REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
18-08-2014	Related Material	-

4. TITLE AND SUBTITLE
Probability in High Dimension

5a. CONTRACT NUMBER
W911NF-14-1-0094

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611103

6. AUTHORS
Ramon van Handel

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Princeton University
PO Box 36
87 Prospect Avenue
Princeton, NJ 08544 -2020

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
62276-MA-PCS.5

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
Lecture notes from course ORF 570 - Probability in High Dimension (educational material made freely available on my website)

15. SUBJECT TERMS
---

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Ramon van Handel
UU	UU	UU	UU		19b. TELEPHONE NUMBER 609-258-0973

Report Title

Probability in High Dimension

ABSTRACT

Lecture notes from course ORF 570 - Probability in High Dimension (educational material made freely available on my website)

Ramon van Handel

Probability in High Dimension

ORF 570 Lecture Notes
Princeton University

This version: June 30, 2014

Preface

These lecture notes were written for the course ORF 570: *Probability in High Dimension* that I taught at Princeton in the Spring 2014 semester. The aim was to introduce in as cohesive a manner as I could manage a set of methods, many of which have their origin in probability in Banach spaces, that arise across a broad range of contemporary problems in different areas.

The notes are necessarily incomplete. The ambitious syllabus for the course was laughably beyond the scope of Princeton's 12-week semester. As a result, there are regrettable omissions, as well as many fascinating topics that I would have liked to but could not cover in the context of this course. These include:

a. Bernstein's inequality does not appear anywhere in these notes (disgraceful!), nor do any Bernstein-type concentration inequalities (such as concentration of the exponential distribution and Talagrand's concentration inequality for empirical processes) and the notion of modified log-Sobolev inequalities. These should be included at the end of Part I.

b. Chaining with adaptive truncation and entropy with brackets. Beyond being a classical topic in empirical process theory, the power of the idea of adaptive truncation has again proven its value in the recent solution of the long-standing Bernoulli problem due to Bednorz and Latała.

c. Universality (prematurely included in Chapter 1 as a topic to be covered though I did not have time to do so) and an introduction to Stein's method.

d. Hypercontractivity and its applications, particularly to concentration inequalities and to sharp thresholds (the latter should be promoted to a fourth "general principle" in Chapter 1 in view of the ubiquity of phase transition phenomena in high-dimensional problems).

e. No doubt this list will grow even longer if I don't stop typing.

Hopefully the opportunity will arise in the future to fill in some of these gaps, in which case I will post an updated version of these notes on my website. For now, as always, these notes are made available as-is.

Please note that these are lecture notes, not a monograph. Many important ideas that I did not have the time to cover are included as problems at the end of each section. Doing the problems is the best way to learn the material. To avoid distraction I have on a few occasions ignored some minor technical issues (such as measurability issues of empirical processes or domain issues of Markov generators), but I have tried to give the reader a fair warning when this is the case. The notes at the end of each chapter do not claim to give a comprehensive historical account, but rather to indicate the immediate origin of the material that I used and to serve as a starting point for further reading.

Many thanks are due to the 30 or so regular participants of the course. These lecture notes are loosely based on notes scribed by the students during the lectures. While they have been almost entirely rewritten, the scribe notes served as a crucial motivation to keep writing. I am particularly grateful to Maria Avdeeva, Mark Cerenzia, Jacob Funk, Danny Gitelman, Max Goer, Jiequn Han, Daniel Jiang, Mitchell Johnston, Haruko Kato, George Kerchev, Dan Lacker, Che-Yu Liu, Yuan Liu, Huanran Lu, Junwei Lu, Tengyu Ma, Efe Onaran, Zhaonan Qu, Patrick Rebeschini, Max Simchowitz, Weichen Wang, Igor Zabukovec, Tianqi Zhao, and Ziwei Zhu for serving as scribes.

Princeton,
June 2014

Contents

1

Introduction

1.1 What is this course about?

What is probability in high dimension? There is no good answer to this question. High-dimensional probabilistic problems arise in numerous areas of science, engineering, and mathematics. A (very incomplete) list might include:

- Large random structures: random matrices, random graphs, ...
- Statistics and machine learning: estimation, prediction and model selection for high-dimensional data.
- Randomized algorithms in computer science.
- Random codes in information theory.
- Statistical physics: Gibbs measures, percolation, spin glasses, ...
- Random combinatorial structures: longest increasing subsequence, spanning trees, travelling salesman problem, ...
- Probability in Banach spaces: probabilistic limit theorems for Banach-valued random variables, empirical processes, local theory of Banach spaces, geometric functional analysis, convex geometry.
- Mixing times and other phenomena in high-dimensional Markov chains.

At first sight, these different topics appear to have limited relation to one another. Each of these areas is a field in its own right, with its own unique ideas, mathematical methods, etc. In fact, even the high-dimensional nature of the problems involved can be quite distinct: in some of these problems, "high dimension" refers to the presence of many distinct but interacting random variables; in others, the problems arise in high-dimensional spaces and probabilistic methods enter the picture indirectly. It would be out of the question to cover all of these topics in a single course.

Despite this wide array of quite distinct areas, there are the some basic probabilistic principles and techniques that arise repeatedly across a broad range of high-dimensional problems. These ideas, some of which will be described at a very informal level below, typically take the form of nonasymptotic probabilistic inequalities. Here *nonasymptotic* means that we are not

concerned with limit theorems (as in many classical probabilistic results), but rather with explicit estimates that are either dimension-free, or that capture precisely the dependence of the problem on the relevant dimensional parameters. There are at least two reasons for the importance of such methods. First, in many high-dimensional problems there may be several different parameters of interest; in asymptotic results one must take all these parameters to the limit in a fixed relation to one another, while the nonasymptotic viewpoint allows to express the interrelation between the different parameters in a much more precise way. More importantly, high-dimensional problems typically involve interactions between a large number of degrees of freedom whose aggregate contributions to the phenomenon of interest must be accounted for in the mathematical analysis; the explicit nature of nonasymptotic estimates makes them particularly well suited to be used as basic ingredients of the analysis, even if the ultimate result of interest is asymptotic in nature.

The goal of this course is to develop a set of tools that are used repeatedly in the investigation of high-dimensional random structures across different fields. Our aim will not only be to build up this common toolbox in a systematic way, but we will also attempt to show how these tools fit together to yield a surprisingly cohesive set of probabilistic ideas. Of course, one should not expect that any genuinely interesting problem that arises in one of the various fascinating areas listed above can be resolved by an immediate application of a tool in our toolbox; the solution of such problems typically requires insights that are specific to each area. However, the common set of ideas that we will develop provides key ingredients for the investigation of many high-dimensional problems, and forms an essential basis for work in this area.

1.2 Some general principles

The toolbox that we will develop is equipped to address a number of different phenomena that arise in high dimension. To give a broad overview of some of the ideas to be developed, and to set the stage for coming attractions, we will organize our theory around three *informal* "principles" to be described presently. None of these principles corresponds to one particular theorem or admits a precise mathematical description; rather, each principle encompasses a family of conceptually related results that appear in different guises in different settings. The bulk of this course is aimed at making these ideas precise.

1.2.1 Concentration

If $X_1, X_2, \ldots$ are i.i.d. random variables, then

$$\frac{1}{n}\sum_{k=1}^{n} X_k - \mathbf{E}\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right] \to 0 \quad \text{as} \quad n \to \infty$$

by the law of large numbers. Another way of stating this is as follows: if we define the function $f(x_1, \ldots, x_n) = \frac{1}{n} \sum_{k=1}^{n} x_k$, then for large n the random variable $f(X_1, \ldots, X_n)$ is close to its mean (that is, its fluctuations are small).

It turns out that this phenomenon is not restricted to linear functions f: it is a manifestation of a general principle, the *concentration* phenomenon, by virtue of which it is very common for functions of many independent variables to have small fluctuations. Let us informally state this principle as follows.

> *If $X_1, \ldots, X_n$ are independent (or weakly dependent) random variables, then the random variable $f(X_1, \ldots, X_n)$ is "close" to its mean $\mathbf{E}[f(X_1, \ldots, X_n)]$ provided that the function $f(x_1, \ldots, x_n)$ is not too "sensitive" to any of the coordinates x_i.*

Of course, to make such a statement precise, we have to specify:

- What do we mean by "sensitive"?
- What do we mean by "close"?

We will develop a collection of results, and some general methods to prove such results in different settings, in which these concepts are given a precise meaning. In each case, such a result takes the form of an explicit bound on a quantity that measures the size of the fluctuations $f(X_1, \ldots, X_n) - \mathbf{E}[f(X_1, \ldots, X_n)]$ (such as the variance or tail probabilities) in terms of "dimension" n and properties of the distribution of the random variables X_i.

At this point, it is perhaps far from clear why a principle of the above type might be expected to hold. We will develop a number of general tools to prove such results that provide insight into the nature of concentration, as well as its connection with other topics. One theme that will arise repeatedly in the sequel is the connection between concentration and the rate of convergence to equilibrium of Markov processes. At first sight, these appear to be entirely different questions: the concentration problem is concerned with the fluctuations of $f(X)$ for a given (vector-valued) random variable X and (possibly very nonlinear) function f, with no Markov process in sight. Nonetheless, it turns out that one can prove concentration properties by investigating Markov processes that have the distribution of X as their stationary distribution. Conversely, functional inequalities closely connected to concentration can be used to investigate the convergence of Markov processes to the stationary distribution (which is of interest in its own right in many areas, for example, in non-equilibrium statistical mechanics or Markov chain Monte Carlo algorithms). Once this connection has been understood, it will also become clear in what manner such results can be systematically improved. This will lead us to the notion of hypercontractivity of Markov semigroups, which is in turn of great interest in various other probabilistic problems. Several other connections that yield significant insight into the concentration phenomenon, including to isoperimetric problems and problems in optimal transportation and information theory, will be developed along the way.

1.2.2 Suprema

The concentration principle is concerned with the deviation of a random function $f(X_1, \ldots, X_n)$ from its mean $\mathbf{E}[f(X_1, \ldots, X_n)]$. However, it does not provide any information on the value of $\mathbf{E}[f(X_1, \ldots, X_n)]$ itself. In fact, the two problems of estimating the magnitude and the fluctuations of $f(X_1, \ldots, X_n)$ prove to be quite distinct, and must be treated by different methods.

A remarkable feature of the concentration principle is that it provides information on the fluctuations for very general functions f: even in cases where the function f is very complicated to compute (for example, when it is defined in terms of a combinatorial optimization problem), it is often possible to estimate its sensitivity to the coordinates by elementary methods. When it comes to estimating the magnitude of the corresponding random variable, there is no hope to develop a principle that holds at this level of generality: the functions f that arise in the different areas described in the previous section are very different in nature, and we cannot hope to develop general tools to address such problems without assuming some additional structure.

A structure that proves to be of central importance in many high-dimensional problems is that of random variables F defined as the supremum

$$F = \sup_{t \in T} X_t$$

of a random process $\{X_t\}_{t \in T}$ (that is, a family of random variables indexed by a set T that is frequently high- or infinite-dimensional). The reason that such quantities play an important role in high-dimensional problems is twofold. On the one hand, problems in high dimension typically involve a large number of interdependent degrees of freedom; the need to obtain simultaneous control over many random variables thus arises frequently as an ingredient of the mathematical analysis. On the other hand, there are many problems in which various quantities of interest can be naturally expressed in terms of suprema. Let us consider a few simple examples for sake of illustration.

Example 1.1 (Random matrices). Let $M = (M_{ij})_{1 \le i,j \le n}$ be a random matrix whose entries M_{ij} are independent (let us assume they are Gaussian for sake of illustration). One question of interest in this setting is to estimate the magnitude of the matrix norm $\|M\|$ (the largest singular value of M), which is a nontrivial function of matrix entries. But recall from linear algebra that

$$\|M\| = \sup_{v,w \in B_2} \langle v, Mw \rangle,$$

where B_2 is the (Euclidean) unit ball and $\langle \cdot, \cdot \rangle$ denotes the usual inner product in $\mathbb{R}^n$. We can therefore treat the matrix norm $\|M\|$ as the supremum of the Gaussian process $\{X_{v,w} = \langle v, Mw \rangle\}_{v,w \in B_2}$ indexed by $B_2 \times B_2$.

Example 1.2 (Norms of random vectors). Let X be a random vector in $\mathbb{R}^n$, and let $\|\cdot\|_B$ be any norm on $\mathbb{R}^n$ (where B denotes the unit ball of $\|\cdot\|_B$). The duality theory of Banach spaces implies that we can write

$$\|X\|_B = \sup_{t \in B^\circ} \langle t, X \rangle,$$

where B° denotes the dual ball. In this manner, the supremum of the random process $\{X_t = \langle t, X \rangle\}_{t \in B^\circ}$ arises naturally in probability in Banach spaces.

Example 1.3 (Empirical risk minimization). Many problems in statistics and machine learning may be formulated as the problem of computing

$$\operatorname*{argmin}_{\theta \in \Theta} \mathbf{E}[l(\theta, X)]$$

given only observed "data" consisting of i.i.d. samples $X_1, \ldots, X_n \sim X$ (that is, without knowledge of the law of X). Here l is a given loss function and Θ is a given parameter space, which depend on the problem at hand.

Perhaps the simplest general way to address this problem is to reason as follows. By the law of large numbers, we can approximate the risk for a fixed parameter θ by the empirical risk which depends only on the data:

$$\mathbf{E}[l(\theta, X)] \approx \frac{1}{n} \sum_{k=1}^{n} l(\theta, X_k).$$

On might therefore naturally expect that

$$\operatorname*{argmin}_{\theta \in \Theta} \mathbf{E}[l(\theta, X)] \approx \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{n} \sum_{k=1}^{n} l(\theta, X_k).$$

This approach to estimating the optimal parameter θ from data is called empirical risk minimization. The problem is now to estimate how close the empirical risk minimizer is to the optimal parameter as a function of the number of samples n, the dimension of the parameter space Θ, the dimension of the state space of X, etcetera. The resolution of this question leads naturally to the investigation of quantities such as the uniform deviation

$$\sup_{\theta \in \Theta} \frac{1}{n} \sum_{k=1}^{n} \{l(\theta, X_k) - \mathbf{E}[l(\theta, X)]\},$$

which is the supremum of a random process. Estimating the magnitude of suprema arises in a similar manner in a wide array of statistical problems.

Example 1.4 (Convex functions). In principle, we can formulate the problem of estimating $\mathbf{E}[f(X_1, \ldots, X_n)]$ as a supremum problem whenever f is *convex*. Indeed, by convex duality, we can express any convex function $f : \mathbb{R}^n \to \mathbb{R}$ as

$$f(x) = \sup_{y \in \mathbb{R}^n} \{\langle y, x \rangle - f^*(y)\},$$

where f^* denotes the convex conjugate of f. The function $f(X_1, \ldots, X_n)$ can therefore be expressed as the supremum of the random process $\{X_y =$

$\langle y, X\rangle\}_{y\in\mathbb{R}^n}$ after subtracting the "penalty" $f^*(y)$ (alternatively, f^* can be absorbed in the definition of X_y). This shows that the investigation of suprema is in fact surprisingly general; this general point of view is very useful in some applications, while more direct methods might be more suitable in other cases.

In all these cases, the process X_t itself admits a simple description, and the difficulty lies in obtaining good estimates on the magnitude of the supremum (for example, to estimate the mean or the tail probabilities). In this setting, a second general principle appears that provides a key tool in many high-dimensional problems. We informally state this principle as follows.

> *If the random process $\{X_t\}_{t\in T}$ is "sufficiently continuous," then the magnitude of the supremum $\sup_{t\in T} X_t$ is controlled (in the sense that we have estimates from above, and in some cases also from below) by the "complexity" of the index set T.*

Of course, to make this precise, we have to specify:

- What do we mean by "sufficiently continuous"?
- What do we mean by "complexity"?

These concepts will be given a precise meaning in the sequel. In particular, let us note that while the supremum of a random process is a probabilistic object, complexity is not: we will in fact consider different geometric (packing and covering numbers and trees) and combinatorial (shattering and combinatorial dimension) notions of complexity. We will develop a collection of powerful tools, such as chaining and slicing methods, that make the connection between these probabilistic, geometric, and combinatorial notions in a general setting. A number of other useful tools will be developed along the way, such as basic methods for bounding Gaussian and Rademacher processes.

