal{S}_{x^*}$ is nonempty. Hence, if one can find a high probability event on which (9.5) holds for every $x^* \in B^\circ$, then $\bigcap_{x^* \in B_{X^*}} \mathcal{S}_{x^*}$ is nonempty—it contains, at the very least, the true mean μ. Moreover, if z is an arbitrary point in the intersection, one has that

$$\|z - \mu\| = \sup_{x^* \in \mathcal{B}^\circ} |x^*(z) - x^*(\mu)| \leq \sup_{x^* \in \mathcal{B}^\circ} (|x^*(z) - \alpha_{x^*}| + |x^*(\mu) - \alpha_{x^*}| \leq 2A.$$

Therefore, a uniform solution to the mean estimation problem of the marginals almost gets us to where we want to be. The only problem is that there is not enough information on X to "guess" the right value of A. Thankfully, that is an issue that is easily resolved.

The mean estimation procedure explored here as follows:

- Set $\varepsilon > 0$.
- Let $n = \log(2/\delta)$ and split the sample $(X_i)_{i=1}^N$ to n blocks I_j, each of cardinality N/n. Set $Z_j = \frac{1}{m} \sum_{i \in I_j} X_i$.
- For every $x^* \in \mathcal{B}^\circ$ set

$$\mathcal{S}_{x^*} = \left\{ v \in \mathbb{R}^d : |x^*(Z_j) - x^*(v)| \leq \varepsilon \text{ for more than } \frac{n}{2} \text{ blocks} \right\}. \tag{9.6}$$

- Set $\mathbb{S}(\varepsilon) = \bigcap_{x^* \in \mathcal{B}^\circ} \mathcal{S}_{x^*}$. Let ε_0 be the smallest such that $\mathbb{S}(\varepsilon) \neq \emptyset$ and select $\hat{\mu}$ to be any point in $\mathbb{S} = \bigcap_{\varepsilon > \varepsilon_0} \mathbb{S}(\varepsilon)$.

Thus, the values α_{x^*} are indeed given by the median of means estimators for each marginal $x^*(X)$, which, in turn, are used to define the slabs of width ε.

Exercise 26. *Show that the set $\mathbb{S}$ is not empty.*

The fact that $\mathbb{S}$ is not empty is not enough. Thanks to its monotonicity in ε, it suffices to show that there is a good choice of $\varepsilon = A$ for which $\mathbb{S}(\varepsilon)$ contains μ; then, automatically, any point in $\mathbb{S}$ also belongs to $\mathbb{S}(A)$, and $\|\hat{\mu} - \mu\| \leq 2A$—without the need to know explicitly the value of A.

Question 9.5. *Consider the family of random variables $\{x^*(X) : x^* \in \mathcal{B}^\circ\}$. Given $X_1, ..., X_N$ and $Z_1, ..., Z_n$ as above, find A such that with probability at least $1 - \delta$*

$$\sup_{x^* \in \mathcal{B}^\circ} |\mathrm{Med}(x^*(Z_j)) - x^*(\mu)| \leq A.$$

In other words, show that with probability at least $1 - \delta$, for every $x^* \in \mathcal{B}^\circ$,

$$|\{j : |x^*(Z_j) - \mathbb{E}x^*(Z)| \leq A\}| > \frac{n}{2}. \tag{9.7}$$

Clearly, (9.7) fits perfectly the set up of Theorem 7.15 for the class of functions $H = \{x^*(\cdot) : x^* \in \mathcal{B}^\circ\}$ and for the n independent copies of the centred random vectors $\bar{Z}_1, ..., \bar{Z}_n$, where $n \sim \log(2/\delta)$.

Let Λ to be named later, and set κ to be

$$\kappa = \Lambda + c\sqrt{\frac{\log(2/\delta)}{N}} \cdot \sup_{x^* \in \mathcal{B}^\circ} \left(\mathbb{E}(x^*(\bar{X}))^2\right)^{1/2},$$

for a suitable absolute constant c.

It follows that with this choice of κ, $Pr(|Z| \leq \kappa) \geq 0.99$, by invoking Chebyshev's inequality and since

$$\|x^*(\bar{Z})\|_{L_2} = \frac{\|x^*(\bar{X})\|_{L_2}}{\sqrt{m}} \lesssim \sup_{x^* \in \mathcal{B}^\circ} \|x^*(\bar{X})\|_{L_2} \sqrt{\frac{\log(2/\delta)}{N}}.$$

At the same time, using the notation of Theorem 7.15 and since H is a convex, centrally-symmetric set, one has that $(H - H) \cap \rho D \subset 2H$ and

$$\mathbb{E} \sup_{u \in 2H} \left| \frac{1}{n} \sum_{j=1}^{n} \varepsilon_i u(\bar{Z}_j) \right| = \frac{2}{n} \mathbb{E} \sup_{x^* \in \mathcal{B}^\circ} \left| x^* \left(\sum_{i=1}^{N} \varepsilon_i \bar{Z}_j \right) \right| = \frac{2}{n} \mathbb{E} \left\| \sum_{j=1}^{N} \varepsilon_j \bar{Z}_j \right\|.$$

Recall that $\bar{Z}_j = \frac{1}{m} \sum_{i \in I_j} X_i$, and thus,

$$\sum_{j=1}^{N} \varepsilon_i \bar{Z}_j = \frac{1}{m} \sum_{j=1}^{n} \varepsilon_j \sum_{i \in I_j} \bar{X}_i,$$

and by a standard symmetrization argument,

$$\mathbb{E} \left\| \sum_{j=1}^{n} \varepsilon_j \sum_{i \in I_j} \bar{X}_i \right\| \leq 4\mathbb{E} \left\| \sum_{i=1}^{N} \varepsilon_i \bar{X}_i \right\|. \tag{9.8}$$

Hence, it suffices to ensure that

$$\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i\bar{X}_i\right\| \lesssim \kappa,$$

which is the case as long as one sets

$$\Lambda \geq \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i\bar{X}_i\right\|,$$

and κ is that the natural "strong term" in a strong-weak inequality.

To complete the proof, one must estimate the covering numbers of H with respect to the $L_2(\bar{Z})$ norm.

Exercise 27. (1) *Prove* (9.8).

(2) *Complete the proof of the theorem, first when* $\|\ \|$ *is Euclidean norm and then for a general norm (Hint: Sudakov type inequalities may be helpful here).*

Part IV

Complexity measures of sets

Chapter 10

Introduction

What is the right way of measuring the size of a subset of $\mathbb{R}^d$? Obviously, there is more than one answer to this question, as it depends a great deal on the context in which it is asked. Here we study three seemingly different such notions: the Bernoulli mean-width, the ℓ_2 metric entropy and the combinatorial dimension. Understanding their roles has been of central importance to the development of statistical learning theory.

A crucial facet of the questions explored here is, once again, *structure*. Most notably, whether the fact that a set is large is exhibited by the appearance of some canonical structure that is "hidden" in the set. As it happens, that is a general phenomenon which seems, at first, rather surprising:

> The fact that a set is both "well bounded" and extremal in the right sense is exhibited by structure.

This statement, albeit vague, is of extreme importance. Let us mention two examples of this phenomenon, both originating from asymptotic geometric analysis.

The Milman-Dvoretzky Theorem

Let $K \subset B_2^d$ be a convex body, making the notion of "bounded" clear in this case. So what would be considered extremal? One option is the set's mean-width—say, relative to directions in the Euclidean unit sphere S^{d-1}. Obviously, the width of K in direction $\theta \in S^{d-1}$, is $\sup_{x\in K}\langle x, \theta\rangle \leq 1$, because $K \subset B_2^d$. Therefore, the mean width cannot be bigger than 1, and a reasonable choice of "extremal" is a constant lower bound on the mean width:

$$\int_{S^{d-1}} \sup_{x\in K} \langle x, \theta\rangle d\sigma(\theta) \geq \delta,$$

for a fixed parameter δ independent of K or of d. If the idea that "bounded + extremal implies structure" is to be believed, K must be hiding some structure; and in this case an optimistic guess is that K is hiding a large Euclidean ball (of radius $\gtrsim \delta$).

Note that K may be very far from a Euclidean ball; for example, the normalized cube, $d^{-1/2}B_\infty^d$ satisfies both conditions, but does not look anything like a Euclidean ball[1].

[1] "Does not look anything like" can be made accurate, using for example, the notion of the Banach-Mazur distance between convex bodies, see, e.g., [**?**].

Exercise 28. *Show that indeed*

$$\int_{S^{d-1}} \sup_{x \in d^{-1/2} B_\infty^d} \langle x, \theta \rangle d\sigma(\theta) \geq c$$

for a suitable absolute constant c that is independent of the dimension d.

Hint: *Use the gaussian representation of the Haar measure on* S^{d-1}.

The outcome of Milman's version of Dvoretzky's Theorem (see, e.g., [**?**]) is that a random orthogonal projection of K onto a subspace of dimension $\ell = d\delta^2$ will satisfies that

$$c_1 \delta B_2^\ell \subset PK.$$

In other words, being both bounded and extremal is exhibited by a typical orthogonal projection of K of dimension ℓ (i.e., projecting K onto an element of the Grassmann manifold $\mathcal{G}_{\ell,d}$ selected according to the Haar measure) contains a Euclidean ball whose diameter is at least the guarantee one has on the mean-width of K.

We shall not formulate Milman's version of Dvoretzky's Theorem here, only mention that it actually implies a two-sided estimate: for a typical orthogonal projection P of the 'right dimension', PK is very close to a ball whose radius is very close to the mean-width of K. And the dimension in which the structure is exposed is the effective dimension of K: the algebraic dimension multiplied by $(M/R)^2$, where M is the mean-width and R is the Euclidean radius. We strongly encourage the reader to explore the beautiful theory behind this result and other result of its kind (e.g., the books [**?**] are an excellent start).

Sign-embeddings of ℓ_1^d

Let $X = (\mathbb{R}^d, \| \ \|)$ be a norm and denote by B_X its unit ball. Let $x_1, ..., x_d \in B_X$, (which captures the notion of being bounded in this case. Note that by the triangle inequality, $\mathbb{E}\| \sum_{i=1}^d \varepsilon_i x_i \| \leq \sum_{i=1}^d \|x_i\| \leq d$; therefore, a possible notion of "being extremal" is that

$$\mathbb{E} \left\| \sum_{i=1}^d \varepsilon_i x_i \right\| \geq \delta d$$

for some $0 < \delta < 1$.

One trivial example of a norm and vectors that are both bounded and extremal in this sense is the ℓ_1^d norm and the unit basis vectors $e_1, ..., e_d$. Therefore, a natural question is whether the fact that the normed space $(\mathbb{R}^d, \| \ \|)$ has $\{x_1, ..., x_d\}$ that are bounded and extremal implies that it is 'hiding' an subspace, spanned by a large subset of $x_1, ..., x_d$ that looks like ℓ_1. More accurately,

Question 10.1. *Are there constants* $c_1(\delta)$ *and* $c_2(\delta)$ *and a subset* $I \subset \{1, ..., d\}$ *such than* $|I| \geq c_1(\delta)d$, *and for every* $v \in \mathbb{R}^I$,

$$c_2(\delta) \sum_{i \in I} |v_i| \leq \left\| \sum_{i \in I} v_i x_i \right\| \leq \sum_{i \in I} |v_i| ? \tag{10.1}$$

The answer to this question is "yes", with $c_1(\delta) \sim \delta^2$ and $c_2(\delta) \sim \delta$, which is the best one can hope for. This problem has been addressed by several authors, most notably Elton [**?**], Pajor [**?**] and Talagrand [**?**]; it was solved in [**?**].

The reader might be disappointed by the formulation of Question 10.1, because the appearance of structure seems a little vague and unrelated to statistical learning theory. However, all one has to do is to look at the formulation of the dual result to see a canonical structure. Indeed, another way of writing (10.1) is that the coordinate projection

$$P_I B_{X^*} = \left\{ (x^*(x_i))_{i \in I} : x^* \in B_{X^*} \right\} \tag{10.2}$$

satisfies that

$$c_2(\delta) B_\infty^I \subset P_I B_{X^*} \subset B_\infty^I,$$

where as always, B_{X^*} is the unit ball in the dual space. Thus, a high dimensional coordinate projection of B_{X^*} contains the promised structure: a large cube.

Exercise 29. *Prove the dual formulation of Question 10.1.*

Remark 10.2. *Note that $P_I B_{X^*}$ has a familiar structure: if $F = \{x^*(\cdot) : x^* \in B_{X^*}\}$ then F consists of linear functions on $\mathbb{R}^d$. Setting $\sigma = \{x_i : i \in I\}$ and using the notation of the previous chapters, $P_\sigma F$ is precisely* (10.2)*. As it happens, the answer to Question 10.1 is actually far more general, and deals with coordinate projections of function classes; it has nothing to do with the linearity of the functions involved. Also, at it will become clear, the general result will prove to be extremely useful in the context of statistical learning theory.*

Let us now turn to some preliminary facts what is arguably the most natural of the complexity parameter we focus on—the *metric entropy* (or covering numbers) of a set.

Definition 10.3. *Let (T, d) be a metric space. The ε-covering number of T is the minimal number of open balls of radius ε (with respect to the metric) needed to cover T. The ε-packing number of T is the largest cardinality of a subset of T that is ε-separated; i.e., for every x_i, x_j, $d(x_i, x_j) \geq \varepsilon$.*

In what follows, covering numbers are denoted by $\mathcal{N}(T, \varepsilon)$ and the packing numbers are denoted by $\mathcal{M}(T, \varepsilon)$—where in both cases, for the sake of simplicity the metric is omitted. If T is a subset of a normed space and B is the unit ball is the space, the covering numbers are denoted by $\mathcal{N}(T, \varepsilon B)$ and the packing numbers are denoted by $\mathcal{M}(T, \varepsilon B)$.

The following exercise shows that the notions of covering numbers and packing numbers are interchangeable if one has the flexibility of a constant factor in the scale.

Exercise 30. *Show that for every $\varepsilon > 0$,*

$$\mathcal{N}(T, \varepsilon) \leq \mathcal{M}(T, \varepsilon) \leq \mathcal{N}(T, \varepsilon/2).$$

A comprehensive exposition on covering/packing numbers can easily occupy an entire volume. As this is just a 'tasting', let us present to relatively simple yet very useful facts on covering numbers of subsets of a normed space, beginning with estimates that are based on volumetric arguments.

10.1 Volume based estimates

The link between various notions of volume and the cardinality of a separated set is rather natural: if a set T contains an ε-separated set $\{t_1, ..., t_k\}$, then by the triangle inequality, the open balls $B(t_1, \varepsilon/2), ..., B(t_k, \varepsilon/2)$ are disjoint. Therefore, if T is a subset of a normed space and B is the unit ball in that space then the sets $t_i + (\varepsilon/2)B$ are disjoint. On the other hand, all these sets are contained in

$$T + (\varepsilon/2)B = \{x : x = t + v,\ t \in T,\ \ \|v\| \le \varepsilon/2\}.$$

Hence,

$$\bigcup_{i=1}^{k} (t_i + (\varepsilon/2)B) \subset T + (\varepsilon/2)B, \tag{10.3}$$

and it is a union of disjoint sets. If μ is a measure on the space, one has

$$\mu\left(\bigcup_{i=1}^{k} (t_i + (\varepsilon/2)B)\right) = \sum_{i=1}^{k} \mu\left(t_i + (\varepsilon/2)B.\right),$$

but at the same time, by (10.3),

$$\mu\left(\bigcup_{i=1}^{k} (t_i + (\varepsilon/2)B)\right) \le \mu\left(T + (\varepsilon/2)B\right).$$

Combining these two simple observations,

$$\sum_{i=1}^{k} \mu\left(t_i + (\varepsilon/2)B\right) \le \mu\left(T + (\varepsilon/2)B\right). \tag{10.4}$$

The simplest example in which (10.4) leads to an interesting bound is when $T \subset \mathbb{R}^d$, B is the unit ball of a norm in $\mathbb{R}^d$ and μ is the volume (Lebesgue) measure in $\mathbb{R}^d$.