1.2.3 Universality

Let $X_1, X_2, \ldots$ be i.i.d. random variables with finite variance. As in our discussion of concentration, let us recall once more the law of large numbers

$$\frac{1}{n}\sum_{k=1}^{n}\{X_k - \mathbf{E}X_k\} \to 0 \quad \text{as} \quad n\to\infty.$$

In this setting, however, we do not only know that the fluctuations are of order $n^{-1/2}$ (as is captured by the concentration phenomenon), but we have much more precise information as well: by the central limit theorem, we have a precise description of the distribution of the fluctuations, as

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\{X_k - \mathbf{E}X_k\} \approx \text{Gaussian}$$

when n is large. A different way of phrasing this property is that

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\{X_k - \mathbf{E}X_k\} \approx \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\{G_k - \mathbf{E}G_k\},$$

where G_k are independent Gaussian random variables with the same mean and variance of X_k (here $\approx$ denotes closeness of the distributions). Beside the fact that this gives precise distributional information, what is remarkable about such results is that they become insensitive to the distribution of the original random variables X_k as $n \to \infty$. The phenomenon that the detailed features of the distribution of the individual components of a problem become irrelevant in high dimension is often referred to a *universality*.

As in the case of concentration, it turns out that this phenomenon is not restricted to linear functions of independent random variables, but is in fact a manifestation of a more general principle. We state it informally as follows.

> *If $X_1, \ldots, X_n$ are independent (or weakly dependent) random variables, then the expectation $\mathbf{E}[f(X_1, \ldots, X_n)]$ is "insensitive" to the distribution of $X_1, \ldots, X_n$ when the function f is "sufficiently smooth."*

Of course, to make this precise, we have to specify:

- What do we mean by "insensitive"?
- What do we mean by "sufficiently smooth"?

We will develop some basic quantitative methods to prove universality in which these concepts are given a precise meaning.

The interest of the universality phenomenon is twofold. First, the presence of the universality property suggests that the high-dimensional phenomenon under investigation is in a sense robust to the precise details of the model ingredients, a conclusion of significant interest in its own right (of course, there are also many high-dimensional phenomena that are *not* universal!) Second, there are often situations in which the quantities of interest can be evaluated by explicit computation when the underlying random variables have a special distribution, but where such explicit analysis would be impossible in a general setting. For example, in random matrix theory, many explicit computations are possible for appropriately defined Gaussian random matrices due to the invariance of the distribution under orthogonal transformations, while such computations would be completely intractable for other distributions of the entries. In such cases, universality properties provide a crucial tool to reduce the proofs of general results to those in a tractable special case.

Let us note that the universality phenomenon is not necessarily related to the Gaussian distribution: universality simply states that certain probabilistic quantities do not depend strongly on the distribution of the individual components. However, Gaussian distributions do appear frequently in many high-dimensional problems that involve the aggregate effect of many independent degrees of freedom, as do several other distributions (such as Poisson

distributions in discrete problems and extreme value distributions for maxima of independent random variables; a much less well understood phenomenon is the appearance of the Tracy-Widom distribution in many complex systems that are said to belong to the "KPZ universality class," a topic of intense recent activity in probability theory.) Thus the related but more precise question of when the distribution a random variable F is close to Gaussian or to some other distribution also arises naturally in this setting. Explicit nonasymptotic estimates in terms of dimensional parameters of the problem can be obtained using a set of tools (collectively known as Stein's method) that have proved to be very useful in a number of high-dimensional problems.

1.3 Organization of this course

We have introduced above three "principles" to motivate some of the general probabilistic ideas that arise in high-dimensional problems. These principles should not be taken too seriously, but rather an informal guide to place into perspective the topics that we will cover in the sequel. In the following lectures, we will proceed to develop these ideas in a precise manner, and to exhibit the many interconnections between these topics.

Unfortunately, there is a lot of ground to cover, probably way too much for a single semester. Thus some hard choices will likely have to be made, depending on the interests of the audience. Let us start out ambitiously, and see how things develop as the course progresses.

Part I

Concentration

2

Variance bounds and Poincaré inequalities

Recall the informal statement of the concentration phenomenon from Ch. 1:

> *If $X_1, \ldots, X_n$ are independent (or weakly dependent) random variables, then the random variable $f(X_1, \ldots, X_n)$ is "close" to its mean $\mathbf{E}f(X_1, \ldots, X_n)$ provided that the function $f(x_1, \ldots, x_n)$ is not too "sensitive" to any of the coordinates x_i.*

In this chapter, we will make a modest start towards making this principle precise by investigating bounds on the variance

$$\mathrm{Var}[f(X_1, \ldots, X_n)] := \mathbf{E}[(f(X_1, \ldots, X_n) - \mathbf{E}f(X_1, \ldots, X_n))^2]$$

in terms of the "sensitivity" of the function f to its coordinates. Various fundamental ideas and a rich theory already arise in this setting, and this is therefore our natural starting point. In the following chapters we will show how to go beyond the variance to obtain bounds on the distribution of the fluctuations of $f(X_1, \ldots, X_n)$ that are useful in many settings.

2.1 Tensorization and bounded differences

At first sight, it might seem that the concentration principle is rather trivial when stated in terms of variance. Indeed, the variance of a constant function is zero, and it is easy to show that the variance of a function that is almost constant is almost zero. For example, we have the following simple lemma:

Lemma 2.1. *Let X be any (possibly vector-valued) random variable. Then*

$$\mathrm{Var}[f(X)] \le \tfrac{1}{4}(\sup f - \inf f)^2 \quad \textit{and} \quad \mathrm{Var}[f(X)] \le \mathbf{E}[(f(X) - \inf f)^2].$$

Proof. Note that

$$\mathrm{Var}[f(X)] = \mathrm{Var}[f(X) - a] \le \mathbf{E}[(f(X) - a)^2] \quad \text{for any } a \in \mathbb{R}.$$

For the first inequality, let $a = (\sup f + \inf f)/2$ and note that $|f(X) - a| \le (\sup f - \inf f)/2$. For the second inequality, let $a = \inf f$. □

The problem with this trivial result is that it does not capture at all the *high-dimensional* phenomenon that we set out to investigate. For example, it gives a terrible bound for the law of large numbers.

Example 2.2. Let $X_1, \ldots, X_n$ be independent random variables with values in $[-1, 1]$, and let $f(x_1, \ldots, x_n) = \frac{1}{n} \sum_{k=1}^n x_k$. Then a direct computation gives

$$\mathrm{Var}[f(X_1, \ldots, X_n)] = \frac{1}{n} \sum_{k=1}^n \mathrm{Var}[X_k] \le \frac{1}{n}.$$

That is, the average of i.i.d. random variables concentrates increasingly well around its mean as the dimension is increased. On the other hand, both bounds of Lemma 2.1 give $\mathrm{Var}[f(X_1, \ldots, X_n)] \lesssim 1$: for example,

$$\mathrm{Var}[f(X_1, \ldots, X_n)] \le \frac{1}{4} (\sup f - \inf f)^2 = 1.$$

Thus Lemma 2.1 provides a reasonable bound on the variance in one dimension, but is grossly inadequate in high dimension.

Of course, this should not be surprising: no independence was assumed in Lemma 2.1, and so there is no reason which we should obtain a sharper concentration phenomenon at this level of generality. For example, if $X_1, \ldots, X_n$ are random variables that are totally dependent $X_1 = X_2 = \ldots = X_n$, then the variance of $\frac{1}{n} \sum_{k=1}^n X_k$ is indeed of order one regardless of the "dimension" n, and Lemma 2.1 captures this situation accurately. The idea that concentration should improve in high dimension arises when there are many *independent* degrees of freedom. To capture this high-dimensional phenomenon, we must develop a method to exploit independence in our inequalities.

To this end, we presently introduce an idea that appears frequently in high-dimensional problems: we will deduce a bound for functions of independent random variables $X_1, \ldots, X_n$ (i.e., in high dimension) from bounds for functions of each individual random variable X_i (i.e., in a single dimension). It is not at all obvious that this is possible: in general, one cannot expect to deduce high-dimensional inequalities from low-dimensional ones without introducing additional dimension-dependent factors. Those quantities for which this is in fact possible are said to *tensorize*.[1] Quantities that tensorize behave well in high dimension, and are therefore particularly important in high-dimensional problems. We will presently prove that the variance is such a quantity. With the tensorization inequality for the variance in hand, we will have reduced the proof of concentration inequalities for functions of many independent random variables to obtaining such bounds for a single random variable.

[1] The joint law $\mu_1 \otimes \cdots \otimes \mu_n$ of independent random variables $X_1, \ldots, X_n$ is the tensor product of the marginal laws $X_i \sim \mu_i$: the terminology "tensorization" indicates that a quantity is well behaved under the formation of tensor products.

To formulate the tensorization inequality, let $X_1, \dots, X_n$ be independent random variables. For each function $f(x_1, \dots, x_n)$, we define the function

$$\mathrm{Var}_i f(x_1, \dots, x_n) := \mathrm{Var}[f(x_1, \dots, x_{i-1}, X_i, x_{i+1}, \dots, x_n)].$$

That is, $\mathrm{Var}_i f(x)$ is the variance of $f(X_1, \dots, X_n)$ with respect to the variable X_i only, the remaining variables being kept fixed.

Theorem 2.3 (Tensorization of variance). *We have*

$$\mathrm{Var}[f(X_1, \dots, X_n)] \le \mathbf{E}\left[\sum_{i=1}^n \mathrm{Var}_i f(X_1, \dots, X_n)\right]$$

whenever $X_1, \dots, X_n$ are independent.

Note that when f is a linear function, it is readily checked that the inequality of Theorem 2.3 holds with equality: in this sense, the result is sharp.

The proof of Theorem 2.3 is a first example of the *martingale method*, which will prove useful for obtaining more general inequalities later on.

Proof. Define

$$\Delta_k = \mathbf{E}[f(X_1, \dots, X_n)|X_1, \dots, X_k] - \mathbf{E}[f(X_1, \dots, X_n)|X_1, \dots, X_{k-1}].$$

Then

$$f(X_1, \dots, X_n) - \mathbf{E}f(X_1, \dots, X_n) = \sum_{k=1}^n \Delta_k,$$

and $\mathbf{E}[\Delta_k|X_1, \dots, X_{k-1}] = 0$, that is, $\Delta_1, \dots, \Delta_k$ are martingale increments. In particular, as $\mathbf{E}[\Delta_k \Delta_l] = \mathbf{E}[\mathbf{E}[\Delta_k|X_1, \dots, X_{k-1}]\Delta_l] = 0$ for $l < k$, we have

$$\mathrm{Var}[f(X_1, \dots, X_n)] = \mathbf{E}\left[\left(\sum_{k=1}^n \Delta_k\right)^2\right] = \sum_{k=1}^n \mathbf{E}[\Delta_k^2].$$

It remains to show that $\mathbf{E}[\Delta_k^2] \le \mathbf{E}[\mathrm{Var}_k f(X_1, \dots, X_n)]$ for every k.

To this end, note that

$$\begin{aligned}
&\mathbf{E}[f(X_1, \dots, X_n)|X_1, \dots, X_{k-1}] \\
&= \mathbf{E}[\mathbf{E}[f(X_1, \dots, X_n)|X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n]|X_1, \dots, X_{k-1}] \\
&= \mathbf{E}[\mathbf{E}[f(X_1, \dots, X_n)|X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n]|X_1, \dots, X_k],
\end{aligned}$$

where we have used the tower property of the conditional expectation in the first equality, and that X_k is independent of $X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n$ in the second equality. In particular, we can write $\Delta_k = \mathbf{E}[\tilde{\Delta}_k|X_1, \dots, X_k]$ with

$$\tilde{\Delta}_k = f(X_1, \dots, X_n) - \mathbf{E}[f(X_1, \dots, X_n)|X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n].$$

But as X_k and $X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n$ are independent, we have

$$\mathrm{Var}_k f(X_1, \ldots, X_n) = \mathbf{E}[\tilde{\Delta}_k^2 | X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n].$$

We can therefore estimate using Jensen's inequality

$$\mathbf{E}[\Delta_k^2] = \mathbf{E}[\mathbf{E}[\tilde{\Delta}_k | X_1, \ldots, X_k]^2] \le \mathbf{E}[\tilde{\Delta}_k^2] = \mathbf{E}[\mathrm{Var}_k f(X_1, \ldots, X_n)],$$

which completes the proof. □

One can view tensorization of the variance in itself as an expression of the concentration phenomenon: $\mathrm{Var}_i f(x)$ quantifies the sensitivity if the function $f(x)$ to the coordinate x_i in a distribution-dependent manner. Thus Theorem 2.3 already the expresses the idea that if the sensitivity of f to each coordinate is small, then $f(X_1, \ldots, X_n)$ is close to its mean. Unlike Lemma 2.1, however, Theorem 2.3 holds with equality for linear functions and thus captures precisely the behavior of the variance in the law of large numbers. The tensorization inequality generalizes this idea to arbitrary nonlinear functions, and constitutes our first nontrivial concentration result.

However, it may not be straightforward to compute $\mathrm{Var}_i f$: this quantity depends not only on the function f, but also on the distribution of X_i. In many cases, Theorem 2.3 is the most useful in combination with a suitable bound on the variances $\mathrm{Var}_i f$ in each dimension. Even the trivial bounds of Lemma 2.1 already suffice to obtain a variance bound that is extremely useful in many cases. To this end, let us define the quanities

$$D_i f(x) := \sup_z f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n) - \inf_z f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n)$$

and

$$D_i^- f(x) := f(x_1, \ldots, x_n) - \inf_z f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n).$$

Then $D_i f(x)$ and $D_i^- f(x)$ quantify the sensitivity of the function $f(x)$ to the coordinate x_i in a distribution-independent manner. The following bounds now follow immediately from Theorem 2.3 and Lemma 2.1.

Corollary 2.4 (Bounded difference inequalities). *We have*

$$\mathrm{Var}[f(X_1, \ldots, X_n)] \le \frac{1}{4} \mathbf{E}\left[\sum_{i=1}^n (D_i f(X_1, \ldots, X_n))^2\right]$$

and

$$\mathrm{Var}[f(X_1, \ldots, X_n)] \le \mathbf{E}\left[\sum_{i=1}^n (D_i^- f(X_1, \ldots, X_n))^2\right]$$

whenever $X_1, \ldots, X_n$ are independent.

Let us illustrate the utility of these inequalities in a nontrivial example.

Example 2.5 (Random matrices). Let M be an $n \times n$ symmetric matrix where $\{M_{ij} : i \geq j\}$ are i.i.d. symmetric Bernoulli random variables $\mathbf{P}[M_{ij} = \pm 1] = \frac{1}{2}$. We are interested in $\lambda_{\max}(M)$, the largest eigenvalue of M. This is a highly nonlinear function of the entries: it is not immediately obvious what is the order of magnitude of either the mean or the variance of $\lambda_{\max}(M)$.

Recall from linear algebra that

$$\lambda_{\max}(M) = \sup_{v \in B_2} \langle v, Mv \rangle = \langle v_{\max}(M), M v_{\max}(M) \rangle,$$

where $B_2 = \{v \in \mathbb{R}^n : \|v\|_2 \leq 1\}$ is the Euclidean unit ball in $\mathbb{R}^n$ and $v_{\max}(M)$ is any eigenvector of M with eigenvalue $\lambda_{\max}(M)$. Since $\lambda_{\max}(M)$ is the supremum of a random process, we will be able to use tools from the second part of this course to estimate its mean: it will turn out that $\mathbf{E}[\lambda_{\max}(M)] \sim \sqrt{n}$. Let us now use Corollary 2.4 to estimate the variance.

Let us consider for the time being a fixed matrix M and indices $i \geq j$. Choose a symmetric matrix M^- such that

$$\lambda_{\max}(M^-) = \inf_{M_{ij}} \lambda_{\max}(M),$$

that is, $M^-_{ij} = M^-_{ji}$ is chosen to minimize $\lambda_{\max}(M^-)$ while the remaining entries $M^-_{kl} = M_{kl}$ with $\{k, l\} \neq \{i, j\}$ are kept fixed. Then we can estimate

$$\begin{aligned}
D^-_{ij}\lambda_{\max}(M) &= \lambda_{\max}(M) - \lambda_{\max}(M^-) \\
&= \langle v_{\max}(M), M v_{\max}(M) \rangle - \sup_{v \in B_2} \langle v, M^- v \rangle \\
&\leq \langle v_{\max}(M), (M - M^-) v_{\max}(M) \rangle \\
&= 2 v_{\max}(M)_i v_{\max}(M)_j (M_{ij} - M^-_{ij}) \\
&\leq 4 |v_{\max}(M)_i| |v_{\max}(M)_j|,
\end{aligned}$$

where the penultimate line holds as $M_{kl} = M^-_{kl}$ unless $k = i, l = j$ or $k = j, l = i$, and the last line holds as M_{ij}, M^-_{ij} only take the values ± 1. As this inequality holds for every matrix M and indices i, j, Corollary 2.4 yields

$$\mathrm{Var}[\lambda_{\max}(M)] \leq \mathbf{E}\left[\sum_{i \geq j} 16 |v_{\max}(M)_i|^2 |v_{\max}(M)_j|^2 \right] \leq 16,$$

where we have used that $\sum_{i=1}^n v_{\max}(M)_i^2 = 1$. Thus the variance of the maximal eigenvalue of an $n \times n$ symmetric random matrix with Bernoulli entries is bounded uniformly in the dimension n (in contrast to the mean $\sim \sqrt{n}$).

Remark 2.6. It is natural to ask whether the result of Example 2.5 is sharp: is $\mathrm{Var}[\lambda_{\max}(M)]$ in fact of constant order as $n \to \infty$? It turns out that this is not

the case: using highly nontrivial specialized tools from random matrix theory, it can be shown that in fact $\mathrm{Var}[\lambda_{\max}(M)] \sim n^{-1/3}$, that is, the fluctuations of the maximal eigenvalue in high dimension are *even smaller* than is predicted by Corollary 2.4. In this example, the suboptimal bound already arises at the level of the tensorization inequality: none of the methods developed here can beat dimension-free rate obtained in Example 2.5.

Thus this example highlights the fact that one cannot always expect to obtain an *optimal* bound by the application of a general theorem. However, this in no way diminishes the utility of these inequalities, whose aim is to provide *general principles* for obtaining concentration properties in high dimension. Indeed, even in the present example, we already obtained a genuinely nontrivial result—a dimension-free bound on the variance—using a remarkably simple analysis that did not use any special structure of random matrix problems. In many applications such dimension-free bounds suffice, or provide essential ingredients for a more delicate problem-specific analysis. It should also be noted that there are many problems in which results such as Corollary 2.4 *do* give bounds of the optimal order (for example, for linear functions f). Whether there exist general principles that can capture the improved order of the fluctuations in settings such as Example 2.5—the *superconcentration* problem—remains a largely open question. This is an active research area.

The bounded difference inequalities of Corollary 2.4, and the tensorization inequality 2.3, are very useful in many settings. On the other hand, these inequalities can often be restrictive due to various drawbacks:

- Due to the supremum and infimum in the definition of $D_i f$ or $D_i^- f$, bounds using bounded difference inequalities are typically restricted to situations where the random variables X_i and/or the function f are bounded. For example, the computation in Example 2.5 is useless for random matrices with Gaussian entries. On the other hand, the tensorization inequality itself does not require boundedness, but in nontrivial problems such as Example 2.5 it is typically far from clear how to bound $\mathrm{Var}_i f$.

- Bounded difference inequalities do not capture any information on the distribution of X_i. For example, suppose $X_1, \ldots, X_n$ are i.i.d., and consider $f(x) = \frac{1}{\sqrt{n}} \sum_{k=1}^n x_k$. Then $\mathrm{Var}[f(X_1, \ldots, X_n)] = \mathrm{Var}[X_1]$, but the bounded difference inequality only gives $\mathrm{Var}[f(X_1, \ldots, X_n)] \le \|X_1\|_\infty^2$. The latter will be very pessimistic when $\mathrm{Var}[X_1] \ll \|X_1\|_\infty^2$. On the other hand, the tensorization inequality is *too* distribution-dependent in that it is often unclear how to bound $\mathrm{Var}_i f$ directly for a given distribution.

- The tensorization method depends fundamentally on the independence of $X_1, \ldots, X_n$: it is not clear how this method can be extended beyond independence to treat more general classes of high-dimensional distributions.

To address these issues, we must develop a more general framework for understanding and proving variance inequalities.

Let us note that the inequalities obtained in this section can be viewed as special cases of a general family of inequalities that are informally described as follows. We can interpret $D_i f$ as a type of "discrete derivative of the function $f(x)$ with respect to the variable x_i." Similarly, $D_i^- f$ can be viewed as a one-sided version of the discrete derivative. More vaguely, one could also view $\mathrm{Var}_i f$ as a type of squared discrete derivative. Thus the inequalities of this section are, rougly speaking, of the following form:

$$\text{“ variance}(f) \lesssim \mathbf{E}[\,\|\text{gradient}(f)\|^2\,]. \text{ ”}$$

Inequalities of this type are called *Poincaré inequalities* (after H. Poincaré who first published such an inequality for the uniform distribution on a bounded domain in $\mathbb{R}^n$ and for the classical notion of gradient, ca. 1890). It turns out that the validity of a Poincaré inequality for a given distribution is intimately connected the convergence rate of a Markov process that admits that distribution as a stationary measure. This fundamental connection between two probabilistic problems provides a powerful framework to understand and prove a broad range of Poincaré inequalities for different distributions and with various different notions of "gradient" (and, conversely, a powerful method to bound the convergence rate of Markov processes in high dimension—an important problem in its own right with applications in areas ranging from statistical mechanics to Markov Chain Monte Carlo algorithms in computer science and in computational statistics). We therefore set out in the sequel to develop this connection in some detail. Before we can do that, however, we must first recall some basic elements of the theory of Markov processes.

Problems

2.1 (Banach-valued sums). Let $X_1, \ldots, X_n$ be random variables with values in a Banach space $(B, \|\cdot\|_B)$. Suppose that that these random variables are bounded in the sense that $\|X_i\|_B \le C$ a.s. for every i. Show that

$$\mathrm{Var}\left(\left\|\frac{1}{n}\sum_{k=1}^n X_k\right\|_B\right) \le \frac{4C^2}{n}.$$

This is a simple vector-valued variant of the elementary fact that the variance of $\frac{1}{n}\sum_{k=1}^n X_k$ for real-valued random variables X_k is of order $\frac{1}{n}$.

2.2 (Rademacher processes). Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent symmetric Bernoulli random variables $\mathbf{P}[\varepsilon_i = \pm 1] = \frac{1}{2}$ (also called Rademacher variables), let $T \subseteq \mathbb{R}^n$. The following identity is completely trivial:

$$\sup_{t\in T} \mathrm{Var}\left[\sum_{k=1}^n \varepsilon_k t_k\right] = \sup_{t\in T} \sum_{k=1}^n t_k^2.$$

Prove the following nontrivial fact:

$$\mathrm{Var}\left[\sup_{t\in T}\sum_{k=1}^{n}\varepsilon_k t_k\right] \le 4\sup_{t\in T}\sum_{k=1}^{n} t_k^2.$$

Thus taking the supremum inside the variance costs at most a constant factor.

2.3 (Bin packing). This is a classical application of bounded difference inequalities. Let $X_1,\dots,X_n$ be i.i.d. random variables with values in $[0,1]$. Each X_i represents the size of a package to be shipped. The shipping containers are bins of size 1 (so each bin can hold a set packages whose sizes sum to at most 1). Let $B_n = f(X_1,\dots,X_n)$ be the minimal number of bins needed to store the packages. Note that computing B_n is a hard combinatorial optimization problem, but we can bound its mean and variance by easy arguments.

a. Show that $\mathrm{Var}[B_n] \le n/4$.
b. Show that $\mathbf{E}[B_n] \ge n\mathbf{E}[X_1]$.

Thus the fluctuations $\sim\sqrt{n}$ of B_n are much smaller than its magnitude $\sim n$.

2.4 (Order statistics and spacings). Let $X_1,\dots,X_n$ be independent random variables, and denote by $X_{(1)} \ge \dots \ge X_{(n)}$ their decreasing rearrangement (so $X_{(1)} = \max_i X_i$, $X_{(n)} = \min_i X_i$, etc.) Show that

$$\mathrm{Var}[X_{(k)}] \le k\mathbf{E}[(X_{(k)} - X_{(k+1)})^2] \quad \text{for } 1 \le k \le n/2,$$

and that

$$\mathrm{Var}[X_{(k)}] \le (n-k+1)\mathbf{E}[(X_{(k-1)} - X_{(k)})^2] \quad \text{for } n/2 < k \le n.$$

Hint: use Corollary 2.4 creatively.

2.5 (Convex Poincaré inequality). Let $X_1,\dots,X_n$ be independent random variables taking values in $[a,b]$. The bounded difference inequalities of Corollary 2.4 estimate the variance $\mathrm{Var}[f(X_1,\dots,X_n)]$ in terms of *discrete* derivatives $D_i f$ or $D_i^- f$ of the function f. The goal of this problem is to show that if the function f is *convex*, then one can obtain a similar bound in terms of the ordinary notion of derivative $\nabla_i f(x) = \partial f(x)/\partial x_i$ in $\mathbb{R}^n$.

a. Show that if $g:\mathbb{R}\to\mathbb{R}$ is convex, then

$$g(y) - g(x) \ge g'(x)(y-x) \quad \text{for all } x,y\in\mathbb{R}.$$

b. Show using part a. and Corollary 2.4 that if $f:\mathbb{R}^n\to\mathbb{R}$ is convex, then

$$\mathrm{Var}[f(X_1,\dots,X_n)] \le (b-a)^2\,\mathbf{E}[\|\nabla f(X_1,\dots,X_n)\|^2].$$

c. Conclude that if f is convex and L-Lipschitz, i.e., $|f(x)-f(y)| \le L\|x-y\|$ for all $x,y\in[a,b]^n$, then $\mathrm{Var}[f(X_1,\dots,X_n)] \le L^2(b-a)^2$.

2.2 Markov semigroups

A (homogeneous) Markov process $(X_t)_{t\in\mathbb{R}_+}$ is a random process that satisfies the *Markov property*: for every bounded measurable function f and $s,t \in \mathbb{R}_+$, there is a bounded measurable function $P_s f$ such that

$$\mathbf{E}[f(X_{t+s})|\{X_r\}_{r\le t}] = (P_s f)(X_t).$$

[We do not put any restrictions on the state space: X_t can take values in any measurable space E, and the functions above are of the form $f : E \to \mathbb{R}$.] The interpretation, of course, is classical: the behavior of the process in the future X_{t+s} depends only on the history to date $\{X_r\}_{r\le t}$ through the current state X_t, and is independent of the prior history; that is, the dynamics of the Markov processes are memoryless. The assumption that $P_s f$ does not also depend on t in the above expression (the homogeneity property) indicates that the same dynamical mechanism is used at each time.

A probability measure μ is called *stationary* or *invariant* if

$$\mu(P_t f) = \mu(f) \quad \text{for all } t \in \mathbb{R}_+, \text{ bounded measurable } f.$$

To interpret this notion, suppose that $X_0 \sim \mu$. Then

$$\mathbf{E}[f(X_t)] = \mathbf{E}[\mathbf{E}[f(X_t)|X_0]] = \mathbf{E}[P_t f(X_0)] = \mu(P_t f).$$

Thus if μ is stationary, then $\mathbf{E}[f(X_t)] = \mu(f)$ for every $t \in \mathbb{R}_+$ and f: in particular, if the process is initially distributed according to the stationary measure $X_0 \sim \mu$, then the process remains distributed according to the stationary measure $X_t \sim \mu$ for every time t. In other words, stationary measures describe the "steady-state" or "equilibrium" behavior of a Markov process.

Let us describe a few basic facts about the functions $P_t f$.

Lemma 2.7. *Let μ be a stationary measure. Then the following hold for all $p \ge 1$, $t,s \in \mathbb{R}_+$, $\alpha,\beta \in \mathbb{R}$, bounded measurable functions f,g:*

1. $\|P_t f\|_{L^p(\mu)} \le \|f\|_{L^p(\mu)} := \mu(f^p)^{1/p}$ *(contraction).*
2. $P_t(\alpha f + \beta g) = \alpha P_t f + \beta P_t g$ *μ-a.s. (linearity).*
3. $P_{t+s} f = P_t P_s f$ *μ-a.s. (semigroup property).*
4. $P_t 1 = 1$ *μ-a.s. (conservativeness).*

In particular, $\{P_t\}_{t\in\mathbb{R}_+}$ defines a semigroup of linear operators on $L^p(\mu)$.

Proof. Assume that $X_0 \sim \mu$. To prove contraction, note that

$$\|P_t f\|^p_{L^p(\mu)} = \mathbf{E}[\mathbf{E}[f(X_t)|X_0]^p] \le \mathbf{E}[\mathbf{E}[f(X_t)^p|X_0]] = \|f\|^p_{L^p(\mu)},$$

where we have used Jensen's inequality. Linearity follows similarly as

$$\mathbf{E}[\alpha f(X_t) + \beta g(X_t)|X_0] = \alpha\mathbf{E}[f(X_t)|X_0] + \beta\mathbf{E}[g(X_t)|X_0].$$

To prove the semigroup property, note that

$$\mathbf{E}[f(X_{t+s})|X_0] = \mathbf{E}[\mathbf{E}[f(X_{t+s})|\{X_r\}_{r\le t}]|X_0] = \mathbf{E}[P_s f(X_t)|X_0].$$

The last property is trivial. □

Remark 2.8. Let μ be a stationary measure. In view of Lemma 2.7, it is easily seen that the definition and basic properties of $P_t f$ make sense not only for bounded measurable functions f, but also for every $f \in L^1(\mu)$. From now on, we will assume the $P_t f$ is defined in this manner for every $f \in L^1(\mu)$.

As an illustration of these basic properties, let us prove the following elementary observation. In the sequel, we will write $\mathrm{Var}_\mu(f) := \mu(f^2) - \mu(f)^2$.

Lemma 2.9. *Let μ be a stationary measure. Then $t \mapsto \mathrm{Var}_\mu(P_t f)$ is a decreasing function of time for every function $f \in L^2(\mu)$.*

Proof. Note that

$$\begin{aligned}\mathrm{Var}_\mu(P_t f) &= \|P_t f - \mu f\|^2_{L^2(\mu)} = \|P_t(f-\mu f)\|^2_{L^2(\mu)} = \|P_{t-s}P_s(f-\mu f)\|^2_{L^2(\mu)} \\ &\le \|P_s(f-\mu f)\|^2_{L^2(\mu)} = \|P_s f - \mu f\|^2_{L^2(\mu)} = \mathrm{Var}_\mu(P_s f)\end{aligned}$$

for every $0 \le s \le t$. □

We now turn to an important notion for Markov processes in continuous time. If you are familiar with Markov chains in discrete time with a finite state space, you will be used to the idea that the dynamics of the chain is defined in terms of a matrix of transition probabilities. This matrix describes with what probability the chain moves from one state to another in one time step, and forms the basic ingredient in the analysis of the behavior of Markov chains. This idea does not make sense in continuous time, as a Markov process evolves continuously and not in individual steps. Nonetheless, there is an object that plays the analogous role in continuous time, called the *generator* of a Markov process. We will first describe the general notion, and then investigate the finite state space case as an example (in which case the generator can be interpreted as a matrix of transition *rates* rather than probabilities).

From now on, we will fix a Markov process with stationary measure μ and consider $\{P_t\}_{t\in\mathbb{R}_+}$ as a semigroup of linear operators on $L^2(\mu)$.

Definition 2.10 (Generator). *The* generator $\mathscr{L}$ *is defined as*

$$\mathscr{L} f := \lim_{t\downarrow 0} \frac{P_t f - f}{t}$$

for every $f \in L^2(\mu)$ for which the above limit exists in $L^2(\mu)$. The set of f for which $\mathscr{L} f$ is defined is called the domain $\mathrm{Dom}(\mathscr{L})$ *of the generator, and $\mathscr{L}$ defines a linear operator from $\mathrm{Dom}(\mathscr{L}) \subseteq L^2(\mu)$ to $L^2(\mu)$.*

*Remark 2.11 (**Warning**).* For Markov processes whose sample paths are of pure jump type (i.e., piecewise constant as a function of time) it is often the case that $\mathrm{Dom}(\mathscr{L}) = L^2(\mu)$. This is the simplest setting for the theory of Markov processes in continuous time, and here many computations can be done without any technicalities. On the other hand, for Markov processes with continuous sample paths (such as Brownian motion, for example), it is an unfortunate fact of life that $\mathrm{Dom}(\mathscr{L}) \subsetneq L^2(\mu)$. In this setting, a rigorous treatment of semigroups, generators, and domains requires functional analytic machinery that is not assumed as a prerequisite for this course. While we should therefore ideally restrict attention to the pure jump case, many important applications (for example, the proof of the Poincaré inequality for Gaussian variables) will require the use of continuous Markov processes.

Fortunately, it turns out that domain problems prove to be of a purely technical nature in all the applications that we will encounter: results that we will derive for the case $\mathrm{Dom}(\mathscr{L}) = L^2(\mu)$ will be directly applicable even when this condition fails. While a rigorous proof would require to check carefully that no domain issues arise, addressing such issues would take significant time and does not provide much insight into the high-dimensional phenomena that are of interest in this course. As a compromise, we will therefore generally ignore domain problems and assume implicitly that $\mathrm{Dom}(\mathscr{L}) = L^2(\mu)$ when deriving general results, while we will still apply these results in more general cases. The interested reader should be aware when a shortcut is being taken, and refer to the literature for a careful treatment of such technical issues.