10.1.1 The Lebesgue measure

With a minor abuse of notation, let $|\ |$ denote the volume measure on $\mathbb{R}^d$, and clearly it is shift-invariant. Therefore,

$$\sum_{i=1}^{k} |t_i + (\varepsilon/2)B| = k \cdot (\varepsilon/2)^d |B|,$$

implying that

$$k \le \left(\frac{2}{\varepsilon}\right)^d \cdot \frac{|T + (\varepsilon/2)B|}{|B|}. \tag{10.5}$$

For $T = B$, by (10.5)

$$\mathcal{M}(B, \varepsilon B) \le \left(\frac{2}{\varepsilon}\right)^d \cdot (1 + \varepsilon/2)^d \le \left(\frac{5}{\varepsilon}\right)^d. \tag{10.6}$$

Remark 10.4. *It is clear from* (10.5) *that obtaining an upper estimate on the volume* $|T+\varepsilon B|$ *leads to an upper bound on* $\mathcal{M}(T, 2\varepsilon B)$. *Bounding* $|T + \varepsilon B|$ *for convex sets* T *and* B *is one of the classical questions in convex geometry, especially when* $B = B_2^d$ *(see, for example, the books [?] for the beautiful theory behind such estimates).*

It turns out that up to the constants involved, and as long as ε is sufficiently smaller than 1, say, $0 < \varepsilon \leq 1/2$, (10.6) is sharp. Indeed, if T is covered by k shifts of B, then since a measure is sub-additive,

$$\mu(T) \leq \mu\left(\bigcup_{i=1}^{k}(t_i + \varepsilon B)\right) \leq \sum_{i=1}^{k} \mu(t_i + \varepsilon B).$$

Hence, in the case of the volume measure, one has

$$|T| \leq k\varepsilon^d |B|,$$

and for $T = B$,

$$\mathcal{N}(B, \varepsilon B) \geq \left(\frac{1}{\varepsilon}\right)^d.$$

Corollary 10.5. *If* $B \subset \mathbb{R}^d$ *is a convex body then for any* $0 < \varepsilon < 1/2$,

$$\left(\frac{1}{\varepsilon}\right)^d \leq \mathcal{N}(B, \varepsilon B) \leq \left(\frac{5}{\varepsilon}\right)^d$$

Exercise 31. *Show that a set* B *as in Corollary 10.5 is a unit ball of a norm, and deduce the corollary from that.*

Let us turn to another application of the volumetric estimate, though this time, a more subtle one—with respect to a different measure which is not shift invariant.

10.1.2 The gaussian measure

Recall that for $T \subset \mathbb{R}^d$ the polar of T is the set

$$T^\circ = \left\{x \ : \ |\langle x, t\rangle| \leq 1 \ \ \forall t \in T\right\},$$

set $\|x\|_{T^\circ} = \sup_{t\in T} |\langle x, t\rangle|$, and let $\ell_*(T)$ be the gaussian mean width of T, that is,

$$\ell_*(T) = \mathbb{E} \sup_{t\in T} |\langle G, t\rangle| = \mathbb{E}\|G\|_{T^\circ},$$

where here, as always, G is the standard gaussian vector in $\mathbb{R}^d$.

Theorem 10.6. *There exist an absolute constant* c *such that, for any* $T \subset \mathbb{R}^d$ *and* $\varepsilon > 0$,

$$c\varepsilon \log^{1/2} N(B_2^d, \varepsilon T^\circ) \leq \ell_*(T).$$

For reasons that will become clear later, Theorem 10.6 is called the *dual Sudakov inequality*—and it is an extremely useful fact. The proof presented here is due to Talagrand [?].

Proof. Set μ_G to be the standard gaussian measure on $\mathbb{R}^d$. Let $u \geq 2\ell_*(T) = 2\mathbb{E}\|G\|_{T^\circ}$ and by Markov's inequality, $\mu_G(\|x\|_{T^\circ} \geq u) \leq 1/2$. Therefore,

$$\mu_G(2\ell(T) \cdot T^\circ) \geq 1/2. \tag{10.7}$$

Observe that if $\{x_1, ..., x_k\} \subset B_2^d$ is an ε-separated set with respect to $\| \ \|_{T^\circ}$, then $x_i + (\varepsilon/2)T^\circ$ have disjoint interiors, and the same holds for the sets $\alpha(x_i + (\varepsilon/2)T^\circ)$ for any $\alpha > 0$. Let α satisfy that $\alpha(\varepsilon/2) = 2\ell_*(T)$ and put $y_i = \alpha x_i$. Thus, the sets $y_i + 2\ell_*(T)T^\circ$ have disjoint interiors and

$$\sum_{i=1}^k \mu_G(y_i + 2\ell_*(T)T^\circ) = \mu_G\left(\bigcup_{i=1}^k (y_i + 2\ell_*(T)T^\circ)\right) \leq 1, \tag{10.8}$$

because μ_G is a probability measure.

Now let us estimate each $\mu_G(y_i + 2\ell_*(T)T^\circ)$ from below. For any measurable $A \subset \mathbb{R}^d$,

$$\mu_G(A) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \mathbb{1}_A \exp(-\|x\|_2^2/2)dx.$$

Let $A = y_i + \beta T^\circ$, note that $\mathbb{1}_{y_i+\beta T^\circ}(x) = \mathbb{1}_{\beta T^\circ}(x - y_i)$ and by a change of variables,

$$\begin{aligned}
\mu_G(y_i + \beta T^\circ) =& \frac{1}{(2\pi)^{d/2}} \int_{\beta T^\circ} \exp\left(-\|x - y_i\|_2^2/2\right) \\
=& \exp\left(-\frac{\|y_i\|_2^2}{2}\right) \frac{1}{(2\pi)^{d/2}} \int_{\beta T^\circ} \exp(-\langle x, y_i\rangle) \exp(-\|x\|_2^2/2)dx \\
=& \exp\left(-\frac{\|y_i\|_2^2}{2}\right) \mu_G(\beta T^\circ)\mathbb{E}\exp(-\langle Z, y_i\rangle),
\end{aligned}$$

where Z is the gaussian vector G conditioned on the set βT°.

Since Z is a symmetric random vector (G is symmetric and T° is centrally-symmetric), it is evident that the linear forms $\langle Z, y_i\rangle$ are symmetric random variables. Hence, $\mathbb{E}\langle Z, y_i\rangle = 0$; by Jensen's inequality, $\mathbb{E}\exp(-\langle Z, y_i\rangle) \geq 1$; and for every $\beta > 0$,

$$\mu_G(y_i + \beta T^\circ) \geq \exp\left(-\frac{\|y_i\|_2^2}{2}\right) \mu_G(\beta T^\circ).$$

Set $\beta = 2\ell_*(T)$, and by (10.7), $\mu_G(\beta T^\circ) \geq 1/2$. Moreover, $y_i = \alpha x_i$, and thus $\|y_i\|_2 \leq 4\ell_*(T)/\varepsilon$. Therefore,

$$\sum_{i=1}^k \mu_G(y_i + \beta T^\circ) \geq \frac{k}{2} \exp\left(-8\frac{\ell_*^2(T)}{\varepsilon^2}\right). \tag{10.9}$$

The claim follows by combining (10.8) and (10.9). ∎

10.2 The Maurey Lemma

The second example is due to B. Maurey [**?**]: an estimate the covering numbers of the convex hull of a set of n points in a normed space which has a nontrivial *type*.

The type constant of a normed space is a generalization of the parallelogram equality: if H is an inner product space and $x_1, ..., x_m \in H$ then

$$\mathbb{E}\left\|\sum_{i=1}^m \varepsilon_i x_i\right\|_H^2 = 2\sum_{i=1}^m \|x_i\|_H^2.$$

Definition 10.7. *A normed space X has type $1 < p \leq 2$ with constant $T_p(X)$ if it is the smallest constant for which, for every finite set $x_1, ..., x_m \in X$,*

$$\mathbb{E}\left\|\sum_{i=1}^m \varepsilon_i x_i\right\|_X \leq T_p(X)\left(\sum_{i=1}^m \|x_i\|_X^p\right)^{1/p}.$$

Remark 10.8. *One may show that for any normed space X and any $1 \leq q < \infty$,*

$$\left(\mathbb{E}\left\|\sum_{i=1}^m \varepsilon_i x_i\right\|_X^q\right)^{1/q} \leq C\sqrt{q}\mathbb{E}\|\sum_{i=1}^m \varepsilon_i x_i\|_X.$$

This is a vector-valued version of Khintchine's inequality—the so-called Kahane-Khintchine inequality*, see, e.g., [**?**]. It particular, one may replace the L_1 norm $\mathbb{E}\|\sum_{i=1}^m \varepsilon_i x_i\|_X$ in the definition of type with any L_q norm.*

Remark 10.9. *Observe that every space has type 1 with constant $T_1(X) = 1$.*

There is a very well developed theory describing the structure of normed spaces that have a non-trivial type – specifically, the connection between the type and the ability to embed the spaces ℓ_p^n in X. We refer the reader to [**?**].

Lemma 10.10. *Let X be a Banach space of type p with constant $T_p(X)$. Let $A = \{x_1, ..., x_n\} \subset X$ and set $a = \max \|x\|$. Then for every $\varepsilon > 0$*

$$\mathcal{N}\left(\mathrm{conv}(A), 2aT_p(X)B_X\right) \leq \left(e + en\varepsilon^{\frac{p}{p-1}}\right)^{\frac{p}{p-1}}.$$

Proof. Let $(\lambda_i)_{i=1}^n$ be nonnegative numbers that satisfy $\sum_{i=1}^n \lambda_i = 1$, and for an integer k let $Y_1, ...Y_k$ be independent random variables defined by $Pr(Y = x_j) = \lambda_j$. Note that $\mathbb{E}Y = \sum_{i=1}^n \lambda_i x_i$, and by a symmetrization argument,

$$\mathbb{E}_Y\left\|k^{-1}\sum_{i=1}^k Y_i - \mathbb{E}Y\right\| \leq \frac{2}{k}\mathbb{E}_Y\mathbb{E}_\varepsilon\left\|\sum_{i=1}^k \varepsilon_i Y_i\right\| \leq \frac{2}{k}\mathbb{E}_Y T_p(X)\left(\sum_{i=1}^k \|Y_i\|^p\right)^{\frac{1}{p}} \leq \frac{2aT_p(X)}{k^{1-1/p}}.$$

Hence, there exists a realization of $k^{-1}\sum_{i=1}^k Y_i$ whose distance from $\sum_{i=1}^n \lambda_i x_i$ is smaller than $2aT_p(X)/k^{1-1/p}$. Observe that every realization of Y_i belongs to $\{x_1, ..., x_n\}$, and therefore the number of possible realizations is at most

$$\binom{n+k-1}{k} \leq e^k(1+n/k)^k.$$

Hence,

$$\mathcal{N}\left(\mathrm{conv}(A), \frac{2aT_p}{k^{1-\frac{1}{p}}}B_X\right) \leq e^k\left(1+\frac{n}{k}\right)^k,$$

from which the claim immediately follows. ∎

Exercise 32. *Use Lemma 10.10 to estimate $\mathcal{N}(B_1^d, \varepsilon B_2^d)$. Is this estimate sharp for every $0 < \varepsilon < 1/2$? Is it sharp in some range of $0 < \varepsilon < 1/2$?*

Chapter 11

Generic Chaining

This chapter is devoted to a systematic study of upper estimates on the expectation of the supremum of random process $\{Z_t : t \in T\}$ indexed by sets T. The focus is on the way the structure of the indexing set T is reflected in the estimates on the expectation of the supremum, and for obvious reasons, the processes that interests us the most are gaussian processes, Bernoulli processes and empirical processes.

The analysis is based on Talagrand's *generic chaining* mechanism, and the reader is strongly encouraged to read Talagrand's treasured manuscript [**?**]. It is far more extensive and instructive than the brief outline presented in what follows, and is a must for anyone who wants to truly understand the chaining mechanism.

11.1 The natural metrics

Earlier in these notes we explored some features of random variables. Here, the goal is to study properties of collections of random variables. To be more accurate, let V be a set and assume that one associates to each $v \in V$ a random variable Z_v. One would like to obtain high probability, sharp estimates on

$$\mathbb{E} \sup_{v \in V} Z_v,$$

and in particular understand the way the structure of V is reflected in upper and lower bounds on the expectation.

But what is the meaning of "structure of V"? At this point, V is just an indexing set, with no apparent structure. The main message of this chapter is:

> The random process $v \to Z_v$ endows a sequence of natural metrics on V, and understanding the geometry of V with respect to those metrics plays a crucial role in estimating $\mathbb{E} \sup_{v \in V} Z_v$.

At this point it should be stressed that "natural" may seem at times totally unnatural. It may be the case that V is a metric space with respect to some underlying metric d (e.g., if $V \subset (\mathbb{R}^d, \| \ \|_2)$, but the Euclidean metric has absolutely nothing to do with the metrics endowed on V by the process $\{Z_v : v \in V\}$.

Remark 11.1. *In what follows we completely avoid the question of measurability, even when V is an infinite set. There are various ways of justifying that but the simplest one is to define*

$$\mathbb{E}\sup_{v\in V} Z_v = \sup\left\{\mathbb{E}\max_{v\in V'} Z_v : \ V'\subset V,\ V' \text{ is finite}\right\}.$$

Under some mild assumptions on the process this coincides with the standard notion.

Although this presentation holds for rather general processes, it helps to keep in mind several specific examples:

- Let X be a symmetric random vector in $\mathbb{R}^d$, set $T\subset\mathbb{R}^d$ and define $Z_t = \langle X, t\rangle$. Among the natural processes that belong to this family are when X is a centred gaussian random vector in $\mathbb{R}^d$, leading to a gaussian process indexed by a subset of $\mathbb{R}^d$, and when X is the uniform measure on $\{-1,1\}^d$, leading to a Bernoulli process indexed by a subset of $\mathbb{R}^d$.

- Let $F\subset L_2(\mu)$ be a class of functions and set $X_1,...,X_N$ to be independent, distributed according to μ. The (centred) empirical process indexed by F is given by the choice of

$$Z_f = \frac{1}{N}\sum_{i=1}^N f(X_i) - \mathbb{E}f,$$

and the centred quadratic process is given by

$$Z_f = \frac{1}{N}\sum_{i=1}^N f^2(X_i) - \mathbb{E}f^2;$$

naturally, one can, and often will consider the non-centred versions of the empirical and quadratic processes.

Note that in all these examples, the indexing set is naturally endowed with various metric structures that could be seen as natural, e.g. some ℓ_p metric for $T\subset\mathbb{R}^d$ and the $L_2(\mu)$ metric for F. However, there is no reason to expect that these metrics have any relevance to questions regarding the random processes indexed by the sets; the intuitive metrics are nothing more than distractions one would be wise to ignore.

Let us begin with the question of obtaining sharp upper bounds on $\mathbb{E}\sup_{v\in V} Z_v$; specifically, given $u>0$, one would like to estimate

$$Pr(\sup_{v\in V}|Z_v|>u).$$

Obviously, there are various ways of bounding $Pr(|Z_v|>u)$ individually, ,but the question here is how can those individual estimates be 'combined', leading to the wanted estimate on the supremum. As a first step, assume that V is finite, and consider a very crude estimate: using the union bound

$$Pr\left(\max_{v\in V}|Z_v|>u\right)\leq\sum_{v\in V}Pr(|Z_v|>u). \tag{11.1}$$

While (11.1) can be very costly if performed without thought, it gives information on the smallest u one can take that would still be useful. For example, it suffices that $\max_{v\in V} Pr(|Z_v| > u) \leq 1/|V|$ to ensure a nontrivial estimate. One way of obtaining such a bound is via Chebyshev's inequality, as for any $p \geq 1$ one has

$$Pr(|Z| \geq te\|Z\|_{L_p}) \leq \frac{e^{-p}}{t^{-p}}. \tag{11.2}$$

In particular, one may set $p = \log|V|$ and $u = t\max_{v\in V}\|Z_v\|_{L_p}$ to obtain

$$Pr\left(\max_{v\in V}|Z_v\| \geq t\max_{v\in V}\|Z_v\|_{L_p}\right) \leq t^{-\log|V|}. \tag{11.3}$$

The almost trivial combination of (11.2) and (11.3) is of crucial importance: it implies that one may use the L_p norm to obtain a probability estimate that is better than $\exp(-p)$. In return, that allows one to control $\exp(p)$ random variables simultaneously.

The estimate in (11.3) is indeed very crude. It is based on the belief that the events $\Omega_v = \{|Z_v| > u\}$ are essentially disjoint, and at the same time that they are roughly of the same 'size'. Thus, the bound follows by setting u large enough to ensure that the largest of the sets Ω_v is small enough. While there is no reason to expect that either part of these beliefs is true, the idea of using the union bound can be optimized. At the heart of the matter is finding a notion that captures when two random variables are 'close': roughly put, the fact that random variables Z_x and Z_y are close should imply that the events $\{Z_x$ is large$\}$ and $\{Z_y$ is large$\}$ have a large overlap. As a result, there would be no need to take both events into account in the union bound. One way of ensuring that the events in question have a large overlap is that the 'tails' $Pr(|Z_x - Z_y| > u)$ are small enough.