How can one use the generator $\mathscr{L}$? We have defined the generator in terms of the semigroup; however, it is in fact possible to define the semigroup in terms of the generator, in analogy to the definition of a discrete Markov chain in terms of its transition probability matrix. To see this, note that

$$\frac{d}{dt}P_t f = \lim_{\delta\downarrow 0}\frac{P_{t+\delta}f - P_t f}{\delta} = \lim_{\delta\downarrow 0} P_t\left(\frac{P_\delta f - f}{\delta}\right) = P_t\mathscr{L}f.$$

Thus P_t can be recovered as the solution of the *Kolmogorov equation*

$$\frac{d}{dt}P_t f = P_t\mathscr{L}f, \qquad P_0 f = f.$$

This computation could also have been performed in a different order:

$$\frac{d}{dt}P_t f = \lim_{\delta\downarrow 0}\frac{P_{t+\delta}f - P_t f}{\delta} = \lim_{\delta\downarrow 0}\frac{P_\delta P_t f - P_t f}{\delta} = \mathscr{L}P_t f.$$

Thus we have demonstrated a basic property: the generator and the semigroup commute, that is, $\mathscr{L}P_t = P_t\mathscr{L}$. [These statements are entirely clear when $\mathrm{Dom}(\mathscr{L}) = L^2(\mu)$, and must be given a careful interpretation otherwise.]

Example 2.12 (Finite state space). Let $(X_t)_{t\in\mathbb{R}_+}$ be a Markov process with values in a finite state space $X_t \in \{1,\dots,d\}$. Such processes are typically described in terms of their *transition rates* $\lambda_{ij} \geq 0$ for $i \neq j$:

$$\mathbf{P}[X_{t+\delta} = j | X_t = i] = \lambda_{ij}\delta + o(\delta) \quad \text{for } i \neq j.$$

Evidently, the transition rates λ_{ij} describe the infinitesimal rate of growth of the probability of jumping from state i to state j (informally, if $X_t = i$, then the probability that $X_{t+dt} = j$ is $\lambda_{ij}dt$).

Let us organize the transition probabilities $q_{t,ij} = \mathbf{P}[X_t = j | X_0 = i]$ and rates λ_{ij} into matrices $Q_t = (q_{t,ij})_{1\le i,j\le d}$ and $\Lambda = (\lambda_{ij})_{1\le i,j\le d}$, respectively, where we define the diagonal entries of Λ as $\lambda_{ii} = -\sum_{j\neq i}\lambda_{ij} \le 0$. Then

$$\lim_{t\downarrow 0} \frac{q_{t,ij} - q_{0,ij}}{t} = \lambda_{ij}$$

for every $1 \le i, j \le d$ (the diagonal entries λ_{ii} were chosen precisely to enforce the law of total probability $\sum_j q_{t,ij} = 1$). In particular, we have

$$\mathscr{L}f(i) = \lim_{t\downarrow 0} \sum_{j=1}^{d} f(j)\, \frac{q_{t,ij} - q_{0,ij}}{t} = \sum_{j=1}^{d} \lambda_{ij} f(j) = (\Lambda f)_i,$$

where we identify the function f with the vector $(f(1), \ldots, f(d)) \in \mathbb{R}^d$. We therefore conclude that the generator of a Markov process in a finite state space corresponds precisely to the matrix of transition rates. The Kolmogorov equation now reduces to the matrix differential equation

$$\frac{d}{dt} Q_t = Q_t \Lambda, \qquad Q_0 = I.$$

This differential equation is the basic tool for computing probabilities of finite state space Markov processes. The solution is in fact easily obtained as

$$Q_t = e^{t\Lambda},$$

from which we readily see why P_t and $\mathscr{L}$ must commute.

The above example provides some intuition for the notion of a generator. Further examples of Markov semigroups will be given in the next section.

Remark 2.13. In analogy with the above example, we can formally express the relation between the semigroup and generator of a Markov process as $P_t = e^{t\mathscr{L}}$. This expression is readily made precise in the case $\mathrm{Dom}(\mathscr{L}) = L^2(\mu)$ by interpreting $e^{t\mathscr{L}}$ as a power series. While this does not work in the case $\mathrm{Dom}(\mathscr{L}) \subsetneq L^2(\mu)$, the intuition extends also to this setting; however, in this case the meaning of the exponential function must be carefully defined.

We conclude this section by introducing one more fundamental idea in the theory of Markov processes. Recall that we have defined semigroup P_t as a family of linear operators on $L^2(\mu)$. The latter is a Hilbert space, and we denote its inner product as $\langle f, g\rangle_\mu := \mu(fg)$ (so that $\|f\|^2_{L^2(\mu)} = \langle f, f\rangle_\mu$).

Definition 2.14 (Reversibility). *The Markov semigroup P_t with stationary measure μ is called* reversible *if $\langle f, P_t g\rangle_\mu = \langle P_t f, g\rangle_\mu$ for every $f, g \in L^2(\mu)$.*

Thus the Markov process is reversible if the operators P_t are self-adjoint on $L^2(\mu)$. Equivalently, as $P_t = e^{t\mathscr{L}}$, the Markov process is reversible if its generator $\mathscr{L}$ is self-adjoint. The reversibility property has a probabilistic interpretation: if the Markov property is reversible, then (assuming $X_0 \sim \mu$)

$$\begin{aligned}\langle P_t f, g\rangle_\mu = \langle f, P_t g\rangle_\mu &= \mathbf{E}[f(X_0)\mathbf{E}[g(X_t)|X_0]] \\ &= \mathbf{E}[f(X_0)g(X_t)] = \mathbf{E}[\mathbf{E}[f(X_0)|X_t]g(X_t)]\end{aligned}$$

for every $f, g \in L^2(\mu)$, so that in particular

$$P_t f(x) := \mathbf{E}[f(X_t)|X_0 = x] = \mathbf{E}[f(X_0)|X_t = x].$$

This implies that when the Markov process $(X_t)_{t\in[0,a]}$ is viewed backwards in time $(X_{a-t})_{t\in[0,a]}$, it has the same law: that is, the law of the Markov process is invariant under time reversal; hence the name *reversibility*.

We will see in the following section that reversible Markov processes are the most natural objects connected to Poincaré inequalities (and to other functional inequalities that we will encounter in later chapters). However, the notion of time reversal will not play any role in our proofs. Rather, for reasons that will become evident in the next section, the self-adjointness of the generator $\mathscr{L}$ will allow us to obtain a very complete characterization of exponential convergence of the Markov semigroup to the stationary measure.

Example 2.15 (Finite state space continued). In the setting of Example 2.12, it is evident that the Markov process is reversible if and only if

$$\sum_{i,j=1}^d \mu_i f_i \Lambda_{ij} g_j = \sum_{i,j=1}^d \mu_j g_j \Lambda_{ji} f_i$$

for all $f, g \in \mathbb{R}^d$, or equivalently

$$\mu_i \Lambda_{ij} = \mu_j \Lambda_{ji} \quad \text{for all } i, j \in \{1, \ldots, d\},$$

where μ denotes the stationary measure of the Markov process. The latter condition is often called "detailed balance" in the physics literature.

Problems

2.6 (Some elementary identities). Let P_t be a Markov semigroup with generator $\mathscr{L}$ and stationary measure μ. Prove the following elementary facts:

a. Show that $\mu(\mathscr{L} f) = 0$ for every $f \in \mathrm{Dom}(\mathscr{L})$.
b. If $\phi : \mathbb{R} \to \mathbb{R}$ is convex, then $P_t\phi(f) \geq \phi(P_t f)$ when $f, \phi(f) \in L^2(\mu)$.
c. If $\phi : \mathbb{R} \to \mathbb{R}$ is convex, then $\mathscr{L}\phi(f) \geq \phi'(f)\mathscr{L} f$ when $f, \phi(f) \in \mathrm{Dom}(\mathscr{L})$.
d. Let $f \in \mathrm{Dom}(\mathscr{L})$. Show that the following process is a martingale:

$$M_t^f := f(X_t) - \int_0^t \mathscr{L} f(X_s)\, ds$$

2.3 Poincaré inequalities

Throughout this section, we fix a Markov semigroup P_t with generator $\mathscr{L}$ and stationary measure μ. As was discussed in the previous section, the stationary measure describes the "steady-state" behavior of the Markov process: that is, if $X_0 \sim \mu$, then $X_t \sim \mu$ for all times t. It is natural to ask whether the Markov process will in fact eventually end up in its steady state even if it is not started there, but rather at some fixed initial condition $X_0 = x$: that is, is it true that

$$\mathbf{E}[f(X_t)|X_0 = x] \to \mu f \quad \text{as } t \to \infty?$$

If this is the case, the Markov process is said to be *ergodic*. There are various different notions of ergodicity in the theory of Markov processes; as we are working in $L^2(\mu)$, the following will be natural for our purposes.

Definition 2.16 (Ergodicity). *The Markov semigroup is called* ergodic *if* $P_t f \to \mu f$ *in* $L^2(\mu)$ *as* $t \to \infty$ *for every* $f \in L^2(\mu)$.

Recall that a Poincaré inequality for μ is, informally, of the form

$$\text{“ variance}(f) \lesssim \mathbf{E}[\,\|\text{gradient}(f)\|^2\,]. \text{ ”}$$

At first sight, such an inequality has nothing to do with Markov processes. Remarkably, however, the validity of a Poincaré inequality for μ turns out to be intimately related to the *rate of convergence* of an ergodic Markov process for which μ is the stationary distribution. Still informally, we have the following:

> *A measure μ satisfies a Poincaré inequality for a certain notion of "gradient" if and only if an ergodic Markov semigroup associated to this "gradient" converges exponentially fast to μ.*

The following definition and result makes this principle precise.

Definition 2.17 (Dirichlet form). *Given a Markov process with generator* $\mathscr{L}$ *and stationary measure* μ*, the corresponding* Dirichlet form *is defined as*

$$\mathcal{E}(f,g) = -\langle f, \mathscr{L} g\rangle_\mu.$$

Theorem 2.18 (Poincaré inequality). *Let* P_t *be reversible ergodic Markov semigroup with stationary measure* μ*. The following are equivalent given* $c \geq 0$*:*

1. $\mathrm{Var}_\mu[f] \leq c\mathcal{E}(f,f)$ *for all* f (Poincaré inequality).
2. $\|P_t f - \mu f\|_{L^2(\mu)} \leq e^{-t/c}\|f - \mu f\|_{L^2(\mu)}$ *for all* f, t.
3. $\mathcal{E}(P_t f, P_t f) \leq e^{-2t/c}\mathcal{E}(f,f)$ *for all* f, t.
4. *For every* f *there exists* $\kappa(f)$ *such that* $\|P_t f - \mu f\|_{L^2(\mu)} \leq \kappa(f) e^{-t/c}$.
5. *For every* f *there exists* $\kappa(f)$ *such that* $\mathcal{E}(P_t f, P_t f) \leq \kappa(f) e^{-2t/c}$.

Remark 2.19. As will be seen in the proof of this Theorem, the implications $5 \Leftarrow 3 \Rightarrow 1 \Leftrightarrow 2 \Rightarrow 4$ remain valid even when P_t is not reversible. The remaining implications $5 \Rightarrow 3$, $4 \Rightarrow 2$ and $2 \Rightarrow 3$ require reversibility.

At this point, the interpretation of Theorem 2.18 is probably far from clear. There are several questions we must address:

- Why do we call $\mathrm{Var}_\mu[f] \le c\mathcal{E}(f,f)$ a Poincaré inequality? In what sense can $\mathcal{E}(f,f)$ be interpreted as an "expected square gradient" of f?
- It there any relation between Theorem 2.18 and the discrete Poincaré inequalities that we already derived in section 2.1?
- Why should we expect any connection between Poincaré inequalities and Markov processes in the first place?

The quickest way to get a feeling for the first two questions is to consider some illminating examples. To this end, we will devote the remainder of this section to developing two applications of Theorem 2.18. First, we will prove one of the most important examples of a Poincaré inequality, the *Gaussian Poincaré inequality*, using the machinery of Theorem 2.18. Along the way, we will introduce an important Markov process, the *Ornstein-Uhlenbeck process*, that will appear again in later chapters. Second, we will show that the tensorization inequality that we already proved in Theorem 2.3 is itself a special case of Theorem 2.18; this again requires the introduction of an suitable Markov process. Of course, this is not the easiest proof of the tensorization inequality, and it is not suggested that Theorem 2.18 should be used when an easier proof is available. Rather, this example highlights that Theorem 2.18 is not distinct from the inequalities that we developed in section 2.1, but rather provides a unified framework for all the Poincaré inequalities that we encounter.

The proof of Theorem 2.18 will be postponed to the next section. When we begin developing the proof, it will quickly become apparent why Poincaré inequalities are connected to Markov processes, and why $\mathrm{Var}_\mu[f] \le c\mathcal{E}(f,f)$ is the "right" notion of a Poincaré inequality. The ideas used in the proof are of interest in their own right and can be used to prove other interesting results.

Remark 2.20. The properties 2–5 of Theorem 2.18 should all be viewed as different notions of exponential convergence of the Markov semigroup P_t to the stationary measure μ. Properties 2 and 4 measure directly the rate of convergence of $P_t f$ to μf in $L^2(\mu)$ (cf. Definition 2.16). On the other hand, properties 3 and 5 measure the rate of convergence of the "gradient" of $P_t f$ to zero. As ergodicity implies that $P_t f(x)$ becomes insensitive to x as $t \to \infty$ (that is, the Markov process "forgets" its initial condition), the latter is also a natural formulation of the ergodicity property. The properties 4 and 5 are often easier to prove than properties 2 and 3, as they only require control of the rate of convergence and not of the constant in the inequality.

Remark 2.21. Let μ be a measure for which we would like to prove a Poincaré inequality. In order to apply Theorem 2.18, we must construct a suitable Markov process for which μ is the stationary measure. There is not a unique way to do this: there are many different Markov processes that admit the same stationary measure μ. However, each Markov process gives rise to a *different* Dirichlet form $\mathcal{E}(f,f)$, and thus to a Poincaré inequality for μ with respect to a

different notion of gradient! By choosing different Markov processes, Theorem 2.18 therefore provides us with a flexible mechanism to prove a whole family of different Poincaré inequalities for the same distribution μ.

Conversely, Theorem 2.18 can be used in the opposite direction. Suppose that we are interested in ergodicity of a given Markov process with stationary measure μ. If we can prove, by some means, that μ satisfies a Poincaré inequality with respect to the Dirichlet form induced by the given Markov process, then we have immediately established exponential convergence of the Markov process to its stationary measure. This is important in many applications, including nonequilibrium statistical mechanics and in the analysis of Markov Chain Monte Carlo algorithms for sampling from the stationary measure μ.

We now turn to the examples announced above. We begin with an important inequality that has many applications: the Gaussian Poincaré inequality.

Example 2.22 (Gaussian Poincaré inequality). Our aim is to obtain a Poincaré inequality for the standard Gaussian distribution $\mu = N(0,1)$ in one dimension (we can subsequently use tensorization to extend to higher dimensions). Of course, there is no unique Poincaré inequality: for example, the trivial Lemma 2.1 applies to the Gaussian distribution as it does to any other. However, we will see that for the Gaussian, we can obtain a nontrivial Poincaré inequality with respect to the classical calculus notion of gradient. This inequality is usually referred to as *the* Gaussian Poincaré inequality.

By Theorem 2.18, the key to obtaining a Poincaré inequality for μ with a specific notion of gradient is to construct a Markov process whose Dirichlet form corresponds to the desired notion of gradient and for which μ is the stationary distribution. For the Gaussian distribution, the appropriate Markov process is the *Ornstein-Uhlenbeck process*, which is one of the most important tools in the study of Gaussian distributions and which we will encounter again in later chapters. Given a standard Brownian motion $(W_t)_{t\in\mathbb{R}_+}$, the Ornstein-Uhlenbeck process can be defined as

$$X_t = e^{-t}X_0 + e^{-t}W_{e^{2t}-1}, \qquad X_0 \perp\!\!\!\perp W.$$

It is evident that if $X_0 \sim N(0,1)$, then $X_t \sim N(0,1)$ for all $t \in \mathbb{R}_+$. Let us collect some basic properties of the Ornstein-Uhlenbeck process.

Lemma 2.23 (Ornstein-Uhlenbeck process). *The process $(X_t)_{t\in\mathbb{R}_+}$ defined above is a Markov process with semigroup*

$$P_t f(x) = \mathbf{E}[f(e^{-t}x + \sqrt{1-e^{-2t}}\xi)], \qquad \xi \sim N(0,1).$$

The process admits $\mu = N(0,1)$ as its stationary measure and is ergodic. Moreover, its generator and Dirichlet form are given by

$$\mathscr{L}f(x) = -xf'(x) + f''(x), \qquad \mathcal{E}(f,g) = \langle f', g'\rangle_\mu.$$

In particular, the Ornstein-Uhlenbeck process is reversible.

Before we can prove this result, we need an elementary property of the Gaussian distribution: the Gaussian integration by parts formula.

Lemma 2.24 (Gaussian integration by parts). *If $\xi \sim N(0,1)$, then*

$$\mathbf{E}[\xi f(\xi)] = \mathbf{E}[f'(\xi)].$$

Proof. If f is smooth with compact support, then we have

$$\int_{-\infty}^{\infty} f'(x)\frac{e^{-x^2/2}}{\sqrt{2\pi}}dx = -\int_{-\infty}^{\infty} f(x)\left(\frac{d}{dx}\frac{e^{-x^2/2}}{\sqrt{2\pi}}\right)dx$$

by integration by parts, and the result follows readily. We can now extend to any f with $\xi f(\xi), f'(\xi) \in L^1(\mu)$ by a routine approximation argument. □

Proof (Lemma 2.23). Let $s \le t$. By the definition of X_t, we have

$$\begin{aligned} X_t &= e^{-(t-s)}X_s + e^{-t}(W_{e^{2t}-1} - W_{e^{2s}-1}) \\ &= e^{-(t-s)}X_s + \sqrt{1-e^{-2(t-s)}}\,\xi, \end{aligned}$$

where $\xi = (W_{e^{2t}-1} - W_{e^{2s}-1})/\sqrt{e^{2t}-e^{2s}} \sim N(0,1)$ is independent of $\{X_r\}_{r\le s}$. It follows immedately that we can write

$$\mathbf{E}[f(X_t)|\{X_r\}_{r\le s}] = P_{t-s}f(X_s),$$

with $P_t f$ as defined in the statement of the Lemma. In particular, $(X_t)_{t\ge 0}$ satisfies the Markov property. Moreover, it is evident by inspection that $\mu = N(0,1)$ is stationary and that the semigroup is ergodic.

With the semigroup in hand, we can now compute the generator and the Dirichlet form. To compute the generator, note that

$$\begin{aligned} \frac{d}{dt}P_t f(x) &= \mathbf{E}\left[f'(e^{-t}x + \sqrt{1-e^{-2t}}\xi)\left\{\frac{e^{-2t}}{\sqrt{1-e^{-2t}}}\,\xi - e^{-t}x\right\}\right] \\ &= \mathbf{E}[-e^{-t}xf'(e^{-t}x + \sqrt{1-e^{-2t}}\xi) + e^{-2t}f''(e^{-t}x + \sqrt{1-e^{-2t}}\xi)], \end{aligned}$$

where we have used Lemma 2.24 in the second line. We therefore have

$$\frac{d}{dt}P_t f(x) = \left\{-x\frac{d}{dx} + \frac{d^2}{dx^2}\right\}P_t f(x).$$

Letting $t \downarrow 0$ yields the expression for $\mathscr{L}$ given in the statement of the Lemma. To compute the Dirichlet form, it suffices to note that

$$\mathcal{E}(f,g) = -\langle f, \mathscr{L}g\rangle_\mu = \mathbf{E}[f(\xi)\{\xi g'(\xi) - g''(\xi)\}] = \mathbf{E}[f'(\xi)g'(\xi)],$$

where we have used Lemma 2.24 once more. Finally, $\langle f, \mathscr{L}g\rangle_\mu = \langle \mathscr{L}f, g\rangle_\mu$ as $\mathcal{E}(f,g)$ is symmetric, so the Ornstein-Uhlenbeck process is reversible. □

Remark 2.25. Our definition of the Ornstein-Uhlenbeck process may seem a little mysterious. Perhaps a more intuitive definition of the Ornstein-Uhlenbeck process is as the solution of the stochastic differential equation

$$dX_t = -X_t\,dt + \sqrt{2}\,dB_t,$$

where $(B_t)_{t\in\mathbb{R}_+}$ is standard Brownian motion: that is, the Ornstein-Uhlenbeck process is obtained by subjecting a Brownian motion to linear forcing that keeps it from going off to infinity. While this approach is more insightful and is more readily generalized to other distributions, our elementary approach has the advantage that it avoids the use of stochastic calculus.

From Lemma 2.23, it follows immediately that

$$\mathcal{E}(f,f) = \|f'\|^2_{L^2(\mu)} = \mathbf{E}[\{f'(\xi)\}^2], \qquad \xi \sim N(0,1).$$

Thus the Dirichlet form for the Ornstein-Uhlenbeck process *is precisely the expected square gradient* for the classical calculus notion of gradient! Thus an inequality of the form $\mathrm{Var}_\mu[f] \le c\mathcal{E}(f,f)$ is indeed a Poincaré inequality in the most classical sense. By Theorem 2.18, proving such an inequality is equivalent to proving exponential ergodicity of the Ornstein-Uhlenbeck process. With Lemma 2.23 in hand, this is a remarkably easy exercise.

Theorem 2.26. *Let* $\mu = N(0,1)$. *Then* $\mathrm{Var}_\mu[f] \le \|f'\|^2_{L^2(\mu)}$.

This is the Gaussian Poincaré inequality in one dimension.

Proof. It follows immediately from the expression for $P_t f$ in Lemma 2.23 that

$$\frac{d}{dx} P_t f(x) = e^{-t} P_t f'(x).$$

Thus

$$\begin{aligned}\mathcal{E}(P_t f, P_t f) = \|(P_t f)'\|^2_{L^2(\mu)} &= e^{-2t}\|P_t f'\|^2_{L^2(\mu)} \\ &\le e^{-2t}\|f'\|^2_{L^2(\mu)} = e^{-2t}\mathcal{E}(f,f).\end{aligned}$$

The result follows by the implication $3 \Rightarrow 1$ of Theorem 2.18. □

Remark 2.27. Let us emphasize once more that there is nothing special about the Ornstein-Uhlenbeck process *per se* in the context of Theorem 2.18: there are many Markov processes for which $\mu = N(0,1)$ is stationary. Different Markov processes could be used to prove different Poincaré inequalities for the Gaussian distribution for different notions of gradient. What singles out the Ornstein-Uhlenbeck process is that its Dirichlet form $\mathcal{E}(f,f) = \|f'\|^2_{L^2(\mu)}$ is precisely given in terms of the classical calculus notion of gradient, which provides a particularly useful tool in many applications.

Having proved the Gaussian Poincaré inequality in one dimension, we immediately obtain an n-dimensional inequality by tensorization. As this is a very useful inequality in applications, let us state it as a theorem. [We could also have proved this directly without tensorization using an n-dimensional Ornstein-Uhlenbeck process, but this does not add much additional insight.]

Corollary 2.28 (Gaussian Poincaré inequality). *Let $X_1, \ldots, X_n$ be independent Gaussian random variables with zero mean and unit variance. Then*

$$\mathrm{Var}[f(X_1, \ldots, X_n)] \le \mathbf{E}[\|\nabla f(X_1, \ldots, X_n)\|^2].$$

We now turn to our second example: we will show that the tensorization inequality of Theorem 2.3 is a special case of Theorem 2.18. Thus the connection between Poincaré inequalities and Markov semigroups captures in a unified framework all of the inequalities that we have seen so far.

Example 2.29 (Tensorization revisited). Let $\mu = \mu_1 \otimes \cdots \otimes \mu_n$ be any product measure. We aim to investigate the tensorization inequality of Theorem 2.3 from the viewpoint of Theorem 2.18. To this end, we begin by constructing a Markov process for which μ is stationary and whose Dirichlet form corresponds to the right-hand side of the tensorization inequality.

Let $X_t = (X_t^1, \ldots, X_t^n)_{t \in \mathbb{R}_+}$ be a random process constructed as follows. To each coordinate $i = 1, \ldots, n$, we attach an independent Poisson process N_t^i with unit rate. The Poisson process should be viewed as a random clock attached to each coordinate that "ticks" whenever N_t^i jumps. The process $(X_t)_{t \in \mathbb{R}_+}$ is now constructed by the following mechanism:

- Draw $X_0 \sim \mu$ independently from the Poisson process $N = (N^1, \ldots, N^n)$.
- Each time N_t^i jumps for some i, replace the current value of X_t^i by an independent sample from μ_i while keeping the remaining coordinates fixed.

As the Poisson process has independent increments, it is easily verified that $(X_t)_{t \in \mathbb{R}_+}$ satisfies the Markov property and that μ is stationary.

Let us now compute the semigroup of $(X_t)_{t \in \mathbb{R}_+}$. By construction,

$$P_t f(x) = \mathbf{E}[f(X_t)|X_0 = x] =$$

$$\sum_{I \subseteq \{1,\ldots,n\}} \mathbf{P}[N_t^i > 0 \text{ for } i \in I,\ N_t^i = 0 \text{ for } i \notin I] \int f(x_1, \ldots, x_n) \prod_{i \in I} \mu_i(dx_i) =$$

$$\sum_{I \subseteq \{1,\ldots,n\}} (1 - e^{-t})^{|I|} e^{-t(n-|I|)} \int f(x_1, \ldots, x_n) \prod_{i \in I} \mu_i(dx_i).$$

In particular, we can compute the generator as

$$\mathscr{L} f = \lim_{t \downarrow 0} \frac{P_t f - f}{t} = -\sum_{i=1}^n \delta_i f,$$

where we have introduced the notation

$$\delta_i f(x) := f(x) - \int f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n)\, \mu_i(dz).$$

Finally, let us compute the Dirichlet form

$$\mathcal{E}(f,g) = \sum_{i=1}^n \int f \delta_i g \, d\mu = \sum_{i=1}^n \int \delta_i f \, \delta_i g \, d\mu,$$

where we have used that $\int h\, \delta_i g \, d\mu = 0$ if $h(x)$ does not depend on x_i. As $\mathcal{E}(f,g)$ is symmetric, it follows that our Markov process is reversible.

Now note that

$$\mathcal{E}(f,f) = \sum_{i=1}^n \int (\delta_i f)^2 \, d\mu = \sum_{i=1}^n \int \mathrm{Var}_i f \, d\mu.$$

Thus the tensorization inequality of Theorem 2.3 can be expressed as

$$\mathrm{Var}_\mu[f] \le \mathcal{E}(f,f),$$

and we therefore conclude that tensorization is nothing but a special case of Theorem 2.18. In fact, given that we already proved the tensorization inequality, we could now invoke Theorem 2.18 to conclude immediately that our Markov process is exponentially ergodic in the sense that

$$\|P_t f - \mu f\|_{L^2(\mu)} \le e^{-t} \|f - \mu f\|_{L^2(\mu)}.$$

Conversely, if we can give a direct proof of exponential ergodicity of our Markov process, then we obtain by Theorem 2.18 an alternative proof of the tensorization inequality. Let us provide such a proof for sake of illustration. From the explicit formula for $P_t f$ above, it follows that

$$\delta_i P_t f = e^{-t} \sum_{I \not\ni i} (1 - e^{-t})^{|I|} e^{-t(n-1-|I|)} \int \delta_i f(x_1, \ldots, x_n) \prod_{i \in I} \mu_i(dx_i).$$

Evidently each term in the sum has $L^2(\mu)$-norm at most $\|\delta_i f\|_{L^2(\mu)}$, so

$$\mathcal{E}(P_t f, P_t f) = \sum_{i=1}^n \|\delta_i P_t f\|^2_{L^2(\mu)} \le \kappa(f) e^{-2t}$$

for some $\kappa(f) < \infty$ for every $f \in L^2(\mu)$. The tensorization inequality of Theorem 2.3 therefore follows from the implication $5 \Rightarrow 1$ of Theorem 2.18.

Problems

2.7 (Carré du champ). We have interpreted the Dirichlet form $\mathcal{E}(f,f)$ as a general notion of "expected square gradient" that arises in the study of

Poincaré inequalities. There is an analogous quantity $\Gamma(f,f)$ that plays the role of "square gradient" in this setting (without the expectation). In good probabilistic tradition, it is universally known by its French name *carré du champ* (literally, "square of the field"). The carré du champ is defined as

$$\Gamma(f,g) := \frac{1}{2}\{\mathscr{L}(fg) - f\mathscr{L}g - g\mathscr{L}f\}$$

in terms of the generator $\mathscr{L}$ of a Markov process with stationary measure μ.

a. Show that $\mathcal{E}(f,f) = \int \Gamma(f,f)\,d\mu$, and that $\mathcal{E}(f,g) = \int \Gamma(f,g)\,d\mu$ if the Markov process is in addition reversible.
b. Show that $\Gamma(f,f) \geq 0$, so it can indeed be interpreted as a square.
Hint: use $P_t(f^2) \geq (P_t f)^2$ and the definition of $\mathscr{L}$.
c. Prove the Cauchy-Schwarz inequality $\Gamma(f,g)^2 \leq \Gamma(f,f)\Gamma(g,g)$.
Hint: use that $\Gamma(f+tg, f+tg) \geq 0$ for all $t \in \mathbb{R}$.
d. Compute the carré du champ in the various examples of Poincaré inequalities encountered in this chapter, and convince yourself that it should indeed be interpreted as the appropriate notion of "square gradient" in each case.

2.8 (Gaussian Poincaré inequality). The goal of this problem is to develop some simple consequences and insights for the Gaussian Poincaré inequality.

a. Let $X_1, \ldots, X_n$ be i.i.d. standard Gaussians. Show that if f is L-Lipschitz, that is, $|f(x) - f(y)| \leq L\|x - y\|$, then $\mathrm{Var}[f(X_1, \ldots, X_n)] \leq L^2$.

Remark. The power of the above inequality is its dimension-free nature: it depends only on the degree of smoothness of f and not on the dimension n.

b. Let $X \sim N(0, \Sigma)$ be an n-dimensional centered Gaussian vector with arbitrary covariance matrix Σ. Prove the following useful identity:

$$\mathrm{Var}\Big[\max_{i=1,\ldots,n} X_i\Big] \leq \max_{i=1,\ldots,n} \mathrm{Var}[X_i].$$

Hint: write $X = \Sigma^{1/2} Y$ where $Y_1, \ldots, Y_n$ are i.i.d. standard Gaussians.

c. By a miracle, it is possible to derive the Gaussian Poincaré inequality from the bounded difference inequality of Corollary 2.4. To this end, let ε_{ji} be i.i.d. symmetric Bernoulli variables. By the central limit theorem,

$$f\left(\frac{1}{\sqrt{k}}\sum_{i=1}^k \varepsilon_{1i}, \ldots, \frac{1}{\sqrt{k}}\sum_{i=1}^k \varepsilon_{ni}\right) \Longrightarrow f(X_1, \ldots, X_n)$$

in distribution as $k \to \infty$ when f is a bounded continuous function and $X_1, \ldots, X_n$ are i.i.d. standard Gaussians. Apply the bounded difference inequality to the left-hand side and use Taylor expansion to provide an alternative proof the Gaussian Poincaré inequality of Corollary 2.28.

Remark. The central limit theorem proof of the Gaussian Poincaré inequality is very specific to the Gaussian distribution. While it works in this particular case, the proof we have given above using the Ornstein-Uhlenbeck semigroup is much more insightful and can be extended to other distributions (for example, to log-concave distributions as in Problem 2.13 below).

2.9 (Exponential distribution). Let $\mu(dx) = \mathbf{1}_{x\geq 0} e^{-x} dx$ be the one-sided exponential distribution. In this problem, we will derive two different (and not directly comparable) Poincaré inequalities for the distribution μ.

a. Show that

$$\mathrm{Var}_\mu[f] \leq \mathbf{E}[\xi |f'(\xi)|^2], \qquad \xi \sim \mu.$$

Hint: show that $\xi \sim (X^2 + Y^2)/2$ where X, Y are i.i.d. $N(0,1)$.

b. Show that

$$\mathrm{Var}_\mu[f] \leq 4\,\mathbf{E}[|f'(\xi)|^2], \qquad \xi \sim \mu.$$

Hint: use $\int_0^\infty g(x)\, e^{-x}\, dx = g(0) + \int_0^\infty g'(x)\, e^{-x}\, dx$ with $g = (f - f(0))^2$.

These two distinct Poincaré inequalities correspond to two distinct Markov processes. For the two Markov processes defined below, show that their Dirichlet forms do indeed yield the two distinct Poincaré inequalities above:

c. The solution of the Cox-Ingersoll-Ross stochastic differential equation

$$dX_t = 2(1 - X_t)\, dt + 2\sqrt{X_t}\, dB_t,$$

which is a Markov process on $\mathbb{R}_+$ with generator

$$\mathscr{L} f(x) = 2(1 - x) f'(x) + 2x f''(x).$$

d. The solution of the stochastic differential equation

$$dX_t = -\mathrm{sign}(X_t)\, dt + \sqrt{2}\, dB_t,$$

which is a Markov process on $\mathbb{R}$ with generator

$$\mathscr{L} f(x) = -\mathrm{sign}(x) f'(x) + f''(x).$$

This process has the two-sided exponential measure $\mu(dx) = \frac{1}{2} e^{-|x|} dx$ as its stationary distribution, but the one-sided Poincaré inequality is easily deduced from it. Alternatively, one can obtain the one-sided inequality directly by considering the above stochastic differential equation with reflection at 0 (i.e., a Brownian motion with negative drift reflected at 0).