To make this idea more precise, recall the definition of the $\mathbb{E}\sup_{v\in V} Z_v$, as the supremum of such expectations taken over finite sets. Let $V' \subset V$ be a finite set, fix an integer $s_0 \geq 0$ and let us construct a sequence of sets $V_s \subset V'$ for $s \geq s_0$, where the cardinalities $|V_s|$ grow with s, and that for s large enough, $V_s = V'$. The idea is that V_s are increasingly fine approximations of V', in the sense that for each Z_v there is some Z_u such that $u \in V_s$ and $|Z_v - Z_u|$ is likely to be small.

Formally, for every $v \in V'$ let $\pi_s v \in V_s$, and if $v \in V_s$ then $\pi_s v = v$. The collection $(Z_{\pi_s v})_{s\geq s_0}$ is called a *chain*, each $Z_{\pi_{s+1}v} - Z_{\pi_s v}$ is a link in that chain and each Z_v can be represented as the telescopic sum along the chain

$$Z_v = Z_{\pi_{s_0}v} + \sum_{s>s_0}(Z_{\pi_{s+1}v} - Z_{\pi_s v}); \tag{11.4}$$

Since $V_s = V'$ for s large enough, for every v the sum is over a finite set of values s.

Let us return to the 'large overlaps' idea: for Z_v to be large, some of the links in the chain (11.4) must be large. Therefore, one's aim is to find a high probability event such that for every v and every $s \geq s_0$, all the links are relatively small, i.e., $|Z_{\pi_{s+1}v} - Z_{\pi_s v}| \leq \Delta(s, v)$, where we set $\Delta(s_0, v) = |Z_{\pi_{s_0 v}}|$. Indeed, on that event,

$$\max_{v\in V'}|Z_v| \leq \max_{v\in V'}\sum_{s\geq s_0}\Delta(s, v). \tag{11.5}$$

Observe that (11.5) holds on the complement of the event

$$\mathcal{A} = \left\{ \exists s \geq s_0,\ \exists v \in V' \ : |Z_{\pi_{s+1}v} - Z_{\pi_s v}| \geq \Delta(s,v) \right\},$$

and

$$Pr(\mathcal{A}) \leq \sum_{s \geq s_0} Pr\left(\exists v \in V' \ : |Z_{\pi_{s+1}v} - Z_{\pi_s v}| \geq \Delta(s,v) \right). \tag{11.6}$$

For every $s \geq s_0$ the number of possible differences of the form $Z_{\pi_{s+1}v} - Z_{\pi_s v}$ is at most $|V_s| \cdot |V_{s+1}|$. This naturally leads to a growth condition on the allowed cardinality of V_s, namely, $|V_s| \leq 2^{2^s}$, ensuring that $|V_s| \cdot |V_{s+1}| \leq 2^{2^{s+2}}$ which is the restriction on $|V_{s+2}|$. Thus, $\log |V_s| \cdot |V_{s+1}| \leq 2^{s+4}$.

Finally, let us return to the key idea behind the choice of $\Delta(s,v)$: one may use the L_p norm to control $\exp(p)$ random variables uniformly. Hence, for every s set

$$p = c2^s \geq \log |V_s| \cdot |V_{s+1}|$$

and let

$$\Delta(s,v) = et \| Z_{\pi_{s+1}v} - Z_{\pi_s v} \|_{L_p}.$$

Therefore,

$$Pr\left(\exists v \in V' \ : |Z_{\pi_{s+1}v} - Z_{\pi_s v}| \geq et \| Z_{\pi_{s+1}v} - Z_{\pi_s v} \|_{L_{c2^s}} \right) \leq t^{-c2^s}. \tag{11.7}$$

By the union bound over $s \geq s_0$ it follows that with probability at least $1 - t^{-c_1 2^{s_0}}$,

$$\sup_{v \in V'} |Z_v| \leq et \sup_{v \in V'} \left(\| Z_{\pi_{s_0} v} \|_{L_{2^{s_0}+4}} + \sum_{s > s_0} \| Z_{\pi_{s+1}v} - Z_{\pi_s v} \|_{L_{2^{s+4}}} \right).$$

This discussion leads to the following definition:

Definition 11.2. *Given a finite set V, an admissible sequence of V is a collection of subsets $V_s \subset V$ such that $|V_0| = 1$ and for every $s \geq 1$, $|V_s| \leq 2^{2^s}$. For $s_0 \geq 0$ let*

$$\bar{\gamma}(Z_V, s_0) = \inf \sup_{v \in V} \left(\| Z_v \|_{L_{2^{s_0}+4}} + \sum_{s > s_0} \| Z_{\pi_{s+1}v} - Z_{\pi_s v} \|_{L_{2^{s+4}}} \right)$$

where the infimum is taken over all admissible sequences of V.

If V is infinite then set

$$\bar{\gamma}(Z_V, s_0) = \sup \left\{ \bar{\gamma}(Z_{V'}, s_0) : V' \subset V,\ V' \text{ is finite} \right\}.$$

In what follows we denote $\bar{\gamma}(Z_V) = \bar{\gamma}(Z_V, 0)$,

Thus, one has the following *Generic Chaining* upper estimate:

Theorem 11.3. *There exist absolute constants c, c' and c'' for which the following holds. Let $\{Z_v : v \in V\}$ be a random process. Then for $t \geq 1$ and any $s_0 \geq 0$, with probability at least $1 - t^{-cs_0}$,*

$$\sup_{v \in V} |Z_v| \leq c' t \bar{\gamma}(Z_V, s_0).$$

Corollary 11.4. *There exists an absolute constant c_0 such that for every process $\{Z_v : v \in V\}$,*

$$\mathbb{E}\sup_{v\in V}|Z_v| \le c_0\bar{\gamma}(Z_V).$$

Proof. Let c be the constant from Theorem 11.3 and set s_0 to be the smallest such that $c2^{s_0} \ge 2$. By the triangle inequality and since $\|\ \|_{L_p} \le \|\ \|_{L_q}$ for $q \ge p$, it is evident that $\bar{\gamma}(V, s_0) \le \bar{\gamma}(V)$. Hence, Theorem 11.3 implies that for every $t \ge 1$,

$$Pr\left(\sup_{v\in V}|Z_v| \ge t\bar{\gamma}(Z_V)\right) \le \frac{1}{t^2},$$

and integrating the tail,

$$\mathbb{E}\sup_{v\in V}|Z_v| = \int_0^\infty \left(\sup_{v\in V}|Z_v| \ge t\right) \le c_0\bar{\gamma}(Z_V)$$

for a suitable absolute constant c_0. ∎

Exercise 33. *For $p > 1$, estimate $(\mathbb{E}\sup_{v\in V}|Z_v|^p)^{1/p}$.*

Theorem 11.3 leads to a general scheme for upper bounding $\sup_{v\in V}|Z_v|$:

(1) Identify the natural L_p norms associated with the process $v \to Z_v$.

(2) Choose the level s_0 for the starting point of the chaining process, based on the probability estimate one is looking for.

(3) Find an optimal (or almost optimal) admissible sequence of the set V with respect to the L_p metrics endowed by the process.

Out of the three tasks, the third one is, by far, the most difficult one. Finding the right way of constructing approximating sets of V and doing so with respect to metrics that are endowed by the random process is often a real challenge. Moreover, at this point there is no guarantee whatsoever that this chaining scheme leads to an optimal estimate (in fact, the question of whether chaining leads to an optimal estimate is open, except in very special cases).

Before we tackle (3), let us consider (1), which justifies the effort that was invested in exploring the L_p norm of linear forms $\langle X, t\rangle$ for several random vectors X.

The standard gaussian random vector

Let $G = (g_1, ..., g_d)$ be the standard gaussian random vector in $\mathbb{R}^d$. As noted previously, for any $v \in \mathbb{R}^d$,

$$\|\langle G, v\rangle\|_{L_p} \sim \sqrt{p}\|v\|_2.$$

Now fix $V \subset \mathbb{R}^d$ and let $G_v = \langle G, v\rangle$. Observe that by the triangle inequality $\|\pi_{s+1}v - \pi_s v\|_2 \le \|\pi_s v - v\|_2 + \|\pi_{s+1}v - v\|_2$. Thus, for any admissible sequence $(V_s)_{s\ge 0}$ of V and every $v \in V$ one has that

$$\|Z_{\pi_{s+1}v} - Z_{\pi_s v}\|_{L_p} = \|\langle G, \pi_{s+1}v - \pi_s v\rangle\|_{L_p} \sim \sqrt{p}\left(\|\pi_s v - v\|_2 + \|\pi_{s+1}v - v\|_2\right),$$

implying that

$$\bar{\gamma}(G_V, s_0) \sim \inf \sup_{v \in V} \left(2^{s_0/2} \|v\|_2 + \sum_{s \geq s_0} 2^{s/2} \|\pi_s v - v\|_2 \right),$$

where the infimum is taken with respect to all admissible sequences of V. In particular,

$$\mathbb{E} \sup_{v \in V} |\langle G, v \rangle| \leq c_1 \inf \sup_{v \in V} \left(2^{s_0/2} \|v\|_2 + \sum_{s \geq s_0} 2^{s/2} \|\pi_s v - v\|_2 \right). \tag{11.8}$$

The important point to note is that all the metrics involved in the chaining bound for a gaussian process are multiples of a single metric. Indeed, the fact that $\|\langle G, x \rangle\|_{L_p} \sim \sqrt{p} \|\langle G, x \rangle\|_{L_2}$ is a general feature of centred gaussian processes and holds regardless of the covariance of G; however, the fact that $\|\langle G, x \rangle\|_{L_2} = \|x\|_2$ follows because the standard gaussian measure on $\mathbb{R}^d$ is isotropic. If one were to choose a centred gaussian random vector Z whose covariance is not the identity, that would endow a different inner product on $\mathbb{R}^d$ and a different distance—namely $\|x\| = \|\langle Z \rangle\|_{L_2}$. Still, the chaining bound involves multiplies of that single metric and

$$\mathbb{E} \sup_{v \in V} |\langle Z, v \rangle| \leq c_1 \inf \sup_{v \in V} \left(2^{s_0/2} \|\langle Z, v \rangle\|_{L_2} + \sum_{s \geq s_0} 2^{s/2} \|\langle Z, \pi_s v - v \rangle\|_{L_2} \right). \tag{11.9}$$

The fact that in the gaussian case all the metrics 'collapse' to a single one makes the chaining-based estimate on $\mathbb{E} \sup_{v \in V} |\langle Z, v \rangle|$ (at least, potentially) simple: one has to construct an almost optimal admissible sequence with respect to the L_2 norm endowed on $\mathbb{R}^d$ by Z. And even though that norm is Hilbertian, *it has nothing to do with the natural Euclidean norm in $\mathbb{R}^d$ unless Z is isotropic.* That illustrates the comment made earlier—the natural metrics endowed by the process may seem totally unnatural at first glance.

As it happens, the analysis of general gaussian processes indexed by a class of functions $F \subset L_2(\mu)$ is not fundamentally different from the case of gaussian processes indexed by subsets of $\mathbb{R}^d$ (though there are nontrivial technical difficulties in defining the gaussian process and showing that it behaves well, see, e.g. [**?**]). Because the definitions and basic properties of general gaussian processes are a little more involved we will not describe that setup here.

Let us stress once again that at this point there is no reason to expect that (11.8) or (11.9) are sharp in any way. The definition of $\bar{\gamma}$ seems to be artificial: a compact way of presenting the outcome of the chaining mechanism. As it happens, one of the great achievements of Generic Chaining, Talagrand's *majorizing measures theorem* shows that this artificial notion is, in fact, optimal: the expectation of the supremum of a gaussian process, is equivalent to $\bar{\gamma}$:

Theorem 11.5. *There exists absolute constants c and C for which the following holds. Let $F \subset L_2$ and let $\{G_f : f \in F\}$ be the canonical gaussian process indexed by F (i.e.,*

the covariance of the process satisfies $\mathbb{E}G_fG_h = \langle f, h\rangle_{L_2}$*). Then*

$$c\bar{\gamma}(G_F) \le \mathbb{E}\sup_{f\in F} G_f \le C\bar{\gamma}(G_F).$$

Remark 11.6. *An outcome of the majorizing measures theorem is that for an arbitrary set* $T \subset \ell_2$, $\bar{\gamma}(G_T) \sim \bar{\gamma}(G_{\mathrm{conv}(T)})$*. To-date, there is no direct proof of that fact.*

The standard Bernoulli vector

Let $X = (\varepsilon_1, ..., \varepsilon_d)$ be the standard Bernoulli vector, that is, a vector whose coordinates are independent, symmetric $\{-1, 1\}$-valued random variables. Recall that for any $v \in \mathbb{R}^d$,

$$\|\langle \mathcal{E}, v\rangle\|_{L_p} \sim \sum_{i\le p} v_i^* + \sqrt{p}\left(\sum_{i>p}(v_i^*)^2\right)^{1/2}$$

where, as always, (v_i^*) denotes the nonincreasing rearrangement of $(|v_i|)_{i=1}^d$.

Right from the start it is clear that a chaining bound for the supremum of the process $\{\langle \mathcal{E}, v\rangle : v \in V\}$ is far more complex than for its gaussian counterpart. The metrics involved in the chaining bound truly change with s rather than being scaled versions of a single metric. As a result, the construction of an optimal admissible sequence is even more challenging than in the gaussian case.

One easy (yet suboptimal) way forward is to note that if G is the standard gaussian vector in $\mathbb{R}^d$, then for any $p \ge 2$,

$$\|\langle \mathcal{E}, v\rangle\|_{L_p} \le c\|\langle G, v\rangle\|_{L_p}; \tag{11.10}$$

in particular, for any $s_0 \ge 0$, $\bar{\gamma}(\mathcal{E}_V, s_0) \le c\bar{\gamma}(G_V, s_0)$.

Exercise 34. (1) *Show that* $\bar{\gamma}$ *satisfies a Sudakov type inequality: that there is an absolute constant* c *such that for any process* X_V *and* $p \ge 1$*, if* $\{X_v : v \in V\}$ *contains a set of cardinality* $\exp(p)$ *that is* ε*-separated in* L_p *then* $\bar{\gamma}(X_v) \ge c\varepsilon$*.*

(2) *Let* $V = \{e_1, ..., e_d\}$ *be the standard basis in* $\mathbb{R}^d$*. Estimate* $\bar{\gamma}(G_v)$ *and* $\bar{\gamma}(\mathcal{E}_V)$ *from above and below and deduce that the two are not equivalent.*

ψ_α processes

Let $0 < \alpha < 2$. A random process $\{Z_v : v \in V\}$ is called a ψ_α process with constant L if for every $u, v \in V$ and any $p \ge 1$,

$$\|Z_u - Z_v\|_{L_p} \le Lp^{1/\alpha}\|Z_u - Z_v\|_{L_2}, \quad \text{and} \quad \|Z_v\|_{L_p} \le Lp^{1/\alpha}\|Z_v\|_{L_2}$$

Hence, the norm equivalence constant of differences between the L_p norm and the L_2 norm is at most $Lp^{1/\alpha}$.

Example 11.7. *Clearly, both a gaussian process and a Bernoulli process are* ψ_2 *processes with an absolute constant* L*, but all though the estimate* $\|Z_u - Z_v\|_{L_p} \le Lp^{1/\alpha}\|Z_u - Z_v\|_{L_2}$ *is sharp for a gaussian process, it is far from sharp for the Bernoulli process: as noted previously,* $\|\mathcal{E}_{e_i}\|_{L_p} = \|\langle \mathcal{E}, e_i\rangle\|_{L_p} = 1$ *for every* $p \ge 1$*, which is far better than* $\sim \sqrt{p}$*.*

There is a wide variety of ψ_α processes that appear naturally in analysis, geometry and statistics, and the study of such processes is worthy of a detailed exposition in its own right. To give some flavour of the diversity of such processes, a log-concave random vector X on $\mathbb{R}^d$ endows a ψ_1 process $\{\langle X, v\rangle : v \in V\}$ for any $V \subset \mathbb{R}^d$, and all with an absolute constant L. At the same time, it is quite likely that there will be some directions in which $\langle X, v\rangle$ behaves in a far better way than the ψ_1 condition would indicate, making the problem of finding sharp estimates on $\mathbb{E}\sup_{v\in V} X_v$ a very subtle one.