Remark. In Problem 2.12 below, we will encounter yet another distinct Poincaré inequality for the exponential distribution.

2.10 (Dependent random signs). Let $X_1, \ldots, X_n$ be random variables with values in $\{-1, 1\}$ whose joint distribution is denoted by μ. In this problem, *we do not assume that* $X_1, \ldots, X_n$ *are independent*. Thus we cannot use tensorization. Nonetheless, we expect that if $X_1, \ldots, X_n$ are "weakly dependent" then the concentration phenomenon should still arise. We are going to use Theorem 2.18 to develop a precise statement along these lines.

Define the *influence coefficient* of variable j on variable i as

$$C_{ij} := \max_{x \in \{-1,1\}^{n-2}} |\mathbf{P}[X_i = 1 | X_j = 1, \{X_k\}_{k \neq i,j} = x] - \mathbf{P}[X_i = 1 | X_j = -1, \{X_k\}_{k \neq i,j} = x]|$$

for $i \neq j$, and let $C_{ii} = 0$. If the random variables $X_1, \ldots, X_n$ are weakly dependent, then all the influences C_{ij} should be small. The goal of this problem is to prove the following Poincaré inequality:

$$(1 - \|C\|_{\mathrm{sp}}) \operatorname{Var}[f(X_1, \ldots, X_n)] \leq \mathbf{E}\left[\sum_{i=1}^{n} \operatorname{Var}[f(X_1, \ldots, X_n) | \{X_k\}_{k \neq i}]\right],$$

where $\|C\|_{\mathrm{sp}}$ denotes the spectral radius of the matrix C. If $X_1, \ldots, X_n$ are independent, then $C \equiv 0$ and this dependent Poincaré inequality reduces to the tensorization inequality for independent random variables.

The basic idea is to mimick the Markov process construction that we introduced above to prove tensorization. To this end, we attach to every coordinate $i = 1, \ldots, n$ an independent Poisson process N_t^i with unit rate. The random process $Z_t = (Z_t^1, \ldots, Z_t^n)_{t \in \mathbb{R}_+}$ is now constructed as follows:

- Draw $Z_0 \sim \mu$ independently from the Poisson processes $N_t^1, \ldots, N_t^n$.
- Each time N_t^i jumps for some i, replace the current value of Z_t^i by an independent sample from $\mu_i(dx_i | Z_t)$ while keeping the remaining coordinates fixed, where $\mu_i(dx_i|x) := \mathbf{P}[X_i \in \cdot \,| \{X_k\}_{k \neq i} = \{x_k\}_{k \neq i}]$.

The process Z_t is called a *Gibbs sampler* or *Glauber dynamics* for μ.

a. Show that $(Z_t)_{t \in \mathbb{R}_+}$ is Markov and that μ is stationary.

b. Show that the generator of Z_t is given by

$$\mathscr{L} f = -\sum_{i=1}^{n} \delta_i f, \qquad \delta_i f(x) := f(x) - \int f(x)\, \mu_i(dx_i|x),$$

and that the Dirichlet form is given by

$$\mathcal{E}(f, g) = \sum_{i=1}^{n} \int \delta_i f \, \delta_i g \, d\mu.$$

In particular, conclude that $(Z_t)_{t \in \mathbb{R}_+}$ is reversible.

We are now going to show that the Markov semigroup is exponentially ergodic.

c. Define the local oscillation

$$\Delta_i f := \max_{x\in\{-1,1\}^n} |f(x_1,\ldots,x_{i-1},1,x_{i+1},\ldots,x_n) - f(x_1,\ldots,x_{i-1},-1,x_{i+1},\ldots,x_n)|.$$

Show that for $i \neq j$

$$\Delta_j \int f\, d\mu_i \leq \Delta_j f + \Delta_i f\, C_{ji}.$$

d. Deduce from the above inequality that

$$\Delta_j\left(f + \frac{t}{n}\mathscr{L}f\right) \leq \left(1-\frac{t}{n}\right)\Delta_j f + \frac{t}{n}\sum_{i=1}^n \Delta_i f\, C_{ij},$$

or, in terms of the vector $\Delta f := (\Delta_1 f,\ldots,\Delta_n f)$ of local oscillations,

$$\Delta(f + t\mathscr{L}f/n) \leq \Delta f\,\{I - t(I-C)/n\}.$$

e. Show using the power series identity $e^{t\mathscr{L}} = \lim_{n\to\infty}(I + t\mathscr{L}/n)^n$ that

$$\Delta P_t f \leq \Delta f\, e^{-t(I-C)}.$$

f. Complete the proof of the Poincaré inequality (use Theorem 2.18, $5 \Rightarrow 1$).

Remark. The dependent Poincaré inequality extends readily to non-binary random variables (i.e., not in $\{-1,1\}$), provided C_{ij} are suitably redefined.

2.4 Variance identities and exponential ergodicity

The goal of this section is to prove Theorem 2.18, which connects the Poincaré inequality to the exponential ergodicity of a Markov semigroup. At first sight, it is far from clear why Markov semigroups should even enter the picture: what is the relation between $\mathrm{Var}_\mu[f]$ and $\mathcal{E}(f,f)$? In fact, the connection between these quantities is almost trivial, as is shown in the following lemma. Once this connection has been realized, Theorem 2.18 loses most of its mystery.

Lemma 2.30. *The following identity holds:*

$$\frac{d}{dt}\mathrm{Var}_\mu[P_t f] = -2\mathcal{E}(P_t f, P_t f).$$

Proof. Since $\mu(P_t f) = \mu(f)$,

$$\begin{aligned}\frac{d}{dt}\mathrm{Var}_\mu[P_t f] &= \frac{d}{dt}\{\mu((P_t f)^2) - (\mu P_t f)^2\} \\ &= \frac{d}{dt}\mu((P_t f)^2) \\ &= \mu\left(2P_t f \frac{d}{dt} P_t f\right) \\ &= \mu\left(2P_t f \mathscr{L} P_t f\right),\end{aligned}$$

and the result follows from the definition of the Dirichlet form. □

Simple as this result is, it yields many important consequences. Let us record two immediate observations for future reference.

Corollary 2.31. $\mathcal{E}(f, f) \geq 0$ *for every* f.

Proof. Immediate from Lemmas 2.9 and 2.30. □

Corollary 2.32 (Integral representation of variance). *Suppose that the Markov semigroup is ergodic. Then we have for every* f

$$\mathrm{Var}_\mu[f] = 2\int_0^\infty \mathcal{E}(P_t f, P_t f)\, dt.$$

Proof. Note that $P_t f \to \mu f$ implies $\mathrm{Var}_\mu[P_t f] \to \mathrm{Var}_\mu[\mu f] = 0$. Thus

$$\mathrm{Var}_\mu[f] = \mathrm{Var}_\mu[P_0 f] - \lim_{t\to\infty} \mathrm{Var}_\mu[P_t f] = -\int_0^\infty \frac{d}{dt}\mathrm{Var}_\mu[P_t f]\, dt$$

by the fundamental theorem of calculus. Now use Lemma 2.30. □

Remark 2.33. Integral representations of the variance such as the expression in Corollary 2.32 can be very useful in different settings. We will encounter some alternative integral representations in the problems below.

We are now ready to prove the implications $5 \Leftarrow 3 \Rightarrow 1 \Leftrightarrow 2 \Rightarrow 4$ of Theorem 2.18 that do not require reversibility. In fact, once the simple observations made above have been realized, these implications are entirely elementary.

Proof (Theorem 2.18, Part I). The implications $2 \Rightarrow 4$ and $3 \Rightarrow 5$ are trivial. We proceed to consider the remaining implications.

- $3 \Rightarrow 1$: Assuming 3, we have by Corollary 2.32

$$\mathrm{Var}_\mu[f] \leq 2\mathcal{E}(f, f)\int_0^\infty e^{-2t/c}\, dt = c\mathcal{E}(f, f).$$

- $1 \Rightarrow 2$: Assuming 1, we have by Lemma 2.30

$$\frac{d}{dt}\mathrm{Var}_\mu[P_t f] \le -\frac{2}{c}\,\mathrm{Var}_\mu[P_t f],$$

from which we obtain

$$\|P_t f - \mu f\|^2_{L^2(\mu)} = \mathrm{Var}_\mu[P_t f] \le e^{-2t/c}\mathrm{Var}_\mu[f] = e^{-2t/c}\|f - \mu f\|^2_{L^2(\mu)}.$$

- $2 \Rightarrow 1$: Assuming 2, we obtain using Lemma 2.30

$$2\mathcal{E}(f,f) = \lim_{t\downarrow 0}\frac{\mathrm{Var}_\mu[f] - \mathrm{Var}_\mu[P_t f]}{t} \ge \lim_{t\downarrow 0}\frac{1-e^{-2t/c}}{t}\,\mathrm{Var}_\mu[f] = \frac{2}{c}\,\mathrm{Var}_\mu[f].$$

This completes the proof of the implications $5 \Leftarrow 3 \Rightarrow 1 \Leftrightarrow 2 \Rightarrow 4$. □

It remains to prove the implications $2 \Rightarrow 3$, $5 \Rightarrow 3$, and $4 \Rightarrow 2$ of Theorem 2.18. These implications require reversibility, which we have not yet exploited. It turns out that reversibility implies a much finer property of the variance as a function of time than was obtained in Lemma 2.30. The appropriate property is contained in the following useful lemma.

Lemma 2.34. *If the Markov semigroup P_t is reversible, then the functions $t \mapsto \log \|P_t f\|^2_{L^2(\mu)}$ and $t \mapsto \log \mathcal{E}(P_t f, P_t f)$ are convex.*

Proof. Since $\mathscr{L}$ is self-adjoint, we have

$$\begin{aligned}\frac{d}{dt}\mathcal{E}(P_t f, P_t f) &= -\frac{d}{dt}\langle P_t f, \mathscr{L} P_t f\rangle_\mu \\ &= -\langle \mathscr{L} P_t f, \mathscr{L} P_t f\rangle_\mu - \langle P_t f, \mathscr{L}^2 P_t f\rangle_\mu \\ &= -2\|\mathscr{L} P_t f\|^2_{L^2(\mu)}.\end{aligned}$$

A straightforward computation yields

$$\begin{aligned}\frac{d^2}{dt^2}\log\|P_t f\|^2_{L^2(\mu)} &= \frac{4\|\mathscr{L}P_t f\|^2_{L^2(\mu)}}{\|P_t f\|^2_{L^2(\mu)}} - \frac{4\mathcal{E}(P_t f, P_t f)^2}{\|P_t f\|^4_{L^2(\mu)}} \\ &= \frac{4}{\|P_t f\|^4_{L^2(\mu)}}\Big\{\|\mathscr{L}P_t f\|^2_{L^2(\mu)}\|P_t f\|^2_{L^2(\mu)} - \langle P_t f, \mathscr{L}P_t f\rangle^2_\mu\Big\}.\end{aligned}$$

As the right-hand side is nonnegative by the Cauchy-Schwarz inequality, and we have shown that the function $t \mapsto \log\|P_t f\|^2_{L^2(\mu)}$ is convex. The proof for $t \mapsto \log \mathcal{E}(P_t f, P_t f)$ is entirely analogous, once we observe that the Dirichlet form also satisfies the Cauchy-Schwarz inequality $\mathcal{E}(f,g)^2 \le \mathcal{E}(f,f)\mathcal{E}(g,g)$ (to prove this, use that $\mathcal{E}(f+tg, f+tg) \ge 0$ for all $t \in \mathbb{R}$ by Corollary 2.31). □

We can now complete the proof of Theorem 2.18.

Proof (Theorem 2.18, Part II). We first prove $2 \Rightarrow 3$. By Lemma 2.34,

$$t \mapsto \frac{d}{dt} \log \|P_t f\|^2_{L^2(\mu)} = -\frac{2\mathcal{E}(P_t f, P_t f)}{\|P_t f\|^2_{L^2(\mu)}}$$

is increasing. In particular, we have

$$-\frac{2\mathcal{E}(P_t f, P_t f)}{\|P_t f\|^2_{L^2(\mu)}} \geq -\frac{2\mathcal{E}(f,f)}{\|f\|^2_{L^2(\mu)}}.$$

Rearranging this inequality yields

$$\frac{\mathcal{E}(P_t f, P_t f)}{\mathcal{E}(f,f)} \leq \frac{\|P_t f\|^2_{L^2(\mu)}}{\|f\|^2_{L^2(\mu)}}.$$

Thus property 3 follows readily from property 2 in Theorem 2.18.

It remains to prove $4 \Rightarrow 2$ and $5 \Rightarrow 3$. In fact, both these implications follow immediately from Lemma 2.34 by applying the following lemma to the functions $t \mapsto \log \|P_t f\|^2_{L^2(\mu)}$ and $t \mapsto \log \mathcal{E}(P_t f, P_t f)$ are convex. □

Lemma 2.35. *If the function $g : \mathbb{R}_+ \to \mathbb{R}$ is convex and $g(t) \leq K - \alpha t$ for all $t \geq 0$, then in fact $g(t) \leq g(0) - \alpha t$ for all $t \geq 0$.*

Proof. It suffices to show that the assumption implies that $g'(t) \leq -\alpha$ for all $t \geq 0$. Suppose that this is not the case. Then there exists $s \geq 0$ such that $g'(s) = -\beta > -\alpha$. As g is convex, g' is increasing and thus $g'(t) \geq -\beta$ for all $t \geq s$. In particular, it follows that $g(t) \geq g(s) - \beta t$ for all $t \geq s$. As $\beta < \alpha$, this contradicts the assumption that $g(t) \leq K - \alpha t$ for all $t \geq 0$. □

Remark 2.36 (Finite state space and spectral gaps). While the elementary implications in Theorem 2.18 are entirely intuitive, the role of reversibility in the remaining implications may not be entirely obvious: indeed, Lemma 2.34, which containes the essence of the reversibility argument, appears as a bit of a miracle. The aim of this remark is to highlight a complementary viewpoint on Theorem 2.18 that sheds additional light on the interpretation of the Poincaré inequality and on the role of reversibility. While this viewpoint can be developed more generally, we restrict attention for simplicity to the setting of finite state Markov processes as in Examples 2.12 and 2.15 above.

Let $(X_t)_{t \in \mathbb{R}_+}$ be a Markov process in a finite state space $X_t \in \{1, \ldots, d\}$. Denote by Λ the transition rate matrix, by μ the stationary measure, and let us assume that the reversibility condition $\mu_i \Lambda_{ij} = \mu_j \Lambda_{ji}$ holds. For notational simplicity, we will implicitly identify functions and measures on $\{1, \ldots, d\}$ with vectors in $\mathbb{R}^d$ in the obvious fashion. Note that we can write

$$\begin{aligned}\mathcal{E}(f,g) &= -\sum_{i,j=1}^d \mu_i f_i \Lambda_{ij} g_j = \sum_{i,j=1}^d \mu_i f_i \Lambda_{ij} (g_i - g_j) \\ &= \frac{1}{2} \sum_{i,j=1}^d \mu_i \Lambda_{ij} (f_i - f_j)(g_i - g_j),\end{aligned}$$

where we have used $\sum_j \Lambda_{ij} = 0$ in the second equality and that $\mu_i \Lambda_{ij}(g_i - g_j)$ is a skew-symmetric matrix in the third equality. In particular, we have

$$\mathcal{E}(f,f) = \frac{1}{2}\sum_{i,j=1}^{d} \mu_i \Lambda_{ij}(f_i - f_j)^2.$$

Again, $\mathcal{E}(f,f)$ can be naturally interpreted as an expected square gradient.

Let us now consider the Poincaré inequality from the point of view of linear algebra. As the matrix Λ is self-adjoint with respect to the weighted inner product $\langle\cdot,\cdot\rangle_\mu$, it has real eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ and associated eigenvectors $v_1, \ldots, v_d$. The property $\mathcal{E}(f,f) = -\langle f, \Lambda f\rangle_\mu \geq 0$ evidently implies that $\lambda_1 \leq 0$, that is, all the eigenvalues of Λ are nonpositive. Moreover, the property $\sum_j \Lambda_{ij} = 0$ implies that $v_1 = \mathbf{1}$ (the vector of ones) is an eigenvector with maximal eigenvalue $\lambda_1 = 0$. If $\mu f = \langle \mathbf{1}, f\rangle_\mu = 0$, we have

$$\mathcal{E}(f,f) = -\langle f, \Lambda f\rangle_\mu \geq -\lambda_2 \langle f, f\rangle_\mu = (\lambda_1 - \lambda_2)\,\mathrm{Var}_\mu[f],$$

and this inequality is tight for $f = v_2$. Thus *the best constant in the Poincaré inequality is the spectral gap* $\lambda_1 - \lambda_2$ *of the generator* Λ. For this reason, Poincaré inequalities are sometimes called spectral gap inequalities.