In the context of the $\bar{\gamma}$ functionals, the fact that a process is ψ_α yields a trivial upper estimate. Indeed, for an arbitrary admissible sequence of V, one has for every $v \in V$

$$\|Z_v\|_{L_{2^{s_0+4}}} + \sum_{s>s_0} \|Z_{\pi_{s+1}v} - Z_{\pi_s v}\|_{L_{2^{s+4}}} \leq cL \left(2^{s_0/\alpha}\|Z_v\|_{L_2} + \sum_{s\geq s_0} 2^{s/\alpha} \| \|Z_{\pi_{s+1}v} - Z_{\pi_s v}\|_{L_{2^s}} \right),$$

implying that

$$\bar{\gamma}(Z_V, s_0) \leq CL \inf\sup_{v\in V} \left(2^{s_0/\alpha}\|Z_v\|_{L_2} + \sum_{s\geq s_0} 2^{s/\alpha} \| \|Z_{\pi_{s+1}v} - Z_{\pi_s v}\|_{L_{2^s}} \right).$$

Definition 11.8. *Let (V, d) be a metric space. For $\alpha > 0$ and an integer $s_0 \geq 0$ set*

$$\gamma_{\alpha,s_0}(V, d) = \inf\sup_{v\in V} \left(2^{s_0/\alpha}\mathrm{diam}(V, d) + \sum_{s\geq s_0} 2^{s/\alpha} d(\pi_{s+1}v, \pi_s v) \right),$$

where the infimum is taken with respect to all admissible sequences of V.

Clearly, when V is the indexing set of a ψ_α process Z_V, and d is the L_2 metric endowed on V by the process, then

$$\bar{\gamma}(Z_V, s_0) \leq CL\gamma_{\alpha,s_0}(V, d).$$

Again, it should be stressed that $\bar{\gamma}(Z_V, s_0)$ is a way of obtaining an upper estimate on $\mathbb{E}\sup_{v\in V} Z_v$, and if the process is ψ_α, then $\gamma_{\alpha,s_0}(V)$ is a way of obtaining an upper estimate on $\bar{\gamma}(Z_V, s_0)$. Neither one of the two steps need be sharp. The fact that both are sharp for gaussian processes (and a few other canonical processes) is rather a miracle, and Talagrand's proof of this fact is a masterpiece of beautiful mathematics.

Understanding when functionals like $\bar{\gamma}$ or γ_α yield sharp estimates on $\mathbb{E}\sup_{v\in V} Z_v$, and, moreover, what is the right bound when they don't, is a real challenge that is very far from being resolved.

11.2 γ_α and metric entropy

The results described above may seem as rather unsatisfactory, because that has actually been accomplished up to this point was replacing a mysterious quantity, $\mathbb{E}\sup_{v\in V} Z_v$, with another mysterious quantity—the $\bar{\gamma}$ functional. Even taking the substantial leap of faith, that the latter can serve as a useful bound on the former, the key question remains:

Question 11.9. *How one may construct a good admissible sequence for $\bar{\gamma}$? And, assuming one chooses to estimate $\bar{\gamma}$ by γ_α for ψ_α processes, how one may construct an admissible sequence metric spaces (T, d)?*

Constructing *optimal* admissible sequences, even up to a multiplicative constant, is a very difficult task and is not known in general—even when $T \subset \mathbb{R}^d$ and $d(x, y) = \|x-y\|_2$. It follows from the majorizing measures theorem that an optimal admissible sequence exists and that $\gamma_2(T, L_2)$ is equivalent to the expectation of the supremum of the gaussian process indexed by T (with L_2 corresponds to its covariance structure). The proof is constructive in the sense that there is a greedy algorithm that produces the admissible sequence. Unfortunately, examples in which that construction can be clearly specified as few and far-between; $\gamma_2(T, L_2)$ is, by-far, the case where one has the most extensive understanding of what is going on....

If one is willing to be less ambitions, and settle for a bound that is likely to be loose, there is a generic construction of an admissible sequence for an arbitrary metric space. As it happens, that simple construction, which is based on covering numbers, is good enough in many cases.

Let us consider $\gamma_\alpha(T, d)$, as the analogous construction for $\bar{\gamma}$ will be clear from this example.

To define an admissible sequence using a covering, fix an integer $s \geq 0$ and let T_s be an ε_s-cover of T with respect to the metric d. Hence, ε_s is the smallest number for which T has an ε-cover of cardinality at most 2^{2^s} (if the infimum is not attained, one may take ε_s to be twice the infimum). Let $\pi_s t$ be the nearest point to t in T_s, implying that for every $t \in T$, $d(t, \pi_s t) \leq \varepsilon_s$, and therefore, that

$$\gamma_{\alpha,s_0}(T, d) \leq \sum_{s \geq s_0} 2^{s/\alpha} \varepsilon_s. \tag{11.11}$$

Remark 11.10. *Observe that the difference between $\gamma_{\alpha,s_0}(T, d)$ and $\sum_{s \geq s_0} 2^{s/\alpha} \varepsilon_s$ is essentially changing the order of the supremum and the sum.*

Equation (11.11) has a more friendly presentation, frequently (and somewhat inaccurately) called the *Dudley entropy integral bound*.

Theorem 11.11. *For every $0 < \alpha < \infty$ there exists a constant c_α for which the following holds. If (T, d) is a metric space then*

$$\gamma_\alpha(T, d) \leq c_\alpha \int_0^D \log^{1/\alpha} \mathcal{N}(T, \varepsilon) d\varepsilon,$$

where D is the diameter of (T, d).

Proof. Without loss of generality assume that $\varepsilon_s > 0$ and that ε_s is the smallest number for which T contains an ε-cover of cardinality at most 2^{2^s}. Therefore, $\mathcal{N}(T, \varepsilon_s/2) \geq 2^{2^s}$. Hence, if $x \in (\varepsilon_{s+1}/2, \varepsilon_s/2)$ then $\mathcal{N}(T, x) \geq 2^{2^s}$. In particular,

$$2^{s/\alpha}(\varepsilon_s - \varepsilon_{s+1}) \leq 2 \int_{\varepsilon_{s+1}/2}^{\varepsilon_s/2} \log^{1/\alpha} \mathcal{N}(T, x) dx.$$

Summing the left-hand side,

$$\sum_{s\geq 0} 2^{s/\alpha}(\varepsilon_s - \varepsilon_{s+1}) = \sum_{s\geq 0} 2^{s/\alpha}\varepsilon_s - \sum_{s\geq 0} 2^{s/\alpha}\varepsilon_{s+1} = \sum_{s\geq 0} 2^{s/\alpha}\varepsilon_s - \sum_{s\geq 1} 2^{(s-1)/\alpha}\varepsilon_s$$
$$\geq (1-2^{-1/\alpha})\sum_{s\geq 0} 2^{s/\alpha}\varepsilon_s.$$

Therefore,

$$(1-2^{-1/\alpha})\sum_{s\geq 0} 2^{s/\alpha}\varepsilon_s \leq 2\sum_{s\geq 0}\int_{\varepsilon_{s+1}/2}^{\varepsilon_s/2} \log^{1/\alpha}\mathcal{N}(T,x)dx$$
$$\leq 2\int_0^\infty \log^{1/\alpha}\mathcal{N}(T,x)dx,$$

and the claim follows because $\log\mathcal{N}(T,x) = 0$ for $x \geq \operatorname{diam}(T,d)$. ∎

Corollary 11.12. *There exists an absolute constant c for which the following holds. Let X be an isotropic, L-subgaussian random vector in $\mathbb{R}^n$. Then for every $T \subset \mathbb{R}^n$,*

$$\mathbb{E}\sup_{t\in T}\langle X,t\rangle \leq c\int_0^{d_T}\sqrt{\log\mathcal{N}(T,\varepsilon B_2^n)}d\varepsilon.$$

In particular, if $X = (\varepsilon_1,...,\varepsilon_n)$ then

$$\mathbb{E}\sup_{t\in T}\sum_{i=1}^n \varepsilon_i t_i \leq c\int_0^{d_T}\sqrt{\log\mathcal{N}(T,\varepsilon B_2^n)}d\varepsilon.$$

Exercise 35. *Prove Corollary* (11.12).

11.2.1 Example: B_1^d

Let us explore the two bounds on $\gamma_2(B_1^d,\ell_2)$—firstly, an lower estimate based on covering numbers and the Dudley's entropy integral, and secondly, by constructing an optimal admissible sequence. These estimates show that the entropy based bound in suboptimal in this case. It will also be a good indication that constructing an optimal admissible sequence is a nontrivial task, even in seemingly simple situations.

Thanks to the majorizing measures theorem, $\gamma_2(B_1^d,\ell_2)$ is known up to absolute multiplicative constants, since

$$\mathbb{E}\sup_{t\in B_1^d}\langle G,t\rangle = \mathbb{E}\|G\|_\infty \sim \sqrt{\log d}.$$

The next exercise shows that

$$\int_0^\infty \sqrt{\log\mathcal{N}(B_1^d,\varepsilon B_2^d)}d\varepsilon \geq c\log^{3/2} d.$$

Exercise 36. *Let $1 \leq s \leq d$ and set $\mathcal{I}_s$ to be the collection of all subsets of $\{1,...,d\}$ of cardinality s. Show that the set*

$$\left\{\frac{1}{\sqrt{s}}\sum_{i\in I} e_i : i \in \mathcal{I}_s\right\}$$

is a subset of B_1^d and contains a $c_1/\sqrt{s}$ separated subset of cardinality at least $\exp(c_2 s\log(ed/s))$.

Use that to deduce a lower estimate on $\log \mathcal{N}(B_1^d, \varepsilon_j B_2^d)$ for $\varepsilon_j \sim 2^j/\sqrt{d}$ and control the entropy integral from below.

The construction of an optimal admissible sequence is more involved. First of all, assume without loss of generality that $d = 2^{\ell+1}$ for some integer ℓ. For every $0 \le k \le \ell$, let $\mathcal{I}_k$ be the collection of all subsets of $\{1, ..., d\}$ of cardinality 2^k. Thus, using the monotone nonincreasing rearrangement of $x \in \mathbb{R}^d$, one can write

$$x = \sum_{k=0}^{\ell} P_{I_k(x)} x,$$

with $I_k(x) \in \mathcal{I}_k$ and $I_0(x), I_1(x), ...$ are disjoint blocks of cardinality 2^k, $k = 0, 1, ...$—according to the nonincreasing rearrangement of the coordinates of x. Moreover, because of the blocks correspond to the nonincreasing rearrangement, for every $0 \le k \le \ell - 1$ one has

$$\|P_{I_{k+1}(x)} x\|_\infty \le \frac{\|P_{I_k(x)} x\|_1}{2^k} \quad \text{and} \quad \|P_{I_{k+1}(x)} x\|_2 \le 2^{(k+1)/2} \|P_{I_{k+1}(x)} x\|_\infty \le 2 \frac{\|P_{I_k(x)} x\|_1}{2^{k/2}}. \tag{11.12}$$

Hence, for every $x \in B_1^d$, $\|P_{I_k(x)} x\|_2 \lesssim 1/2^{k/2}$. Moreover,

$$\|\sum_{j>k} P_{I_j(x)} x\|_2^2 \le \sum_{j>k} \frac{\|P_{I_j(x)} x\|_1^2}{2^j}. \tag{11.13}$$

From here on, to ease notation, when the identity of x is clear we write P_{I_k} instead of $P_{I_k(x)}$.

Lemma 11.13. *Let $\theta > 1$. There are absolute constants c_0 and c_1 for which the following holds. For every $0 \le k \le \ell$ there exists $W_k \subset B_1^d$ such that*

$$|W_k| \le \exp\left(c_0 \theta 2^k \log\left(\frac{ed}{2^k} \right) \right),$$

and for every $x \in B_1^d$ there is $z \in W_k$ that is supported on $I_k(x)$ such that

$$\|P_{I_k(x)} x - z\|_2 \le c_1 \left\{ \left(\frac{2^k}{d} \right)^{\theta - 1} \|P_{I_k(x)} x\|_2, \frac{1}{\sqrt{d}} \right\}. \tag{11.14}$$

The first step in the construction of the sets W_k is to consider all the subsets I of $\{1, ..., d\}$ of cardinality 2^k, and the collection of balls $2^{-j/2} B_2^I$ for $j \ge k$. The idea is that after the decomposition of each x according to the subsets $I_0(x), ..., I_k(x), ...I_\ell(x)$, there is no information on $\|P_{I_k(x)} x\|_2$ beyond the fact that this norm is at most $\sim 1/2^{k/2}$, and that is where the balls $2^{-j/2} B_2^I$ for $I = I_k(x)$ and $j \ge k$ come in. As (11.14) indicates, the goal is approximate $P_{I_k(x)} x$ up to an error which is either trivial, or, at most, a small order of $\|P_{I_k(x)} x\|_2$.

Proof of Lemma 11.13.

Consider $0 \le k \le \ell$ and $j = k, ..., \ell$, and let us construct covers of

$$U_{k,j} = \bigcup_{I \in \mathcal{I}_k} \frac{1}{2^{j/2}} B_2^I,$$

where for each j the cover is of cardinality $\exp(\theta 2^k \log(ed/2^k))$. Since there are at most $\binom{d}{2^k}$ subsets in $\mathcal{I}_k$, by the volumetric estimate for each ball $2^{-j/2}B_2^I$, one has

$$\mathcal{N}(U_{k,j}, \varepsilon B_2^d) \leq \binom{d}{2^k} \left(\frac{5}{2^{j/2}\varepsilon}\right)^{2^k}.$$

Consider the union of such covers of $U_{k,j}$, for $j = k, ..., \ell$. It follows that as long as the mesh-width of the net of $U_{k,j}$ is at most

$$2^{-j/2}(2^k/d)^{\theta-1}, \tag{11.15}$$

the total number of points in the union of these covers is

$$\sum_{j=k}^{\ell} \exp(\theta 2^k \log(ed/2^k)) = \exp(\log(\ell-k) + \theta 2^k \log(ed/2^k)) \leq \exp(c_0 \theta 2^k \log(ed/2^k)) \tag{11.16}$$

for an absolute constant $c_0 > 1$, because $\ell - k \sim \log(d/2^k)$.

Define the set W_k as the union of these covers with 0.

Add to this cover the point 0 and denote it by W_k. Now fix $x \in B_1^d$ and recall that $\|P_{I_k}x\|_2 \leq 2/2^{k/2}$. Observe that if $\|P_{I_k}x\|_2 \geq 2/2^{\ell/2} \sim 1/\sqrt{d}$ then it satisfies that

$$2^{-(j+1)/2} \leq \|P_{I_k}x\|_2\| \leq 2^{-j/2}$$

for some $k \leq j \leq \ell$. By approximating $P_{I_k}x$ using a appropriate point $z \in U_{k,j}$ that is supported on I_k, it is evident from (11.15) that

$$\|P_{I_k}x - z\|_2 \leq 2^{-j/2}\left(\frac{2^k}{d}\right)^{\theta-1} \lesssim \|P_{I_k}x\|_2\left(\frac{2^k}{d}\right)^{\theta-1},$$

as required. Otherwise, $\|P_{I_k}x\|_2 \lesssim 1/\sqrt{d}$ and one may choose 0 as the approximating point. ∎

Next, for $0 \leq k \leq \ell$ consider the approximating sets V_k obtained by combining the sets $(W_s)_{s\leq k}$ in a reasonable way:

$$V_k = \left\{\sum_{s=0}^{k}\sum_{i\in I_s} w_j : \ |I_s| = 2^s, \ I_1, ..., I_k \text{ are disjoint}, \ \ w_s \in W_s\right\}.$$

In other words, V_k consists of all the points that are of the following form: taking disjoint subsets of $\{1, ..., d\}$, one for each cardinality 2^s, $1 \leq s \leq k$, and on each block I_s select some $w_s \in W_s$ that is supported on that block. The idea is that just like any $x \in B_1^d$ can be approximated by a point in W_k on $I_k(x)$, it can be well approximated by a point $\pi_k x \in V_k$ on $\bigcup_{s\leq k} I_s(x)$, implying that

$$\|x - \pi_k x\|_2 \leq \left\|\sum_{s\leq k} P_{I_s(x)}x - \pi_k x\right\|_2 + \left\|\sum_{s>k} P_{I_s(x)}x\right\|$$

is small. Formally,

Lemma 11.14. *There exist absolute constants c_2 and c_3 for which the following holds. For any $0 \le k \le \ell$,*

$$|V_k| \le \exp\left(c_2\theta 2^k \log\left(\frac{ed}{2^k}\right)\right),$$

and for any $x \in B_1^d$ there is $\pi_k x \in V_k$ such that

$$\|x - \pi_k x\|_2 \le c_3\left(\sum_{s=0}^{k}(2^s/d)^{\theta-1}\frac{\|P_{I_j}x\|_1}{2^{s/2}} + \sum_{s>k}\frac{\|P_{I_s}x\|_1}{2^{s/2}} + \sqrt{\frac{k}{d}}\right)$$

Proof. To estimate the cardinality of V_k, note that it is bounded by the cardinality of a set containing all possible sums $\sum_{s=0}^k w_s$, with $w_s \in W_s$. Thus, by (11.16), the cardinality of the latter is at most

$$\prod_{s=0}^{k}\exp(c_0\theta 2^s\log(ed/2^s)) \le \exp(c_2\theta 2^k\log(ed/2^k)), \tag{11.17}$$

implying that the same holds for V_k.

Observe that by the construction and (11.14), for every $x \in B_1^d$ and every $0 \le s \le k$, there is $w_s \in W_s$ that is supported on $I_s(x)$ and

$$\begin{aligned}\|\sum_{s=0}^{k}P_{I_s}x - v\|_2^2 &= \sum_{s=0}^{k}\|P_{I_s}x - w_s\|_2^2 \le \sum_{s=0}^{k}c_1\max\left\{(2^s/d)^{2(\theta-1)}\|P_{I_s}x\|_2^2, \frac{1}{d}\right\}\\ &\lesssim \sum_{s=0}^{k}(2^s/d)^{2(\theta-1)}\|P_{I_s}x\|_2^2 + \frac{k}{d}\\ &\lesssim \sum_{s=0}^{k}(2^s/d)^{2(\theta-1)}\frac{\|P_{I_s}x\|_1^2}{2^s} + \frac{k}{d},\end{aligned}$$

using that by (11.12)

$$\|P_{I_s}x\|_2^2 \le 2\frac{\|P_{I_s}x\|_1}{2^{s/2}}.$$

Clearly, $\sum_{s=0}^k w_s \in V_k$ and denote that point by $\pi_k x$. Therefore, by (11.13), for every $x \in B_1^d$ there is $\pi_k x \in V_k$ such that

$$\|x - \pi_k x\|_2^2 = \|\sum_{s=0}^{k}P_{I_s}x - \pi_k x\|_2^2 + \|\sum_{s>k}P_{I_s}x\|_2^2 \lesssim \sum_{s=0}^{k}(2^s/d)^{2(\theta-1)}\frac{\|P_{I_s}x\|_1^2}{2^s} + \frac{k}{d} + \sum_{s>k}\frac{\|P_{I_s}x\|_1^2}{2^s}. \tag{11.18}$$

Since $\|a\|_2 \le \|a\|_1$, it is evident that

$$\|x - \pi_k x\|_2 \lesssim \sum_{s=0}^{k}(2^s/d)^{\theta-1}\frac{\|P_{I_s}x\|_1}{2^{j/2}} + \sum_{s>k}\frac{\|P_{I_s}x\|_1}{2^{j/2}} + \sqrt{\frac{k}{d}}, \tag{11.19}$$

as claimed. ∎

Thanks to (11.19), it is possible to obtain an upper estimate for a functional that is close to $\gamma_2(B_1^d, \ell_2)$.

Lemma 11.15. *There exists an absolute constant c for which, for every $x \in B_1^d$,*

$$\sum_{k=0}^{\ell} 2^{k/2} \|x - \pi_k x\|_2 \leq c\sqrt{\log d}.$$

Proof. Fix $x \in B_1^d$. By (11.19) one has to consider the sum of the three terms. Firstly, recalling that $2^\ell \sim d$,

$$\sum_{k=0}^{\ell} 2^{k/2} \cdot \sqrt{\frac{k}{d}} \sim \sqrt{\ell} \sim \sqrt{\log d},$$

as required. Secondly,

by changing the order of summation,

$$\begin{aligned}
&\sum_{k=0}^{\ell} 2^{k/2} \cdot \sum_{s=0}^{k} \left(\frac{2^s}{d}\right)^{\theta-1} \frac{\|P_{I_s} x\|_1}{2^{s/2}} = \sum_{s=0}^{\ell} \left(\frac{2^s}{d}\right)^{\theta-1} \frac{\|P_{I_s} x\|_1}{2^{s/2}} \sum_{k=s}^{\ell} 2^{k/2} \\
\leq &\sum_{s=0}^{\ell} \left(\frac{2^s}{d}\right)^{\theta-1} \left(\frac{d}{2^s}\right)^{1/2} \|P_{I_s} x\|_1 = \sum_{s=0}^{\ell} (2^s/d)^{\theta-3/2} \|P_{I_s} x\|_1 \\
\leq &\left(\max_s (2^s/d)^{\theta-3/2}\right) \cdot \sum_{s=0}^{\ell} \|P_{I_s} x\|_1 \leq c
\end{aligned}$$

provided that $\theta > 3/2$.

Finally,

$$\sum_{k=0}^{\ell} 2^{k/2} \sum_{s>k} \frac{\|P_{I_s} x\|_1}{2^{s/2}} = \sum_{s=0}^{\ell} \frac{\|P_{I_s} x\|_1}{2^{s/2}} \sum_{k=0}^{s} 2^{k/2} = c \sum_{s=0}^{\ell} \|P_{I_s} x\|_1 \leq c'.$$

■

Exercise 37. *Use the sets V_k and Lemma 11.15 to construct an admissible sequence of B_1^d, showing that $\gamma_2(B_1^d, \ell_2) \sim \sqrt{\log d}$.*

Another well-known example in which there is a true gap between the γ_2 functional and the entropy integral are ellipsoids in ℓ_2. We refer the reader to [**?**] for details of this generic example. Despite this gap, there are still many interesting cases in which the entropy integral yields a useful estimate, and in any case, as in the case of B_1^d, understanding how efficient covers of the set can be constructed is a first step towards on optimal admissible sequence.

11.2.2 Random coordinate projections

Keeping in mind that the main focus of these notes is statistical learning theory, in makes sense to see what an entropy integral bound means as far as the Rademacher averages are concerned. Thus, for a class of functions F, let $f^* \in F$. Assume for the sake of simplicity that F is star-shaped around f^*, implying that $F_{f^*,r} = (F - f^*) \cap rD$. Let us explore

$$\mathbb{E} \sup_{u \in F_{f^*,r}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i u(X_i) \right| = \mathbb{E}_X \left(\mathbb{E}_\varepsilon \sup_{v \in P_{\sigma F_{f^*,r}}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i v_i \right| \right).$$

The trivial estimate follows from the fact that a Bernoulli process is L-subgaussian, and in particular, $\|\langle \mathcal{E}, t\rangle\|_{L_p} \lesssim \|\langle G, t\rangle\|_{L_p}$ for every $t \in \mathbb{R}^N$. Therefore, by the chaining argument, given any $\sigma = (X_1, ..., X_N)$,

$$\mathbb{E}_\varepsilon \sup_{v \in P_{\sigma F_{f^*,r}}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i v_i \right| \le c\gamma_2(P_\sigma F_{f^*,r}/\sqrt{N}, \ell_2).$$

Note that when endowed on F, the metric in question depends on σ: for any $u, v \in F_{f^*,r}$,

$$\left\| \frac{1}{\sqrt{N}} P_\sigma u - \frac{1}{\sqrt{N}} P_\sigma v \right\|_2^2 = \frac{1}{N} \sum_{i=1}^N (u-v)^2(X_i),$$

implying that

$$\mathbb{E}_X \mathbb{E}_\varepsilon \sup_{v \in P_{\sigma F_{f^*,r}}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \varepsilon_i v_i \right| \le c\mathbb{E}_X \gamma_2(F_{f^*,r}, L_2^\sigma) = (*).$$

Unfortunately, a useful estimate on $(*)$ calls for the construction of an optimal admissible sequences of $F_{f^*,r}$ with respect to the random L_2 metrics endowed by the sample. And at this point, there is no indication that these metrics are close to the original L_2 one (though that is an extremely important and well-studied question).

One can relax the estimate further by turning to the entropy integral, as

$$\mathbb{E}_X \gamma_2(F_{f^*,r}, L_2^\sigma) \le c\mathbb{E}_X \int_0^\infty \log^{1/2} \mathcal{N}(F_{f^*,r}, \varepsilon B(L_2^\sigma)) d\varepsilon = c\mathbb{E}_X \int_0^R \log^{1/2} \mathcal{N}(F_{f^*,r}, \varepsilon B(L_2^\sigma)) d\varepsilon,$$

where

$$R = \sup_{u \in F_{f^*,r}} \left(\frac{1}{N} \sum_{i=1}^N u^2(X_i) \right)^{1/2}.$$

Once again, relating the diameter of $F_{f^*,r}$ with respect to the random metric L_2^σ to the L_2 diameter (which is at most $2r$) is a highly nontrivial question which is far from being fully understood.

One way around these obstacles is to find some way of obtaining *uniform entropy estimates*, that is, estimates that would holds for any σ. That was the motivation behind the introduction of the combinatorial dimension. However, the real question—the possible equivalence between the spaces (F, L_2) and (F, L_2^σ) for a typical sample σ,—is, in general, open.

Remark 11.16. *Note that the estimate on the Rademacher averages is loose already in the first step: it is based on the fact that a Bernoulli process is subgaussian, implying that the expectation $\mathbb{E}\sup_{v \in P_\sigma F_{f^*,r}} |N^{-1/2} \sum_{i=1}^N \varepsilon_i v_i|$ is dominated by $\gamma_2(P_\sigma F_{f^*,r}/\sqrt{N}, L_2^\sigma)$. However, there are situations in which reverting to the gaussian case comes at a high price, most notably, when trying to analyze the quadratic empirical process (see, e.g., [?]). Although this aspect will not be pursued further in this notes, this fact should be kept in mind.*

Chapter 12

Sudakov type inequalities

The entropy integral is a way of obtaining upper estimates on $\mathbb{E}\sup_{v\in V} Z_v$ using metric properties of V with respect to the right metric structure/s. It is natural to ask whether there exist similar lower bounds. It turns out that it is much harder to obtain lower bounds of that flavour, and understanding when such bounds are possible is an extremely difficult question that is still open. In the chapter we present two such metric-based bounds: for gaussian processes and for Bernoulli processes. There are more general examples of similar bounds (see, e.g. [**?**]), but those are beyond the scope of this presentation.

The first result we present is a simple proof of Sudakov's inequality in $\mathbb{R}^d$, due to N. Tomczak-Jaegermann [**?**]. The argument is based on Theorem 10.6. A different proof is presented in what follows.

Recall that for $T \subset \mathbb{R}^d$, $\ell_*(T) = \mathbb{E}\sup_{t\in T} |\langle G, t\rangle|$.

Theorem 12.1. *There exists an absolute constant c such that, for every $T \subset \mathbb{R}^d$ and every $\varepsilon > 0$,*

$$c\varepsilon \log^{1/2} \mathcal{N}(T, \varepsilon B_2^d) \leq \ell_*(T).$$

Proof. Is suffices to prove Theorem 12.1 under some additional assumptions on T: firstly, that T is a convex and centrally-symmetric, because $\ell_*(T) = \ell_*(\mathrm{absconv}(T))$; that T is bounded—otherwise the statement is trivially true; and, by replacing $\mathbb{R}^d$ by $\mathrm{span}(T)$, that T has a nonempty interior. Thus, without loss of generality, T is the unit ball of a norm on $\mathbb{R}^d$ which is denoted by $\| \ \|_T$, and its dual norm is denoted by $\| \ \|_{T^\circ}$. Observe that

$$2T \cap (\varepsilon^2/2)T^\circ \subset \varepsilon B_2^d, \tag{12.1}$$

because $\|x\|_2^2 = \langle x, x\rangle \leq \|x\|_T \cdot \|x\|_{T^\circ}$. Moreover,

$$\mathcal{N}(T, 2T \cap (\varepsilon^2/2)T^\circ) = \mathcal{N}(T, (\varepsilon^2/2)T^\circ); \tag{12.2}$$

indeed, one direction is clear because $2T \cap (\varepsilon^2/2)T^\circ \subset (T, (\varepsilon^2/2)T^\circ$. In the other direction, if $T \subset \bigcup(y_i + rT^\circ)$ then for every $x \in T$ there is some y_i for which $x - y_i \in rT^\circ$. But T is convex and centrally-symmetric, $x - y_i \in 2T$, and thus $T \subset \bigcup(y_i + 2T \cap (rT^\circ))$.

Combining (12.1) and (12.2), and since covering numbers are sub-multiplicative,

$$\begin{aligned}
\mathcal{N}(T, \varepsilon B_2^d) \leq& \mathcal{N}(T, 2T \cap (\varepsilon^2/2)T^\circ) \leq \mathcal{N}(T, (\varepsilon^2/2)T^\circ) \\
\leq& \mathcal{N}(T, 2\varepsilon B_2^d) \cdot \mathcal{N}(2\varepsilon B_2^d, (\varepsilon^2/2)T^\circ) = \mathcal{N}(T, 2\varepsilon B_2^d) \cdot \mathcal{N}(B_2^d, (\varepsilon/4)T^\circ) \\
\leq& \mathcal{N}(T, 2\varepsilon B_2^d) \exp(c_1 \ell_*^2(T)/\varepsilon^2),
\end{aligned}$$

where the last inequality follows from Theorem 10.6. Thus, setting $\phi(\varepsilon) = \log \mathcal{N}(T, \varepsilon B_2^d)$, one has, for every $\varepsilon > 0$,

$$\phi(\varepsilon) \leq \phi(2\varepsilon) + c_1 \frac{\ell_*^2(T)}{\varepsilon^2}.$$

Clearly, $\lim_{j\to\infty} \phi(2^j\varepsilon) = 0$ and therefore,

$$\phi(\varepsilon) = \sum_{j=1}^{\infty} \big(\phi(2^j\varepsilon) - \phi(2^{j+1}\varepsilon)\big) \leq c_1 \frac{\ell^2(T)}{\varepsilon^2} \sum_{j=1}^{\infty} 2^{-2j} \leq c_2 \frac{\ell^2(T)}{\varepsilon^2}.$$

■

Because Sudakov's inequality is so important, let us present an alternative proof which is more direct. The proof is based on a the fundamental idea that "bounded+being extremal" implies the exitance of structure. One first obtains a seemingly weak entropy bound, but when applied to the right projection of the set—i.e., to a set that will be shown both bounded and extremal, the entropy estimate turns out to be both obvious and sharp.

12.1 A direct proof of Sudakov's inequality

Recall that one may assume that $T \subset \mathbb{R}^d$ is a convex body, and the proof is based a fundamental fact from convex geometry, known as *Urysohn's inequality.*

Theorem 12.2. *Let T be a convex body in $\mathbb{R}^d$. Then*

$$\left(\frac{|T|}{|B_2^d|}\right)^{1/d} \leq \int_{S^{d-1}} \|x\|_{T^\circ} dx(\sigma), \tag{12.3}$$

where integration is with respect to the Haar measure on S^{d-1}. Moreover,

$$\left(\frac{|T|}{|B_2^d|}\right)^{1/d} \leq \frac{\ell_*(T)}{\ell_*(B_2^d)}. \tag{12.4}$$

Exercise 38. (1) *Show that for the standard gaussian vector G, one has that $\|G\|_2$ and $G/\|G\|_2$ are independent.*

(2) *Use the gaussian representation of the Haar measure on S^{d-1},* (1) *and* (12.3) *to deduce* (12.4)*.*

Let m be an integer to be selected later and consider the random operator

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

where $G_1, ..., G_m$ are independent copies of the gaussian random vector in $\mathbb{R}^d$.

Lemma 12.3. *There exist absolute constants c_1, c_2 and c_3 for which the following holds.*

(1) If $x \in \mathbb{R}^d$, then with probability at least $1 - 2\exp(-c_1 m)$, $\|\Gamma x\|_2 \geq c_2 \|x\|_2$.

(2) $\mathbb{E}\ell_*(\Gamma T) \leq c_3 \ell_*(T)$, where the expectation is with respect to $G_1, ..., G_m$.

Before presenting the proof of the Lemma, let us recall the following standard fact: if O is a random orthogonal matrix, that is, if O is distributed according to the Haar measure on the orthogonal group $O(d)$ then Oe_1 is distributed uniformly on S^{d-1}. Indeed, Oe_1 is a random vector taking values on S^{d-1} and is rotation invariant under $O(d)$. Thus, by the uniqueness of the Haar measure, it must coincide with Oe_1. In particular, one has that

$$\ell_*(T) \sim \sqrt{d}\mathbb{E}_O \sup_{t\in T}\langle Oe_i, t\rangle. \tag{12.5}$$

Proof of Lemma 12.3. The first part of the claim is an immediate outcome of the small-ball property of a gaussian variable. Since this argument has already appeared several times in these notes, it is omitted.