We can now also understand why the Poincaré inequality is so closely related to exponential convergence of the Markov semigroup. Indeed, let f be any function, and expand it in the eigenbasis of Λ as

$$f = \sum_{i=1}^{d} a_i v_i.$$

Then the semigroup acts on f as

$$P_t f = e^{t\Lambda} f = \sum_{i=1}^{d} e^{\lambda_i t} a_i v_i.$$

As $\lambda_1 = 0$, we have

$$\sup_f \frac{\|P_t f - \mu f\|^2_{L^2(\mu)}}{\|f - \mu f\|^2_{L^2(\mu)}} = \sup_f \frac{\sum_{i=2}^d e^{2\lambda_i t} a_i^2}{\sum_{i=2}^d a_i^2} = e^{-2(\lambda_1 - \lambda_2)t}.$$

Thus the spectral gap $\lambda_1 - \lambda_2$ controls precisely the exponential convergence rate of the semigroup. The various implications of Theorem 2.18 now become rather elementary from the linear algebra viewpoint. However, the fact that these equivalences can be proved hinges from the outset on the fact that Λ admits a spectral decomposition into eigenvectors with real-valued eigenvalues. This explains why reversibility of the semigroup (that is, the self-adjointness of Λ) is essential to obtain a complete set of equivalences in Theorem 2.18, despite that this fact was not entirely explicit in our general proof given above.

Problems

2.11 (Covariance identities). Let P_t be a reversible ergodic Markov semigroup with stationary measure μ. The goal of this problem is to prove useful integral representations of the covariance $\mathrm{Cov}_\mu(f,g) := \langle f - \mu f, g - \mu g\rangle_\mu$.

a. Prove the following identity:

$$\mathrm{Cov}_\mu(f,g) = 2\int_0^\infty \mathcal{E}(P_t f, P_t g)\,dt.$$

b. Prove the following identity:

$$\mathrm{Cov}_\mu(f,g) = \int_0^\infty \mathcal{E}(f, P_t g)\,dt.$$

c. Let $X \sim N(0,\Sigma)$ be a centered Gaussian vector in $\mathbb{R}^n$ with covariance matrix Σ. Assume that that the entries are positively correlated, that is, $\Sigma_{ij} \geq 0$ for all i,j. Prove that this implies the following much stronger *positive association* property: for every pair of functions f,g that are coordinatewise increasing, we have $\mathrm{Cov}(f(X),g(X)) \geq 0$.

Hint: write $X = \Sigma^{1/2}Y$ for $Y \sim N(0,I)$, and apply one of the above identities for the n-dimensional Ornstein-Uhlenbeck process (which is defined in the precisely the same manner as the one-dimensional Ornstein-Uhlenbeck process but using an n-dimensional Brownian motion).

2.12 (Local Poincaré inequalities I). We have seen that the validity of a Poincaré inequality for a given distribution μ is intimately connected with exponential ergodicity of Markov processes that admit μ as the stationary measure. In this problem, we will develop a method to deduce Poincaré inequalities for the distribution of the Markov process X_t at a *finite* time t, rather than for the stationary distribution (which is obtained as $t \to \infty$). In most cases, the stationary case is more useful, as it is much easier to construct a Markov process that admits a given measure μ as its stationary measure than to construct a Markov process that has distribution μ at a finite time. Nonetheless, there are several situations in which such *local* Poincaré inequalities are useful. In the following problem, we will see that this viewpoint provides significant insight even on the stationary case.

Let P_t be a Markov semigroup with generator $\mathscr{L}$. For the purposes of this problem, we do not assume the existence of a stationary measure.

a. Prove the following variance identity:

$$P_t(f^2) - (P_t f)^2 = 2\int_0^t P_{t-s}\Gamma(P_s f, P_s f)\,ds,$$

where we recall the definition of the carré du champ (Problem 2.7)

$$\Gamma(f,g) := \frac{1}{2}\{\mathscr{L}(fg) - f\mathscr{L}g - g\mathscr{L}f\}.$$

Hint: apply the fundamental theorem of calculus to $P_{t-s}((P_s f)^2)$.

b. Suppose that we can prove an identity of the form

$$\Gamma(P_s f, P_s f) \le \alpha(s)\, P_s \Gamma(f,f)$$

for some function $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$. Conclude that

$$P_t(f^2) - (P_t f)^2 \le c(t)\, P_t \Gamma(f,f), \qquad c(t) = \int_0^t 2\alpha(s)\, ds.$$

Such an inequality is called a *local Poincaré inequality*.

c. Let $(W_t)_{t\in\mathbb{R}_+}$ be standard Brownian motion. Brownian motion is itself a Markov process. Compute an explicit expression for its semigroup and generator (in analogy with Lemma 2.23), and show that in this case

$$\Gamma(P_t f, P_t f) \le P_t \Gamma(f,f).$$

Show that the local Poincaré inequality consequently provides a alternative proof of the Gaussian Poincaré inequality using Brownian motion.

d. The present approach provides a convenient method to derive Poincaré inequalities for infinitely divisible distributions (this part requires some familiarity with Lévy processes). Let ν be a positive measure on $\mathbb{R}$ such that $\int_{\mathbb{R}}(1 \wedge |x|)\,\nu(dx) < \infty$, and let X be an infinitely divisible random variable whose characteristic function has the Lévy-Khintchin representation $\mathbf{E}[e^{iux}] = \exp\{\int (e^{iuz} - 1)\nu(dz)\}$. Then $X \sim X_1$, where $(X_t)_{t\in\mathbb{R}_+}$ is the Lévy process with Lévy measure ν. The latter is Markov with generator

$$\mathscr{L}f(x) = \int D_y f(x)\,\nu(dy), \qquad D_y f(x) := f(x+y) - f(x).$$

Use the above machinery to prove the following Poincaré inequality:

$$\mathrm{Var}[f(X)] \le \mathbf{E}\left[\int (D_y f(X))^2\,\nu(dy)\right].$$

In particular, deduce Poincaré inequalities for the Poisson distribution and for the one-sided exponential distribution (the latter being distinct from both Poincaré inequalities in Problem 2.9 above).

2.13 (Local Poincaré inequalities II). The approach of Problem 2.12 makes it possible to obtain Poincaré inequalities using Markov processes that do not admit a stationary measure. However, even for ergodic Markov processes, it can be useful to develop a Poincaré inequality for the stationary measure μ by letting $t \to \infty$ in a local Poincaré inequality. The reason for this is the following result that will be proved in this problem.

Theorem 2.37 (Local Poincaré inequality). *The following are equivalent:*

1. $c\Gamma_2(f,f) \geq \Gamma(f,f)$ *for all* f (Bakry-Émery criterion).
2. $\Gamma(P_t f, P_t f) \leq e^{-2t/c} P_t \Gamma(f,f)$ *for all* f, t (local ergodicity).
3. $P_t(f^2) - (P_t f)^2 \leq c(1 - e^{-2t/c}) P_t \Gamma(f,f)$ *for all* f, t (local Poincaré).

Here we defined

$$\Gamma_2(f,g) := \frac{1}{2}\{\mathscr{L}\Gamma(f,g) - \Gamma(f,\mathscr{L}g) - \Gamma(\mathscr{L}f,g)\}.$$

This is called the *iterated carré du champ* or Γ_2*-operator*.

Why is this result useful? Suppose that P_t is an ergodic Markov semigroup with stationary measure μ. To prove a Poincaré inequality using Theorem 2.18, we had to be able to prove exponential ergodicity of the semigroup. This is typically a nontrivial matter: one cannot readily read off exponential ergodicity from the expression for the generator $\mathscr{L}$, for example. In contrast, the first property of Theorem 2.37 is an *algebraic identity*

$$c\Gamma_2(f,f) \geq \Gamma(f,f)$$

that can be verified readily from the expression for $\mathscr{L}$. On the other hand, if this identity is valid, letting $t \to \infty$ in property 3 of Theorem 2.37 yields

$$\mathrm{Var}_\mu[f] \leq c\mathcal{E}(f,f)$$

(cf. Problem 2.7). Thus the local approach provides us with an algebraic criterion for the validity of a Poincaré inequality. This can be extremely useful, as we will see below. However, the Bakry-Émery criterion is strictly stronger than the validity of a Poincaré inequality for the stationary measure μ.

Let us begin by proving the various implications of Theorem 2.37

a. Prove 2 $\Rightarrow$ 3. Hint: this follows easily as in Problem 2.12.

b. Prove 1 $\Rightarrow$ 2. Hint: $\frac{d}{ds} P_{t-s}\Gamma(P_s f, P_s f)$.

c. Prove 3 $\Rightarrow$ 1. Hint: $\lim_{t\downarrow 0} t^{-2}\{P_t(f^2) - (P_t f)^2 - c(1 - e^{-2t/c}) P_t\Gamma(f,f)\}$.

We now demonstrate the power of Theorem 2.37 in an important example.

d. Let μ be a probability measure on $\mathbb{R}^n$ with density $\mu(dx) = e^{-W(x)}\,dx$ where W is a smooth convex function. Such distributions are called *log-concave*. Note that is $X \sim \mu$, then $X_1, \ldots, X_n$ are not independent. Nonetheless, we have the following result: if W is ρ-uniformly convex, that is,

$$\sum_{i,j=1}^n v_i v_j \frac{\partial^2 W(x)}{\partial x_i \partial x_j} \geq \rho \sum_{i=1}^n v_i^2 \quad \text{for all } v \in \mathbb{R}^n,$$

then we have the dimension-free Poincaré inequality

$$\mathrm{Var}_\mu[f] \le \frac{1}{\rho} \int \|\nabla f\|^2 \, d\mu.$$

To prove it, we note that μ is the stationary measure of the Langevin stochastic differential equation (B is n-dimensional Brownian motion)

$$dX_t = -\nabla W(X_t)\, dt + \sqrt{2}\, dB_t,$$

which is a Markov process with generator

$$\mathscr{L} f(x) = -\sum_{i=1}^n \frac{\partial W(x)}{\partial x_i} \frac{\partial f(x)}{\partial x_i} + \sum_{i=1}^n \frac{\partial^2 f(x)}{\partial x_i^2}.$$

Prove the log-concave Poincaré inequality using the Bakry-Émery criterion.

Remark. We have shown that ρ-uniformly log-concave measures admit a dimension-free Poincaré inequality with constant ρ^{-1}. This says nothing about general case where ρ may be zero. One of the deepest open problems in the theory of Poincaré inequalities is to understand the situation for general log-concave measures. It has been conjectured by Kannan, Lovász and Simonovits that if μ is a log-concave measure on $\mathbb{R}^n$ with zero mean and identity covariance matrix, then $\mathrm{Var}_\mu[f] \le C \int \|\nabla f\|^2 \, d\mu$ for a *universal* constant C (independent of the dimension!) To date, there is little progress in this direction. However, several interesting ideas have been developed for the investigation of such problems, including a localization method that provides a partial replacement for tensorization in the setting of log-concave distributions.

Notes

§2.1. The tensorization property of the variance is classical. It is sometimes called the Efron-Stein inequality after [33], where it was used to investigate Tukey's jackknife estimator. The importance of tensorization as a fundamental principle was emphasized by Ledoux [48]. The random matrix example was taken from [13]. Problems 2.4 and 2.5 are from [14] and [48], respectively. Much of what is known on superconcentration can be found in [18].

§2.2. The text [52] gives an introduction to Markov processes in continuous time. A comprehensive treatment of Markov semigroups and their connections with functional inequalities is given in [7].

§2.3 and §2.4. The treatment of Poincaré inequalities given here follows [7], as do many of the problems. Problem 2.9 is inspired by [10], and Problem 2.10 is taken from [99]. The application of local Poincaré inequalities to infinitely divisible distributions in Problem 2.12 is inspired by [17]. See [16] for more on the conjecture of Kannan, Lovász and Simonovits for log-concave measures.

3

Subgaussian concentration and log-Sobolev inequalities

In Chapter 2 we investigated the simplest form of the concentration phenomenon: the variance of a function $f(X_1, \ldots, X_n)$ of independent (or weakly dependent) random variables is small if the "gradient" of f small. This is indeed an embodiment of the concentration phenomenon as it was informally presented in Chapter 1: the variance measures the size of the fluctuations of the random variable $f(X_1, \ldots, X_n)$, while the gradient measures the sensitivity of $f(x)$ to its coordinates x_i. While variance bounds can be extremely useful and are of interest in their own right, it is often important in applications to have sharper control on the distribution of the fluctuations.

What type of refined behavior can we expect? Let us recall our original motivating example where $f(X_1, \ldots, X_n) = \frac{1}{n}\sum_{k=1}^n X_k$ is a linear function. By the weak law of large numbers, we expect that the fluctuations are of order

$$f(X_1, \ldots, X_n) - \mathbf{E}f(X_1, \ldots, X_n) \sim \sigma/\sqrt{n},$$

which is indeed captured correctly by the variance bounds developed in the previous chapter. In this case, however, the central limit theorem provides us with much sharper information: it controls not only the *size* of the fluctuations, but also the *distribution* of the fluctuations

$$f(X_1, \ldots, X_n) - \mathbf{E}f(X_1, \ldots, X_n) \approx N(0, \sigma^2/n).$$

In particular, we might expect that

$$\mathbf{P}[|f(X_1, \ldots, X_n) - \mathbf{E}f(X_1, \ldots, X_n)| \geq t] \lesssim e^{-nt^2/2\sigma^2},$$

as would be true if the fluctuations were in fact Gaussian (we will show this below). Such a Gaussian tail inequality provides much more precise control of the fluctuations than a bound on the variance. This will be important, for example, in understanding the behavior of suprema later on in the course.

As in the previous chapter, it turns out that the above idea is not restricted to linear functions, but is in fact a manifestation of a general phenomenon: it is

often possible to obtain Gaussian tail bounds on the fluctuations of nonlinear functions f provided that their "gradient" is small in a suitable sense. In this chapter, we begin the investigation of such *concentration inequalities*.

3.1 Subgaussian variables and Chernoff bounds

Before we can prove any concentration inequalities, we must first consider how one might go about proving that a random variable satisfies a Gaussian tail bound. Most tail bounds in probability theory are proved using some form of Markov's inequality. For example, if we have a bound on the variance as in the previous chapter, we immediately obtain a tail bound of the form

$$\mathbf{P}[|X - \mathbf{E}[X]| \geq t] \leq \frac{\mathrm{Var}[X]}{t^2}.$$

However, this bound only decays as t^{-2}, and we cannot obtain Gaussian tail bounds from Poincaré inequalities in this manner. To obtain Gaussian tail bounds, we must use Markov's inequality in a more sophisticated manner. The basic method is known as the *Chernoff bound*.

Lemma 3.1 (Chernoff bound). *Define the log-moment generating function ψ of a random variable X and its Legendre dual ψ^* as*

$$\psi(\lambda) := \log \mathbf{E}[e^{\lambda(X - \mathbf{E}X)}], \qquad \psi^*(t) = \sup_{\lambda \geq 0}\{\lambda t - \psi(\lambda)\}.$$

Then $\mathbf{P}[X - \mathbf{E}X \geq t] \leq e^{-\psi^*(t)}$ for all $t \geq 0$.

Proof. The idea is strikingly simple: we simply exponentiate inside the probability before applying Markov's inequality. For any $\lambda \geq 0$, we have

$$\mathbf{P}[X - \mathbf{E}X \geq t] = \mathbf{P}[e^{\lambda(X - \mathbf{E}X)} \geq e^{\lambda t}] \leq e^{-\lambda t}\mathbf{E}[e^{\lambda(X - \mathbf{E}X)}] = e^{-\{\lambda t - \psi(\lambda)\}}$$

using Markov's inequality and that $x \mapsto e^{\lambda x}$ is increasing. As the left-hand side does not depend on the choice of $\lambda \geq 0$, we can optimize the right-hand side over λ to obtain the statement of the lemma. □

Remark 3.2. Note that the Chernoff bound only gives the *upper tail*, that is, the probability $\mathbf{P}[X \geq \mathbf{E}X + t]$ that the random variable X exceeds its mean $\mathbf{E}X$ by a fixed amount. However, we can obtain an inequality for the *lower tail* by applying the Chernoff bound to the random variable $-X$, as

$$\mathbf{P}[X \leq \mathbf{E}X - t] = \mathbf{P}[-X \geq \mathbf{E}[-X] + t].$$

In particular, given an upper and lower tail bound, we can obtain a bound on the magnitude of the fluctuations using the union bound

$$\begin{aligned}\mathbf{P}[|X - \mathbf{E}X| \geq t] &= \mathbf{P}[X \geq \mathbf{E}X + t \text{ or } X \leq \mathbf{E}X - t] \\ &\leq \mathbf{P}[X \geq \mathbf{E}X + t] + \mathbf{P}[-X \geq \mathbf{E}[-X] + t].\end{aligned}$$

In many cases, proving an upper tail bound will immediately imply a lower tail bound and a two-sided bound in this manner. On the other hand, sometimes upper or lower tail bounds will be proved under assumptions that are not invariant under negation. For example, if we prove an upper tail bound for convex functions $f(X)$, this does not automatically imply a lower tail bound as $-f(X)$ is concave and not convex; in such cases, a lower tail bound must be proved separately. One should therefore be careful when interpreting tail bounds to check separately the validity of upper and lower tail bounds.