Turning to the second part, let $K = \Gamma T \subset \mathbb{R}^m$ and note that by (12.5),

$$\ell_*(K) \sim \sqrt{m}\mathbb{E}_O \sup_{t\in T}\langle \Gamma^* Oe_1, t\rangle = \sqrt{m}\mathbb{E}_O \sup_{t\in T}\langle \Gamma^* Oe_1, t\rangle.$$

Therefore, taking the expectation with respect to $G_1, ..., G_m$, followed by a Fubini argument,

$$\mathbb{E}\mathbb{E}_O \sup_{t\in T}\langle \Gamma^* Oe_1, t\rangle = \mathbb{E}_O \mathbb{E} \sup_{t\in T}\langle \Gamma^* Oe_1, t\rangle.$$

Moreover, for any fixed orthogonal matrix O, the distribution of $\Gamma^* Oe_1 = (\langle \Gamma^i, Oe_i\rangle)_{i=1}^m$ is the same as for a standard gaussian vector in $\mathbb{R}^m$. Hence, for every fixed orthogonal matrix O,

$$\mathbb{E} \sup_{t\in T}\langle \Gamma^* Oe_1, t\rangle = \mathbb{E} \sup_{t\in T}\langle G, t\rangle = \ell_*(T),$$

completing the proof. ∎

Exercise 39. *Show that indeed, for any $O \in O(d)$, $\Gamma^* Oe_1$ is distributed as the standard gaussian random vector in $\mathbb{R}^m$.*

As was mentioned previously, the first component of this proof of Sudakov's inequality is actually a weak version of the theorem.

Lemma 12.4. *Let $T \subset \mathbb{R}^d$ be a convex body. Then for every $u > 0$,*

$$\log \mathcal{M}(T, u\ell_*(T)B_2^d) \leq d \log\left(1 + \frac{1}{u\sqrt{d}}\right) \leq \frac{\sqrt{d}}{u}.$$

Remark 12.5. *The full version of Sudakov's inequality would imply that* $\log \mathcal{M}(T, u\ell_*(T)B_2^d) \leq c/u^2$*, rather than* $\sqrt{d}/u$*.*

Proof. Recall that by a volumetric estimate,

$$\mathcal{M}(T, sB_2^d) \leq \frac{|T + sB_2^d|}{|sB_2^d|} = (*).$$

By the gaussian version of Urysohn's inequality and the subadditivity of ℓ_* (i.e. $(\ell_*(A+B) \leq \ell_*(A) + \ell_*(B))$,

$$|T + sB_2^d| \leq |B_2^d| \cdot \left(\frac{\ell_*(T + sB_2^d)}{\ell_*(B_2^d)}\right)^d \leq |B_2^d| \cdot \left(\frac{\ell_*(T) + s\ell_*(B_2^d)}{\ell_*(B_2^d)}\right)^d.$$

Therefore,

$$(*) \leq \left(\frac{\ell_*(T) + s\ell_*(B_2^d)}{s\ell_*(B_2^d)}\right)^d \leq\leq \left(1 + \frac{\ell_*(T)}{s\sqrt{d}}\right)^d,$$

where the last inequality holds because $\ell_*(B_2^d) = \mathbb{E}(\sum_{i=1}^d g_i^2)^{1/2} \leq \sqrt{d}$. Hence, setting $s = u\ell_*(T)$ one has

$$\log \mathcal{M}(T, u\ell_*(K)B_2^d) \leq d \log\left(1 + \frac{1}{u\sqrt{d}}\right) \leq \frac{\sqrt{d}}{u},$$

as claimed. ∎

Proof of Sudakov's inequality. To prove Sudakov's inequality, let $A = \{x_1, ..., x_{e^r}\} \subset T$ be $u\ell_*(T)$ separated in ℓ_2^d, and one has to show that there is an absolute constant c for which $r \leq c/t^2$.

To that end, fix an integer m to be named later, and consider the random mapping $\Gamma : \mathbb{R}^d \to \mathbb{R}^m$. Applying the first part of Lemma 12.3, one has that if $r \leq c_1 m$ there is an event of probability at least $1 - 2\exp(2r) \cdot \exp(-c_0 m) \geq 1 - 2\exp(-c_0 m/2)$ on which, for every $x_i \neq x_j$, $x_i, x_j \in A$,

$$\|\Gamma(x_i - x_j)\|_2 \geq c_2 \|x_i - x_j\|_2 \geq c_2 u\ell_*(T).$$

Also, by Markov's inequality, $\ell_*(\Gamma T) \leq 2\ell_*(T)$ with probability at least $1/2$. Hence, there exists Γ for which both properties holds. Therefore,

$$\log \mathcal{M}(\Gamma T, (c_2/2)u\ell_*(\Gamma T)) \geq \log \mathcal{M}(\Gamma T, c_2 u\ell_*(T)) \geq r.$$

On the other hand, by Lemma 12.4 for the set $\Gamma T \subset \mathbb{R}^m$,

$$\log \mathcal{M}(\Gamma T, (c_2/2)u\ell_*(\Gamma T)) \leq \frac{\sqrt{m}}{(c_2/2)u},$$

and taking the minimal 'legal' choice of m, namely, $m = r/c_1$, it follows that $r \leq c_3/u^2$ for a suitable absolute constant c_3, as required. ∎

12.2 Sudakov's inequality for Bernoulli processes

In what follows we present a version of Sudakov's inequality for Bernoulli processes $v \to \langle \mathcal{E}, v \rangle$, where, as always, we denote by $\mathcal{E}$ the standard Bernoulli vector in $\mathbb{R}^d$. The first question that should be asked is on the right formulation of the inequality, because using the gaussian formulation is clearly false:

Example 12.6. *Let $T = \{e_1, ..., e_d\}$. Then for any $\varepsilon < \sqrt{2}$, $\log \mathcal{M}(T, \varepsilon B_2^d) = \log d$. At the same time,*

$$\mathbb{E} \sup_{t \in T} \left| \sum_{i=1}^d \varepsilon_i t_i \right| = 1.$$

As it happens, the reason for the difficulty is the price one pays for the regularity of the gaussian process—that $\|\langle G, t \rangle\|_{L_p} \sim \sqrt{p}\|t\|_2$ for every $t \in \mathbb{R}^d$ and $p \geq 2$. Let us try to re-write Sudakov's inequality while keeping the L_p structure in place and without resorting to the fact that all the L_p norms of $\langle G, t \rangle$ are equivalent to its L_2 norm.

Observe that if $T' \subset T$ is ε-separated in ℓ_2, then $\{\langle G, t\rangle : t \in T'\}$ is $\varepsilon\sqrt{p}$ separated subset in L_p. Hence, identifying T with the set of linear functionals $\{\langle t, \cdot\rangle : t \in T\}$, the gaussian version of Sudakov's inequality can be reformulated as:

$$\sup_{\varepsilon>0} \varepsilon \log^{1/2} \mathcal{M}(T, \varepsilon\sqrt{p}B(L_p)) = \sup_{\delta>0} \frac{\delta}{\sqrt{p}} \log^{1/2} \mathcal{M}(T, \delta B(L_p)) \leq c\ell_*(T). \tag{12.6}$$

Exercise 40. *With* (12.6)*, show that the following is an equivalent formulation of Sudakov's inequality: There exist an absolute constant c such that for every $p \geq 2$, if $|T'| = \exp(p)$ is a δ-separated subset of T, then $\delta \leq c\ell_*(T)$.*

The version of Sudakov's inequality from Exercise 40 actually holds for Bernoulli processes and for other processes as well. In the Bernoulli case one has:

Theorem 12.7. *There exists an absolute constant for which the following holds. Let $V \subset \mathbb{R}^d$, $\delta > 0$ and $p \geq 1$ such that $A_V = \{\langle \mathcal{E}, v\rangle : v \in V\}$ is δ-separated in L_p and $|V| \geq \exp(p)$. Then*

$$\mathbb{E} \sup_{v\in V} \left| \sum_{i=1}^d \varepsilon_i v_i \right| \geq c\delta.$$

The proof of Theorem 12.7 requires some preparation. Its main component is an embedding lemma that allows one to revert to a gaussian argument:

Lemma 12.8. *There are absolute constants c_0 and c_1 for which the following holds. Let V be as in Theorem 12.7. There is an integer m and a map $\Phi : V \to \ell_2^m \cap \ell_\infty^m$ such that $W = \Phi(V)$ satisfies:*

(1) *W is $c_0\delta/\sqrt{p}$-separated with respect to the ℓ_2 norm;*

(2) *$W \subset (\delta/p)B_\infty^m$; and*

(3) $\mathbb{E}\sup_{w\in W} |\sum_{i=1}^m \varepsilon_i w_i| \leq c_1 \mathbb{E}\sup_{v\in V} |\sum_{i=1}^d \varepsilon_i v_i|$.

The idea behind Lemma 12.8 is that Φ 'spreads' the vectors $v \in V$ in a way that ensures that the image of each vector is well-bounded coordinate-wise, and at the same time, the separation in L_p of the A_V is transformed to separation in ℓ_2. Moreover, the Bernoulli average of the resulting set W is dominated by the Bernoulli averages of the original set V, implying that it suffices to lower bound the former. The intuitive reason why such a lower bound is possible is due to (1) and (2): vectors in W are both well-separated and well-spread, which indicates that the Bernoulli average of W should not be too far from the gaussian one, allowing one to invoke Sudakov's inequality for gaussian processes. More accurately, assuming that Lemma 12.8 is true, the first step in the proof of Theorem 12.7 is the following:

Theorem 12.9. *For every $\kappa > 0$ there exists a constant $c(\kappa)$ for which the following holds. Let $W \subset \mathbb{R}^m$ be a bounded, ε-separated in ℓ_2 which also satisfies that*

$$\sup_{w\in W} \|w\|_\infty \leq \kappa \frac{\varepsilon}{\sqrt{\log |W|}}.$$

Then

$$\mathbb{E} \sup_{w\in W} \left| \sum_{i=1}^m \varepsilon_i w_i \right| \geq c(\kappa)\varepsilon\sqrt{\log |W|}.$$

In other words, a standard Sudakov bound happens to be true for the Bernoulli process as long as the indexing set is well bounded in ℓ_∞. Thus, in such cases, the Bernoulli process "behaves" as if it were gaussian.

The combination of Lemma 12.8 and Theorem 12.9 leads to the proof of Theorem 12.7:

Proof of Theorem 12.7. Recall that $W = \Phi(V) \subset \ell_2^m$, and without loss of generality one can assume that $|V| = \exp(p)$. By parts (1) and (2) of Lemma 12.8, $|W|$ is ε-separated in ℓ_2^m for $\varepsilon = c_0\delta/\sqrt{p}$ and

$$\max_{w\in W} \|w\|_\infty \leq \frac{\delta}{p} \leq \kappa \frac{\varepsilon}{\sqrt{\log|W|}}$$

for $\kappa = c_0^{-1}$. Thus, $W = \Phi(V)$ satisfied the conditions of Theorem 12.9, implying that

$$\mathbb{E}\sup_{w\in W}\left|\sum_{i=1}^m \varepsilon_i w_i\right| \geq c(\kappa)\varepsilon\sqrt{\log|W|} = c_1\sqrt{\delta}. \tag{12.7}$$

Moreover, by part (3) of Lemma 12.8,

$$\mathbb{E}\sup_{v\in V}\left|\sum_{i=1}^d \varepsilon_i v_i\right| \geq c_2\mathbb{E}\sup_{w\in W}\left|\sum_{i=1}^m \varepsilon_i w_i\right|,$$

which, combined with (12.7) completes the proof. ∎

Next, let us turn to the proofs of the two components, which are of independent interest in their own right.

12.2.1 The supremum of Bernoulli processes for bounded sets

As mentioned previously, the idea behind the proof of Theorem 12.9 is that if W is a separated set in ℓ_2^m that is also well bounded in ℓ_∞^m then the random variables $\sum_{i=1}^m \varepsilon_i w_i$ behave as if they were gaussian because the vectors $(v_i - w_i)_{i=1}^m$ are 'well-spread': their ℓ_2^m norm is large, but their ℓ_∞^m is relatively small.

Making this intuition more precise requires two preliminary steps. Let $X_1, ..., X_m$ be iid copies of a symmetric random variable X. Set $x > 0$ and $u \in \mathbb{R}^m$ and consider the random variable

$$Q_u^t = \sum_{i=1}^m u_i X_i \mathbb{1}_{\{|X_i|>t\}}.$$

It stands to reason that if X has a reasonable tail decay, that should be reflected in the behaviour of Q_u^t as a function of t. Indeed, in the extreme case, when X is bounded, say by 1, $Q_u^t \equiv 0$ if $t \geq 1$. Quantifying the effect that increasing t has on the tail behaviour of Q_u^t is studied in the next lemma.

Lemma 12.10. *There exist absolute constants c and c_0 for which the following holds. Let X be a symmetric random variable that satisfies $\|X\|_{\psi_1} \leq L$ and let $X_1, ..., X_m$ be independent copies of X. Then for $u \in \mathbb{R}^m$ and every $t, x > 0$*

$$Pr\left(|Q_u^t| \geq x\right) \leq 2\exp\left(-c\min\left\{\frac{x^2}{\|u\|_2^2 L^2\lambda^2(t)}, \frac{x}{L\|u\|_\infty}\right\}\right), \tag{12.8}$$

where $\lambda(t) = Pr^{1/8}(|X| \geq t)$.

Moreover, for any $p \geq 1$,

$$\|Q_u^t\|_{L_p} \leq c_0 L(\sqrt{p}\|u\|_2 \lambda(t) + p\|u\|_\infty).$$

The implication of Lemma 12.10 is that the fast tail decay of X "helps" the subgaussian part of the tail of Q_u^t.

Proof. Fix $p \geq 1$ and let $Y_i = u_i X_i \mathbb{1}_{\{|X_i| \geq t\}}$, which is a symmetric random variable. Using the notation of Bernstein's inequality (Theorem 3.7),

$$\begin{aligned}\mathbb{E}Y_i^p =& \mathbb{E}(u_i X_i)^{p-2} \cdot (u_i X_i)^2 \mathbb{1}_{\{|X| \geq t\}} \leq \left(\mathbb{E}(u_i X_i)^{2(p-2)}\right)^{1/2} \cdot \left(\mathbb{E}(u_i X_i)^4 \mathbb{1}_{\{|X_i| \geq t\}}\right)^{1/2} \\ \leq& \left(\mathbb{E}(u_i X_i)^{2(p-2)}\right)^{1/2} \cdot \left(\mathbb{E}(u_i X_i)^8\right)^{1/4} \cdot Pr^{1/4}(|X| \geq t) \\ \leq& (cL)^{p-2}\|u\|_\infty^{p-2} p! \cdot L^2 u_i^2 Pr^{1/4}(|X| \geq t)\end{aligned}$$

for a suitable absolute constant c, because $\|X\|_{L_q} \lesssim Lq$ and since $p! \sim \sqrt{2\pi p}(p/e)^p$. Therefore, one may set

$$M = cL\|u\|_\infty, \quad \text{and} \quad \sigma_i^2 = L^2 u_i^2 Pr^{1/4}(|X| \geq t).$$

The first part of the claim follows from Theorem 3.7, and the second one from the connection between moments and tail estimates. ∎

The estimate on $\|Q_u^t\|_{L_p}$ immediately leads to the following outcome:

Corollary 12.11. *Let* $U \subset \mathbb{R}^m$ *be a finite set and let* X *be as in Lemma 12.10. Then*

$$\mathbb{E}\max_{u \in U} |Q_u^t| \leq c(L) \max_{u \in U} \left(\lambda(t)\|u\|_2 \sqrt{\log |U|} + \|u\|_\infty \cdot \log |U|\right).$$

Proof. Let $p = \log |U|$. Recall that if $a \in \mathbb{R}^d$ then $\|a\|_\infty \leq \|a\|_{\log d}$; therefore, by Jensen's inequality and the second part of Lemma 12.10,

$$\begin{aligned}\mathbb{E}\max_{u \in U} |Q_u^t| \leq& e\mathbb{E}\left(\sum_{u \in U} |Q_u^t|^p\right)^{1/p} \leq \left(\sum_{u \in U} \mathbb{E}|Q_u^t|^p\right)^{1/p} \\ \leq& |U|^{1/p} \cdot CL \max_{u \in U} (\sqrt{p}\|u\|_2 \lambda(t) + p\|u\|_\infty) \\ =& cL \max_{u \in U} \left(\lambda(t)\|u\|_2 \sqrt{\log |U|} + \|u\|_\infty \log |U|\right).\end{aligned}$$

∎

Let us now prove a Sudakov type inequality indexed by a set that is both bounded in ℓ_∞^m and in ℓ_2^m.

Lemma 12.12. *There exist constants* κ *and* c *for which the following holds. Let* $\rho > 0$ *and* $U \subset 2\rho B_2^m \cap \theta B_\infty^m$ *where* $\theta \leq \kappa\rho/\sqrt{\log |U|}$. *If* U *is* ρ*-separated with respect to the* ℓ_2 *norm then*

$$\mathbb{E}\sup_{u \in U} \left|\sum_{i=1}^m \varepsilon_i u_i\right| \geq c\rho\sqrt{\log |U|}.$$

Proof. First, observe that by Sudakov's inequality there is an absolute constant c_1 such that

$$\mathbb{E}\sup_{u\in U}|\sum_{i=1}^m g_i u_i| \geq c_1\rho\sqrt{\log|U|}.$$

On the other hand, fix $x>0$ and apply Lemma 12.10 for $X=g$—the standard gaussian random variable. One has that for any $t>0$,

$$\begin{aligned}\mathbb{E}\sup_{u\in U}|\sum_{i=1}^m g_i u_i| \leq& \mathbb{E}\sup_{u\in U}|\sum_{i=1}^m \varepsilon_i g_i \mathbb{1}_{\{|g_i|\leq t\}}u_i| + \mathbb{E}\sup_{u\in U}|\sum_{i=1}^m g_i\mathbb{1}_{\{|g_i|\geq t\}}u_i| \\ \leq& t\mathbb{E}\sup_{u\in U}|\sum_{i=1}^m \varepsilon_i u_i| + \mathbb{E}\sup_{u\in U}|\sum_{i=1}^m g_i\mathbb{1}_{\{|g_i|\geq t\}}u_i|,\end{aligned}$$

where the last inequality is an outcome of the contraction principle for Bernoulli processes, conditioned on $g_1,...,g_m$. Therefore, it suffices to shows that

$$\frac{c_1}{2}\rho\sqrt{\log|U|} \geq \mathbb{E}\sup_{u\in U}|\sum_{i=1}^m g_i\mathbb{1}_{\{|g_i|\geq t\}}u_i|, \tag{12.9}$$

to ensure that

$$\mathbb{E}\sup_{u\in U}|\sum_{i=1}^m g_i u_i| \leq 2t\mathbb{E}\sup_{u\in U}|\sum_{i=1}^m \varepsilon_i u_i|,$$

which, in turn, implies

$$\mathbb{E}\sup_{u\in U}|\sum_{i=1}^m \varepsilon_i u_i| \geq \frac{c_1}{2t}\rho\sqrt{\log|U|},$$

as required.

To establish (12.9) one uses Corollary 12.11: recall that

$$\max_{u\in U}\|u\|_2 \leq 2\rho \quad \text{and} \quad \max_{u\in U}\|u\|_\infty \leq \theta.$$

Set t to satisfy that $\lambda(t)\leq c_1/(8c_2)$, where c_2 is the constant from Corollary 12.11. Since X is a standard gaussian random variable, t can be taken to be an absolute constant. Thus,

$$\begin{aligned}\mathbb{E}\sup_{u\in U}|\sum_{i=1}^m g_i\mathbb{1}_{\{|g_i|\geq t\}}u_i| \leq& c_2\max_{u\in U}\left(\lambda(t)\|u\|_2\sqrt{\log|U|}+\|u\|_\infty\cdot\log|U|\right) \\ \leq& \frac{c_1}{8}\rho\sqrt{\log|U|}+c_2\theta\log|U| \leq \frac{c_1}{4}\rho\sqrt{\log|U|}\end{aligned}$$

provided that $\theta\leq c_3\rho/\sqrt{\log|U|}$ for a suitable absolute constant c_3. ∎

Proof of Theorem 12.9. Let W be ε separated in ℓ_2^m, set κ to be as in Lemma 12.12 and recall that

$$W\subset c\frac{\kappa\varepsilon}{\sqrt{\log|W|}}B_\infty^m.$$

The first observation is that for any $\rho\geq\varepsilon/2$

$$\sup_{w\in W}\log\mathcal{M}\left((W-w)\cap 2\rho B_2^m,\rho B_2^m\right) \leq c_2\rho^{-2}\mathbb{E}\sup_{w\in W}\left|\sum_{i=1}^m\varepsilon_i w_i\right|. \tag{12.10}$$

Indeed, fix $w \in W$ and let $U \subset (W-w) \cap 2\rho B_2^m$ be a ρ-separated set for $\rho \geq 2\varepsilon$. Therefore,

$$\sup_{u\in U} \|u\|_\infty \leq 2 \sup_{w\in W} \|w\|_\infty \leq 2\frac{\kappa\varepsilon}{\sqrt{\log|W|}} \leq \kappa\frac{\rho}{\sqrt{\log|U|}},$$

and the conditions of Lemma 12.12 hold. Applying that lemma for the set U, it is evident that

$$\log|U| \leq c\rho^{-2}\left(\mathbb{E}\sup_{u\in U}\left|\sum_{i=1}^m \varepsilon_i u_i\right|\right)^2$$

and (12.10) is valid because each $u \in U$ is of the form $u = w' - w$ for some $w' \in W$; thus

$$\mathbb{E}\sup_{u\in U}\left|\sum_{i=1}^m \varepsilon_i u_i\right| \leq 2\mathbb{E}\sup_{w\in W}\left|\sum_{i=1}^m \varepsilon_i W_i\right|.$$

Next, let us use (12.10) iteratively: denote by $R = \sup_{w\in W}\|w\|_2$ and let $W_1 \subset W$ be a maximal $R/2$-separated subset of W. Clearly,

$$\mathcal{M}(W, \varepsilon B_2^m) \leq \mathcal{M}(W \cap RB_2^m, (R/2)B_2^m) \cdot \max_{w_1\in W_1} \mathcal{M}\left((W-w_1)\cap(R/2)B_2^m, \varepsilon B_2^m\right).$$

Continuing in the same fashion,

$$\mathcal{M}(W, \varepsilon B_2^m) \leq \prod_{k=1}^{\ell} \max_{w\in W} \mathcal{M}\left(W \cap (R/2^{k-1})B_2^m, (R/2^k)B_2^m\right), \tag{12.11}$$

where $R/2^\ell \leq \varepsilon \leq R/2^{\ell-1}$, and in particular, for each $\rho = R/2^k$ one has that $\rho \geq \varepsilon/2$. Hence, combining (12.11) with (12.10),

$$\begin{aligned}\log \mathcal{M}\left(W, \varepsilon B_2^m\right) \leq& \sum_{k=1}^{\ell} \max_{w\in W} \log \mathcal{M}\left(W \cap (R/2^{k-1})B_2^m, (R/2^k)B_2^m\right)\\ \leq& c\varepsilon^2\left(\mathbb{E}\sup_{w\in W}|\sum_{i=1}^m \varepsilon_i w_i|\right)^2.\end{aligned}$$

■

12.2.2 Spreading a set using chopping maps

Now let us turn to the second component needed for the proof of Theorem 12.7, the construction of the embedding Φ that maps V to ℓ_2^m. The construction is based on the idea of chopping maps, introduced in [**?**] and which play a central in the solution of the Bernoulli conjecture in [**?**].

We only define the maps that are used in the proof of Theorem 12.7 and refer the reader to [**?**] for a more detailed exposition on the topic of chopping maps.

Definition 12.13. *Let $\Delta > 0$. For every $k = \pm 1, \pm 2, \ldots..$ define a function ϕ_k as follows: if $k > 0$ and $x \geq 0$ set*

$$\phi_k(x) = \begin{cases} 0 & \text{if } x \in [0, k\Delta), \\ x - \Delta k & \text{if} x \in [k\Delta, (k+1)\Delta), \\ \Delta & \text{if } x \in [(k+1)\Delta, \infty), \end{cases}$$

and if $x < 0$ set $\phi_k(x) = 0$. If $k < 0$ define $\phi_k(x) = -(\phi_{-k}(-x))$ for any $x \in \mathbb{R}$.

Thus, each $\phi_k(x)$ is a 1-Lipschitz function. The functions ϕ_k are called chopping maps because that is what they do: each $x \in \mathbb{R}$ is 'broken' into pieces of size Δ. This can be recorded in the natural vector $(\phi_k(x))_k$: for example, if $x > 0$ then writing $x = m\Delta + \alpha$ for $m \geq 0$ and $0 \leq \alpha < \Delta$, it is evident that

$$(\phi_k(x))_{k\neq 0} = \left(\underbrace{0, ..., 0}_{k<0}, \underbrace{\Delta, ..., \Delta}_{k=1,...,m}, \underbrace{\alpha}_{k=m+1}, \underbrace{0, ...}_{k>m+1} \right).$$

From here on we set for $x \in \mathbb{R}$,

$$\Phi(x) = (\phi_k(x))_{k\neq 0},$$

where we omit the dependence of the chopping maps on the parameter Δ.

The chopping maps naturally define a transformation, mapping finite subsets of $\mathbb{R}^d$ to a high-dimensional space in the following way:

Definition 12.14. *Let $\Phi(x) = (\phi_k(x))_{k\neq 0}$ and for $T \subset \mathbb{R}^d$ set*

$$\Phi(T) = \left\{ (\Phi(t_i))_{i=1}^d : (t_i)_{i=1}^d \in T \right\}.$$

Note the obvious fact that the coordinates of $\Phi(T)$ are all bounded by Δ. With Lemma 12.8 in mind, the choice of Δ has to be $\sim \delta/p$.

The two crucial features of the mapping Φ are:

- If $T \subset \mathbb{R}^d$ is $\sim \delta$ separated with respect to the $L_p(\mathcal{E})$ norm then for the right choice of Δ (again, $\Delta \sim \delta/p$, then $\Phi(T)$ is separated in ℓ_2 at scale $\sim \delta/\sqrt{p}$.
- For an arbitrary $T \subset \mathbb{R}^d$, the Bernoulli process indexed by $\Phi(T)$ is 'smaller' than the Bernoulli process indexed by T.

In other words, if one is interested in lower bounds on Bernoulli processes, one possibility is to study the process indexed by $\Phi(T)$ (defined explicitly in what follows). The added value is that $\Phi(T)$ is 'more regular' than T, because its coordinates are all bounded by Δ. And, as a preliminary indication that the resulting bound is not trivial, separation (in some sense) in T is inherited by $\Phi(T)$.

Of course, all this is shameless hand waiving. Let us turn to a more accurate description of the mapping Φ, beginning with a straightforward (yet a little tedious) observation:

Lemma 12.15. *Let $\Delta > 0$ and set $(\phi_k)_{k\neq 0}$ be the corresponding chopping maps. Then for $x, y \in \mathbb{R}$,*

$$\frac{1}{10}\|\Phi(x) - \Phi(y)\|_2^2 \leq \Delta|x-y|\mathbb{1}_{\{|x-y|>\Delta\}} + |x-y|^2\mathbb{1}_{\{|x-y|<\Delta\}} \leq 10\|\Phi(x) - \Phi(y)\|_2^2.$$

We omit the details of the proof which is based on a case-by-case analysis. To give an illustration of why the statement is correct in one case, assume that $x = m\Delta + \alpha$ and $y = n\Delta + \beta$ for $m \geq n > 0$ and $0 < \alpha, \beta < \Delta$. If $m - n \geq 3$ then

$$\frac{m-n}{2}\Delta \leq |x-y| \leq 2(m-n)\Delta,$$

implying that $\Delta|x-y|\mathbb{1}_{\{|x-y|>\Delta\}} + |x-y|^2\mathbb{1}_{\{|x-y|<\Delta\}} = \Delta|x-y|$, and

$$\frac{1}{2}(m-n)\Delta^2 \leq \Delta|x-y| \leq 2(m-n)\Delta^2.$$

On the other hand

$$\Phi(x) = \left(\underbrace{0,...,0}_{k<0}, \underbrace{\Delta,...,\Delta}_{k=1,...,m}, \underbrace{\alpha,}_{k=m+1}, \underbrace{0,....}_{k>m+1}\right).$$

$$\Phi(y) = \left(\underbrace{0,...,0}_{k<0}, \underbrace{\Delta,...,\Delta}_{k=1,...,n}, \underbrace{\beta}_{k=n+1}, \underbrace{0,...,}_{k>n+1}\right),$$

and thus

$$\|\Phi(x)-\Phi(y)\|_2^2 = (\Delta-\beta)^2 + (m-n-1)\Delta^2 + \alpha^2.$$

Since $m-n-1 \geq (m-n)/3$ if follows that

$$\|\Phi(x)-\Phi(y)\|_2^2 \geq (m-n-1)\Delta^2 \geq \frac{m-n}{3}\Delta^2$$

and

$$\|\Phi(x)-\Phi(y)\|_2^2 \leq (m-n)\Delta^2 + 2\Delta^2 \leq 2(m-n)\Delta^2,$$

proving the wanted equivalence.

Verifying the other cases is equally simple. ∎

Let us turn to the question of separation and the way the vector-valued mapping Φ preserves it.

Lemma 12.16. *There exists an absolute constant c for which the following holds. If $x, y \in \mathbb{R}^d$ satisfy that*

$$\sum_{i=1}^{p}(x-y)_i^* + \sqrt{p}\left(\sum_{i>p}((x-y)_i^*)^2\right)^{1/2} \geq \delta,$$

then for $\Delta = \delta/4p$ one has that

$$\|\Phi(x)-\Phi(y)\|_2 \geq c\delta/\sqrt{p}.$$

In particular, if x, y are δ-separated with respect to the $L_p(\mathcal{E})$ norm endowed on $\mathbb{R}^d$, that is exhibited by the ℓ_2 distance between $\Phi(x)$ and $\Phi(y)$, as long as the scale of the chopping maps is selected wisely.

Proof. Applying Lemma 12.15 and the fact that

$$\|\Phi(x)-\Phi(y)\|_2^2 = \sum_{i=1}^{d}\left(\Phi(x_i)-\Phi(y_i)\right)^2,$$

it suffices to show that for $\Delta = \delta/4p$ we have

$$\sum_{i=1}^{d}\Delta|x_i-y_i|\mathbb{1}_{\{|x_i-y_i|>\Delta\}} + |x_i-y_i|^2\mathbb{1}_{\{|x_i-y_i|<\Delta\}} \gtrsim \frac{\delta^2}{p}.$$

Without loss os generality assume that $(|x_i - y_i|)_{i=1}^d$ is non-increasing and let us consider two cases. First, if $\sum_{i=1}^p |x_i - y_i| \geq \delta/2$, then

$$\sum_{i=1}^p |x_i - y_i| \mathbb{1}_{\{|x_i - y_i| \geq \delta/4p\}} \geq \frac{\delta}{4};$$

indeed, $\sum_{i=1}^p |x_i - y_i| \mathbb{1}_{\{|x_i - y_i| < \delta/4p\}} \leq p \cdot (\delta/4p) = \delta/4$. Therefore,

$$\Delta \cdot \sum_{i=1}^p |x_i - y_i| \mathbb{1}_{\{|x_i - y_i| \geq \delta/4p\}} \geq \frac{\delta}{4p} \cdot \frac{\delta}{4} = \frac{\delta^2}{16p},$$

as required.

If, on the other hand, $\sum_{i=1}^p |x_i - y_i| \leq \delta/2$ then by monotonicity,

$$|x_p - y_p| \leq \frac{1}{p} \sum_{i=1}^p |x_i - y_i| \leq \frac{\delta}{2p}.$$

And, since by the separation condition

$$\sum_{i=1}^p |x_i - y_i| + \sqrt{p} \left(\sum_{i>p} (x_i - y_i)^2 \right)^{1/2} \geq \delta,$$

it is evident that

$$p \sum_{i>p} (x_i - y_i)^2 \geq \delta^2 - \frac{\delta^2}{4} \geq \frac{\delta^2}{2}.$$

Let $I = \{i > p : |x_i - y_i| \geq \delta/4p\}$. If $|I| \geq p/4$ then again,

$$\Delta \sum_{i \in I} |x_i - y_i| \geq \frac{\delta}{4p} \frac{\delta}{4p} \frac{p}{4} = \frac{\delta^2}{64p};$$

otherwise, $|I| \leq p/4$, implying that

$$\sum_{i \in I} |x_i - y_i|^2 \leq |x_p - y_p|^2 |I| \leq \frac{\delta^2}{4p^2} \cdot \frac{p}{4} = \frac{\delta^2}{16p},$$

and thus on the complement I^c in $\{p+1, ..., d\}$ we have $|x_i - y_i| \leq \delta/4p$ and

$$\sum_{i \in I^c} |x_i - y_i|^2 \geq \frac{\delta^2}{4p} - \frac{\delta^2}{16p} \geq \frac{\delta^2}{16p};$$

thus,

$$\sum_{i \in I^c} |x_i - y_i|^2 \mathbb{1}_{\{|x_i - y_i| \leq \delta/4p\}} \geq \frac{\delta^2}{16},$$

completing the proof. ∎

Finally, let us show that the Bernoulli process indexed by T dominates the Bernoulli process indexed by $\Phi(T)$, which is the final component needed for the proof of Theorem 12.7.

Recall that

$$\Phi(T) = \left\{ (\Phi(t_i))_{i=1}^d : (t_i)_{i=1}^d \in T \right\},$$

and view each $\Phi(t) \in \Phi(T)$ as a vector indexed by pairs (i,k) for $i \in \{1,...,d\}$ and $k \in \mathbb{Z}\backslash\{0\}$. This set of indices is denoted by Λ.

Let $\{\varepsilon_{i,k} : (i,k) \in \Lambda\}$ and $(\varepsilon_i')_{i=1}^d$ be independent Bernoulli random variables. Thus, $(\varepsilon_{i,k})_{(i,k)\in\Lambda}$ and $(\varepsilon_i'\varepsilon_{i,k})_{(i,k)\in\Lambda}$ have the same distribution. With that in mind, set

$$Z_{\Phi(t)} = \sum_{(i,k)\in\Lambda} \varepsilon_i'\varepsilon_{i,k}\phi_k(t_i) = \sum_{i=1}^d \varepsilon_i' \left(\sum_{k\neq 0} \varepsilon_{i,k}\phi_k(t_i) \right)$$

and note that $\mathbb{E}\sup_{t\in T} Z_{\Phi(t)}$ is the expectation of the supremum of the Bernoulli process indexed by $\Phi(T)$.

Observe that there exists an absolute constant c for which

$$\mathbb{E}\sup_{t\in T} \left| \sum_{i=1}^d \varepsilon_i t_i \right| \geq c\mathbb{E}\sup_{t\in T} |Z_{\Phi(t)}|.$$

Indeed, fix a realization of $(\varepsilon_{i,k})_{(i,k)\in\Lambda}$ and let $h_i(x) = \sum_{k\neq 0} \varepsilon_{i,k}\phi_k(x)$. Thus,

$$\mathbb{E}\sup_{t\in T} |Z_{\Phi(t)}| = \mathbb{E}\sup_{t\in T} \left| \sum_{i=1}^d \varepsilon_i' h_i(t_i) \right|.$$

For $x, y \in \mathbb{R}$, $|h_i(x) - h_i(y)| \leq 2|x-y|$ and $h(0) = 0$; therefore, by the contraction inequality for Bernoulli processes,

$$\mathbb{E}\sup_{t\in T} |Z_{\Phi(t)}| = \mathbb{E}\sup_{t\in T} \left| \sum_{i=1}^d \varepsilon_i' h_i(t_i) \right| \leq 2\mathbb{E}\sup_{t\in T} \left| \sum_{i=1}^d \varepsilon_i' t_i \right|.$$

■

Exercise 41. *Prove that the process $t \to Z_{\Phi(t)}$ is well defined.*

Chapter 13

The Combinatorial Dimension

In addition to the covering numbers and to the Bernoulli/gaussian mean-widths of the class, there is another way of measuring how "rich" a set is, that is extensively used in statistical learning theory. In its simplest from, the *combinatorial dimension* captures the largest dimension of a combinatorial cube that can be found in a coordinate projection of a binary valued class of functions F.

This simple version of the combinatorial dimension was introduced by Vapnik and Chervonenkis, and is known as the VC dimension.

Definition 13.1. *Let F be a class of $\{0,1\}$-valued functions on a space Ω. The class F shatters $\{x_1,...,x_n\} \subset \Omega$, if for every $I \subset \{1,...,n\}$ there is a function $f_I \in F$ for which $f_I(x_i) = 1$ if $i \in I$ and $f_I(x_i) = 0$ if $i \notin I$. Let*

$$\mathrm{VC}(F,\Omega) = \sup\{|A| \ : \ A \subset \Omega, \ A \text{ is shattered by } F\}.$$

$\mathrm{VC}(F,\Omega)$ is the VC dimension of F, but when the underlying space is clear we denote it by $\mathrm{VC}(F)$.

It is easy to see that $\sigma = \{x_1,...,x_n\}$ is shattered if

$$P_\sigma F = \{(f(x_i))_{i=1}^n \ : \ f \in F\} = \{0,1\}^n.$$

Thus, The VC dimension is the largest cardinality of $\sigma \subset \Omega$ for which $P_\sigma F$ is as big as it can be: the combinatorial cube of dimension $|\sigma|$.

The real-valued counterpart of the VC dimension, called the combinatorial dimension, is defined by extending the notion of a shattered set. Unlike the binary-valued case, when every cube found in $P_\sigma F$ is the combinatorial cube of the appropriate dimension, in the real-valued case there is a tradeoff between the 'size' of the cube and the dimension.

Definition 13.2. *For $\varepsilon > 0$, a set $\sigma = \{x_1,...,x_n\} \subset \Omega$ is ε-shattered by F if there is a function $s : \sigma \to \mathbb{R}$, which satisfies that for every $I \subset \{1,...,n\}$ there is some $f_I \in F$ for which $f_I(x_i) \geq s(x_i) + \varepsilon$ if $i \in I$, and $f_I(x_i) \leq s(x_i) - \varepsilon$ if $i \notin I$. Let*

$$\mathrm{VC}(F,\Omega,\varepsilon) = \sup\{|\sigma| \ : \ \sigma \subset \Omega, \ \sigma \text{ is } \varepsilon\text{−shattered by } F\}.$$

f_I is called the shattering function of the subset I and the vector $(s(x_i))_{i=1}^n$ is a witness to the ε-shattering. In cases where the underlying space is clear we denote the combinatorial dimension by $\mathrm{VC}(F,\varepsilon)$.

There is a clear geometric interpretation of shattering, implying that if σ is ε-shattered by F then $P_\sigma F$ contains a 'cubic structure'. Indeed, consider the "cell"

$$(s(x_i))_{i=1}^n + \varepsilon B_\infty^n \subset \mathbb{R}^n.$$

The $2n$ hyperplanes supporting the $n-1$ dimensional facets of the cell define 2^n 'quadrants' in $\mathbb{R}^n$, with each quadrant corresponding to a vector signs $(\eta_i)_{i=1}^n \in \{-1,1\}^n$. A vector $z \in \mathbb{R}^n$ belongs to the quadrant defined by $(\eta_i)_{i=1}^n$ if $z_i \geq s(x_i) + \varepsilon$ when $\eta_i = 1$, and $z_i \leq s(x_i) - \varepsilon$ otherwise. The set σ is shattered by F with the witness $(s(x_i))_{i=1}^n$ if $P_\sigma F$ contains at least one point in each quadrant. Using this geometric picture it is clear that $\mathrm{conv}(P_\sigma F)$ contains $(s(x_i))_{i=1}^n + \varepsilon B_\infty^n$.

Observe that if F is convex and centrally symmetric and if $\sigma = \{x_1, ..., x_n\}$ is ε-shattered by F then one may take $s(x_i) = 0$ and

$$\varepsilon B_\infty^n \subset P_\sigma F.$$

Indeed, if σ is ε-shattered with a witness $(s(x_i))_{i=1}^n$, then for every $I \subset \{1, ..., n\}$ there is some f_I that satisfies $f_I(x_i) \geq s(x_i) + \varepsilon$ if $i \in I$ and $f_I(x_i) \leq s(x_i) - \varepsilon$ for $i \in I^c$. For every such I, let $g_I = (f_I - f_{I^c})/2$, which clearly belongs to F because the class is convex and centrally symmetric. It follows that $\varepsilon B_\infty^\sigma \subset P_\sigma F$, as claimed.

In particular, if $\{x_1, ..., x_n\}$ is ε-shattered by F then $\varepsilon B_\infty^n \subset \mathrm{absconv}(F)$.

13.1 Metric entropy and the VC dimension

The next lemma, known as the Sauer-Shelah Lemma was proved independently at least three times, by Sauer [**?**], Shelah [**?**] and Vapnik and Chervonenkis [**?**], and then generalized by Karpovsky and Milman [**?**].

Lemma 13.3. *Let F be a class of $\{0,1\}$-valued functions and set $d = \mathrm{VC}(F)$. Then, for every finite subset $\sigma \subset \Omega$ of cardinality n,*

$$|P_\sigma F| \leq \left(\frac{en}{d}\right)^d.$$

In particular, for every $0 < \varepsilon \leq 1/2$,

$$N(\varepsilon, F, L_\infty(\sigma)) = |P_\sigma F| \leq (en/d)^d.$$

The proof of Lemma 13.3 uses the notion of a hereditary class of sets.

Definition 13.4. *Let $\mathcal{U}$ be a class of subsets of Ω. The class is hereditary if for every $A \in \mathcal{U}$ and $B \subset A$, then $B \in \mathcal{U}$.*

An natural example of a hereditary class of subsets of $\{1, ..., n\}$ is the d-cross:

$$\{U \subset \{1, ..., n\} \ : \ |U| \leq d\}.$$

Proof. We identify $P_\sigma F$ with a class of sets $\mathcal{U}$, each belongs to $\{1, ..., n\}$ in the natural way: each $P_\sigma f$ is of the form $f = \mathbb{1}_A$ for $A \subset \{1, ..., n\}$.

The assertion of Lemma 13.3 is immediate if $\mathcal{U}$ is hereditary. Indeed, a hereditary class of subsets of $\{1,...,n\}$ whose VC dimension is at most d is a subset of the d-cross. And, the cardinality of the d-cross is

$$\sum_{i=1}^{d}\binom{n}{i} \leq \left(\frac{en}{d}\right)^d.$$

Therefore, the proof will be complete by constructing a one-to-one map from $\mathcal{U}$ to a hereditary class $\mathcal{U}'$ whose VC dimension is at most d.

Define a family of operators T_x in the following way: $T_x(U) = U\backslash\{x\}$ if $x \in U$ and $U\backslash\{x\} \notin \mathcal{U}$; otherwise $T_x(U) = U$. Thus, when viewing $\mathcal{U}$ as a subset of $\{0,1\}^n$, the operator T_x makes $\mathcal{U} \subset \{0,1\}^n$ more hereditary by 'pushing down' points in the combinatorial cube when there is an empty slot 'below' them.

To see that T_x is one-to-one, assume that $T_x(U_1) = T_x(U_2)$. If $T_x(U_i) \neq U_i$ for $i = 1, 2$, then $U_i = T_x(U_i) \cup \{x\}$ and thus $U_1 = U_2$. Otherwise, if $T_x(U_1) = U_1$ and $T_x(U_2) \neq U_2$, then $U_1 = T_x(U_2)$ and $U_2 = T_x(U_2) \cup \{x\}$, which is impossible, because $U_2\backslash\{x\} \in \mathcal{U}$, and that would imply that $T_x(U_2) = U_2$.

Next, let us show that if I is shattered by $T_x(\mathcal{U})$ then it is also shattered by $\mathcal{U}$. In particular, that means that the VC dimension cannot increase by the application of T_x.

Consider two cases. First, let $x \notin I$, and in which case, for every $U \in \mathcal{U}$, $U \cap I = T_x(U) \cap I$. Thus, if I is shattered by $\{T_x(U) : U \in \mathcal{U}\}$ then I is shattered by $\mathcal{U}$.

Next, assume that $x \in I$, and let $I' \subset I\backslash\{x\}$. Let us show that both I' and $I' \cup \{x\}$ belong to $\mathcal{U}$, implying that I is shattered by $\mathcal{U}$. Indeed, recall that $T_x(\mathcal{U})$ shatters I, and therefore $I', I' \cup \{x\} \in T_x(\mathcal{U})$. By the definition of T_x, $I' \cup \{x\} \in \mathcal{U}$ and since $x \notin I'$ it is evident that $I' \in \mathcal{U}$ as well—showing that the VC dimension does not increase with the application of T_x.

To complete the proof, let $\mathcal{V}$ be a class of subsets of $\{1,...,n\}$ and set

$$H(\mathcal{V}) = \sum_{V \in \mathcal{V}} |V|.$$

Let $\mathcal{T}$ be the set of finite compositions of operators of the form T_x. Since $H(T_x\mathcal{V}) \leq H(\mathcal{V})$, one has that $\inf_{\tilde{T}\in\mathcal{T}} H\left(\tilde{T}(\mathcal{U})\right)$ is attained. The minimizer $\tilde{T}(\mathcal{U})$ is a hereditary class of sets: for every $x \in \{1,...,n\}$, and every $V \in \tilde{T}(\mathcal{U})$, if $x \in V$ then $V\backslash\{x\} \in \tilde{T}(\mathcal{U})$ – otherwise, one may apply T_x and decrease H even further. ∎

Although the L_∞ entropy estimate depends on n and thus on the dimension (cardinality) of the coordinate projection, it is possible to derive dimension free L_p entropy bounds for $1 \leq p < \infty$. The first such bound was proved by Dudley [**?**] and is based on a combination of a dimension reduction argument and the Sauer-Shelah Lemma. The dimension reduction part shows that if $K \subset F$ is "well separated" in L_1, in the sense that every two points are different on a number of coordinates that is proportional to n, one can find a much smaller set of coordinates (whose cardinality depends on the cardinality of K) on which every two points in K are different on at least one coordinate. We prove the claim for $p = 1$; the general case follows because the class contains $\{0,1\}$-valued functions, and for any $f, h \in F$ and any probability measure μ, $\|f - g\|^p_{L_p(\mu)} = \|f - g\|_{L_1(\mu)}$.

Theorem 13.5. *There exists an absolute constant c for which the following holds. Let F be a class of $\{0,1\}$-valued functions on a probability space (Ω, μ). If $\mathrm{VC}(F) \leq d$ then for any*

$0 < \varepsilon < 1/2$,

$$N(\varepsilon, F, L_1(\mu)) \leq \left(\frac{2}{\varepsilon}\right)^{cd}.$$

Proof. Consider first an arbitrary empirical measure $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, and let us prove the entropy estimate for such a measure. Set K_ε to be any ε-separated subset of F with respect to the $L_1(\mu_n)$ norm and denote its cardinality by D.

Let $V = \{f_i - f_j | f_i \neq f_j \in K_\varepsilon\}$; thus, $|V| \leq D^2$, and since K_ε is ε-separated and consist of $\{0,1\}$-valued functions, it follows that every $v \in V$ has at least $n\varepsilon$ coordinates which belong to $\{-1, 1\}$.

Set $(X_i)_{i=1}^t$ to be independent $\{x_1, ..., x_n\}$-valued random variables, where for every $1 \leq i \leq t$ and $1 \leq j \leq n$, $Pr(X_i = x_j) = 1/n$. For any $v \in V$,

$$Pr\left(\forall i,\ v(X_i) = 0\right) = \prod_{i=1}^t Pr\left(v(X_i) = 0\right) \leq (1-\varepsilon)^t,$$

implying that

$$Pr\left(\exists v \in V,\ \forall i,\ v(X_i) = 0\right) \leq |V|(1-\varepsilon)^t \leq D^2(1-\varepsilon)^t.$$

Therefore,

$$Pr\left(\forall v \in V,\ \exists i,\ 1 \leq i \leq t,\ |v(X_i)| = 1\right) \geq 1 - D^2(1-\varepsilon)^t,$$

and if the latter is greater than 0, there is a set $I \subset \{1, ..., n\}$ of cardinality $|I| \leq t$ and the mapping $(f(x_i))_{i=1}^n \to (f(x_i))_{i\in I}$ is one-to-one. In particular, setting

$$P_I K_\varepsilon = \left\{(f(x_i))_{i\in I} \,\middle|\, f \in K_\varepsilon\right\},$$

then $|P_I K_\varepsilon| = D$.

$$\left|\left\{(f(x_i))_{i\in I} \,\middle|\, f \in K_\varepsilon\right\}\right| = D.$$

It is straightforward to verify that the choice of $t = \frac{2\log D}{\varepsilon}$ suffices to ensure the existence of such a set I. Finally, by the Sauer-Shelah Lemma,

$$D = |P_I K_\varepsilon| \leq |P_I F| \leq \left(\frac{e|I|}{d}\right)^d \leq \left(\frac{2e\log D}{d\varepsilon}\right)^d. \tag{13.1}$$

To complete the proof, note that if $\alpha \geq 1$ and $\alpha \log^{-1}\alpha \leq \beta$ then $\alpha \lesssim \beta \log(e\beta)$. By (13.1),

$$\frac{c\log D}{\varepsilon d} \leq \frac{c}{\varepsilon}\log\left(\frac{c\log D}{\varepsilon d}\right),$$

and one may set $\alpha = (c\log D)/\varepsilon d$ and $\beta = 1/\varepsilon$. Thus, $|D| \leq (2/\varepsilon)^{c'd}$, as claimed. ∎

This result was strengthened by Haussler in [**?**]:

Theorem 13.6. *There is a constant C such that, for every class of binary-valued functions F with $\mathrm{VC}(F) = d$ and every $0 < \varepsilon < 1$, $N(\varepsilon, F) \leq Cd(4e)^d\varepsilon^{-d}$.*