Remark 3.3. The utility of the Chernoff bound is by no means restricted to proving Gaussian tails as we will do below. One can obtain many different tail behaviors in this manner. However, the method clearly only works if $\psi(\lambda)$ is finite at least for λ in a neighborhood of 0. Therefore, to apply the Chernoff bound, the random variable X should have at least exponential tails. For random variables with heavier tails an alternative method is needed, for example, one could take powers rather than exponentials in Markov's inequality:

$$\mathbf{P}[X - \mathbf{E}X \geq t] \leq \inf_{p \in \mathbb{N}} \frac{\mathbf{E}[(X - \mathbf{E}X)_+^p]}{t^p}.$$

In fact, even when the Chernoff bound is applicable, it is not difficult to show that this moment bound is at least as good as the Chernoff bound.

Why are Chernoff bounds so useful? There are some simple examples, such as the case of sums of random variables, where the Chernoff bound proves to be easy to manipulate (we will exploit this in the next section). However, the real power of the Chernoff bound is that the log-moment generating function $\lambda \mapsto \psi(\lambda)$ is a *continuous* object, and can therefore be investigated using calculus. We will repeatedly exploit this approach in the sequel. In contrast, the moment bound given above is discrete in nature, and is therefore much more difficult to handle. As we will mostly be interested in Gaussian tail bounds, we will make full use of the convenience of the Chernoff method.

To show how the Chernoff bound can give rise to Gaussian tail bounds, let us first consider the case of an actual Gaussian random variable.

Example 3.4. Let $X \sim N(\mu, \sigma^2)$. Then $\mathbf{E}[e^{\lambda(X - \mathbf{E}X)}] = e^{\lambda^2\sigma^2/2}$, so

$$\psi(\lambda) = \frac{\lambda^2\sigma^2}{2}, \qquad \psi^*(t) = \frac{t^2}{2\sigma^2}.$$

In particular, we have $\mathbf{P}[X - \mathbf{E}X \geq t] \leq e^{-t^2/2\sigma^2}$.

Observe that in order to get the tail bound in Example 3.4, the fact that X is Gaussian was not actually important: it would suffice to assume that the

log-moment generating function is bounded from above by that of a Gaussian $\psi(\lambda) \le \lambda^2\sigma^2/2$. Random variables that satisfy this condition play a central role in the investigation of Gaussian tail bounds.

Definition 3.5 (Subgaussian random variables). *A random variable is called* σ^2-subgaussian *if its log-moment generating function satisfies* $\psi(\lambda) \le \lambda^2\sigma^2/2$ *for all* $\lambda \in \mathbb{R}$ *(and the constant* σ^2 *is called the* variance proxy*).*

Note that if $\psi(\lambda)$ is the log-moment generating function of a random variable X, then $\psi(-\lambda)$ is the log-moment generating function of the random variable $-X$. For a σ^2-subgaussian random variable X, we can therefore apply the Chernoff bound to both the upper and lower tails to obtain

$$\mathbf{P}[X \ge \mathbf{E}X + t] \le e^{-t^2/2\sigma^2}, \qquad \mathbf{P}[X \le \mathbf{E}X - t] \le e^{-t^2/2\sigma^2}.$$

As moment generating functions will prove to be much easier to manipulate than the tail probabilities themselves, we will almost always study Gaussian tail behavior of random variables in terms of the subgaussian property. Fortunately, it turns out that little is lost in making this simplification: any random variable that satisfies Gaussian tail bounds must necessarily be subgaussian (albeit for a slightly larger variance proxy), cf. Problem 3.1 below.

So far, the only examples of subgaussian random variables that we have encountered are Gaussians, which is not terribly interesting. One of the most basic results on subgaussian random variables is that every *bounded* random variable is subgaussian. This statement is made precise by Hoeffding's lemma, which could be viewed as a far-reaching generalization of the trivial Lemma 2.1. Even in this simple setting, the proof provides a nontrivial illustration of the important role of calculus in bounding moment generating functions.

Lemma 3.6 (Hoeffding lemma). *Let* $a \le X \le b$ *a.s. for some* $a, b \in \mathbb{R}$. *Then* $\mathbf{E}[e^{\lambda(X-\mathbf{E}X)}] \le e^{\lambda^2(b-a)^2/8}$*, i.e.,* X *is* $(b-a)^2/4$*-subgaussian.*

Proof. We can assume without loss of generality that $\mathbf{E}X = 0$. In this case we have $\psi(\lambda) = \log \mathbf{E}[e^{\lambda X}]$, and we can readily compute

$$\psi'(\lambda) = \frac{\mathbf{E}[Xe^{\lambda X}]}{\mathbf{E}[e^{\lambda X}]}, \qquad \psi''(\lambda) = \frac{\mathbf{E}[X^2e^{\lambda X}]}{\mathbf{E}[e^{\lambda X}]} - \left[\frac{\mathbf{E}[Xe^{\lambda X}]}{\mathbf{E}[e^{\lambda X}]}\right]^2.$$

Thus $\psi''(\lambda)$ can be interpreted as the variance of the random variable X under the twisted probability measure $d\mathbf{Q} = \frac{e^{\lambda X}}{\mathbf{E}[e^{\lambda X}]}d\mathbf{P}$. But then Lemma 2.1 yields $\psi''(\lambda) \le (b-a)^2/4$, and the fundamental theorem of calculus yields

$$\psi(\lambda) = \int_0^\lambda \int_0^\mu \psi''(\rho)\, d\rho\, d\mu \le \frac{\lambda^2(b-a)^2}{8}$$

using $\psi(0) = \log 1 = 0$ and $\psi'(0) = \mathbf{E}X = 0$. □

Problems

3.1 (Subgaussian variables). There are several different notions of random variables with a Gaussian tail that are all essentially equivalent up to constants. The aim of this problem is to obtain some insight into these notions.

a. As a warmup exercise, show that if X is σ^2-subgaussian, then $\mathrm{Var}[X] \le \sigma^2$.

b. Show that for any increasing and differentiable function Φ

$$\mathbf{E}[\Phi(|X|)] = \Phi(0) + \int_0^\infty \Phi'(t)\,\mathbf{P}[|X| \ge t]\,dt.$$

This elementary identity will be needed below.

In the following, we will assume for simplicity that $\mathbf{E}X = 0$. We now prove that the following three properties are equivalent for suitable constants σ, b, c: (1) X is σ^2-subgaussian; (2) $\mathbf{P}[|X| \ge t] \le 2e^{-bt^2}$; and (3) $\mathbf{E}[e^{cX^2}] \le 2$.

c. Show that if X is σ^2-subgaussian, then $\mathbf{P}[|X| \ge t] \le 2e^{-t^2/2\sigma^2}$.

d. Show that if $\mathbf{P}[|X| \ge t] \le 2e^{-t^2/2\sigma^2}$, then $\mathbf{E}[e^{X^2/6\sigma^2}] \le 2$.
Hint: use part a.

e. Show that if $\mathbf{E}[e^{X^2/6\sigma^2}] \le 2$, then X is $18\sigma^2$-subgaussian.
Hint: use $\mathbf{E}[e^{\lambda X}] \le 1 + \frac{\lambda^2}{2}\mathbf{E}[X^2 e^{|\lambda X|}]$ by Taylor's theorem together with Young's inequality $|\lambda X| \le \frac{a\lambda^2}{2} + \frac{X^2}{2a}$ for a suitable choice of a.

In addition, the subgaussian property of X is equivalent to the fact that the moments of X scale as is the case for the Gaussian distribution.

f. Show that if X is σ^2-subgaussian, then $\mathbf{E}[X^{2q}] \le (4\sigma^2)^q q!$ for all $q \in \mathbb{N}$.
Hint: use part a.

g. Show that if $\mathbf{E}[X^{2q}] \le (4\sigma^2)^q q!$ for all $q \in \mathbb{N}$, then $\mathbf{E}[e^{X^2/8\sigma^2}] \le 2$.
Hint: expand in a power series.

3.2 (Tightness of Hoeffding's lemma). Show that the bound of Hoeffding's lemma is the best possible by considering $\mathbf{P}[X = a] = \mathbf{P}[X = b] = \frac{1}{2}$.

3.3 (Chernoff bound vs. moments). Show that for $t \ge 0$

$$\mathbf{P}[X - \mathbf{E}X \ge t] \le \inf_{p \ge 0} \frac{\mathbf{E}[(X - \mathbf{E}X)_+^p]}{t^p} \le \inf_{\lambda \ge 0} e^{-\lambda t}\mathbf{E}[e^{\lambda(X - \mathbf{E}X)}].$$

Thus the moment bound of Remark 3.3 is at least as good as the Chernoff bound. However, the former is much harder to use than the latter.
Hint: use $\mathbf{E}[e^{\lambda(X-\mathbf{E}X)}] \ge \mathbf{E}[\mathbf{1}_{X-\mathbf{E}X>0} e^{\lambda(X-\mathbf{E}X)}]$ and expand in a power series.

3.4 (Chernoff bound exercises). Compute the explicit form of the Chernoff bound for Poisson and Bernoulli random variables.

3.5 (Maxima of subgaussian variables). Let $X_1, X_2, \ldots$ be (not necessarily independent) σ^2-subgaussian random variables. Show that

$$\mathbf{P}\left[\max_{i\le n}\{X_i - \mathbf{E}X_i\} \ge (1+\varepsilon)\sigma\sqrt{2\log n}\right] \xrightarrow{n\to\infty} 0 \quad \text{for all } \varepsilon > 0.$$

Hint: use the union bound

$$\mathbf{P}[X \vee Y \ge t] = \mathbf{P}[X \ge t \text{ or } Y \ge t] \le \mathbf{P}[X \ge t] + \mathbf{P}[Y \ge t].$$

This problem shows that the maximum $\max_{i\le n}\{X_i - \mathbf{E}X_i\}$ of σ^2-subgaussian random variables is at most of order $\sigma\sqrt{2\log n}$. This is the simplest example of the crucial role played by tail bounds in estimating the size of maxima of random variables. The second part of this course will be entirely devoted to the investigation of such problems (using much deeper ideas).

3.2 The martingale method

Let $X_1, \ldots, X_n$ be independent random variables. In the previous chapter, we showed that the variance of $f(X_1, \ldots, X_n)$ can be bounded in many cases by a "square gradient" of the function f. The aim of this chapter is to obtain a much stronger type of result: we would like to show that $f(X_1, \ldots, X_n)$ is subgaussian with variance proxy controlled by a "square gradient" of f.

A key idea developed in the previous chapter was to use tensorization to reduce the problem to the one-dimensional case. With the tensorization inequality in hand, we could even apply a trivial bound such as Lemma 2.1 to obtain a nontrivial variance inequality in terms of bounded differences. Our first instinct in the present setting is therefore to prove a tensorization inequality for the subgaussian property, which could then be combined with Hoeffding's Lemma 3.6 (which plays the analogous role in the present setting to the trivial Lemma 2.1 for the variance) in order to obtain a concentration inequality in terms of bounded differences. Unfortunately, it turns out that unlike in the case of the variance, the subgaussian property *does not tensorize* in a natural manner, and thus we cannot directly implement this program. One of the most important ideas that will be developed in the following sections is that the proof of subgaussian inequalities can be reduced to a strenghtened form of Poincaré inequalities, called *log-Sobolev inequalities*, that *do* tensorize exactly in the same manner as the variance. This will provide us with a very powerful tool to prove subgaussian concentration.

There is, however, a more elementary approach that should be attempted before we begin introducing new ideas. Even though the subgaussian property does not tensorize in the same manner as the variance, we can still repeat some

of the steps in the proof of the tensorization Theorem 2.3 in the subgaussian setting. Recall that the main idea of the proof of Theorem 2.3 is to write

$$f(X_1,\dots,X_n) - \mathbf{E}f(X_1,\dots,X_n) = \sum_{k=1}^n \Delta_k,$$

where

$$\Delta_k = \mathbf{E}[f(X_1,\dots,X_n)|X_1,\dots,X_k] - \mathbf{E}[f(X_1,\dots,X_n)|X_1,\dots,X_{k-1}]$$

are martingale differences. The following simple result, which exploits the nice behavior of the exponential of a sum, could be viewed as a sort of poor man's tensorization property for sums of martingale increments. By working directly with the martingale increments, we will be able to derive a first concentration inequality. This approach is commonly known as the *martingale method*.

Lemma 3.7 (Azuma). *Let $\{\mathcal{F}_k\}_{k\le n}$ be any filtration, and let $\Delta_1,\dots,\Delta_n$ be random variables that satisfy the following properties for $k = 1,\dots,n$:*

1. *Martingale difference property: Δ_k is $\mathcal{F}_k$-measurable and $\mathbf{E}[\Delta_k|\mathcal{F}_{k-1}] = 0$.*
2. *Conditional subgaussian property: $\mathbf{E}[e^{\lambda\Delta_k}|\mathcal{F}_{k-1}] \le e^{\lambda^2\sigma_k^2/2}$ a.s.*

Then the sum $\sum_{k=1}^n \Delta_k$ is subgaussian with variance proxy $\sum_{k=1}^n \sigma_k^2$.

Proof. For any $1 \le k \le n$, we can compute

$$\mathbf{E}[e^{\lambda\sum_{i=1}^k \Delta_i}] = \mathbf{E}[e^{\lambda\sum_{i=1}^{k-1}\Delta_i}\,\mathbf{E}[e^{\lambda\Delta_k}|\mathcal{F}_{k-1}]] \le e^{\lambda^2\sigma_k^2/2}\mathbf{E}[e^{\lambda\sum_{i=1}^{k-1}\Delta_i}].$$

It follows by induction that $\mathbf{E}[e^{\lambda\sum_{i=1}^n\Delta_i}] \le e^{\lambda^2\sum_{i=1}^n\sigma_i^2/2}$. □

Remark 3.8. While we did not explicitly use the martingale difference property in the proof, $\mathbf{E}[e^{\lambda\Delta_k}|\mathcal{F}_{k-1}] \le e^{\lambda^2\sigma_k^2/2}$ can in fact only hold if $\mathbf{E}[\Delta_k|\mathcal{F}_{k-1}] = 0$ (consider $(\mathbf{E}[e^{\lambda\Delta_k}|\mathcal{F}_{k-1}]-1)/\lambda$ as $\lambda \downarrow 0$). In general, the conditional subgaussian property of X given $\mathcal{F}$ should read $\mathbf{E}[e^{\lambda(X-\mathbf{E}[X|\mathcal{F}])}|\mathcal{F}] \le e^{\lambda^2\sigma^2/2}$ a.s.

In combination with Hoeffding's Lemma 3.6, we now obtain a classical result on the tail behavior of sums of martingale differences.

Corollary 3.9 (Azuma-Hoeffding inequality). *Let $\{\mathcal{F}_k\}_{k\le n}$ be any filtration, and let Δ_k, A_k, B_k satisfy the following properties for $k = 1,\dots,n$:*

1. *Martingale difference property: Δ_k is $\mathcal{F}_k$-measurable and $\mathbf{E}[\Delta_k|\mathcal{F}_{k-1}] = 0$.*
2. *Predictable bounds: A_k, B_k are $\mathcal{F}_{k-1}$-measurable and $A_k \le \Delta_k \le B_k$ a.s.*

Then $\sum_{k=1}^n \Delta_k$ is subgaussian with variance proxy $\frac{1}{4}\sum_{k=1}^n \|B_k - A_k\|_\infty^2$. In particular, we obtain for every $t \ge 0$ the tail bound

$$\mathbf{P}\left[\sum_{k=1}^n \Delta_k \ge t\right] \le \exp\left(-\frac{2t^2}{\sum_{k=1}^n \|B_k - A_k\|_\infty^2}\right).$$

Proof. Applying Hoeffding's Lemma 3.6 to Δ_k conditionally on $\mathcal{F}_{k-1}$ implies $\mathbf{E}[e^{\lambda\Delta_k}|\mathcal{F}_{k-1}] \le e^{\lambda^2(B_k-A_k)^2/8}$. The result now follows from Lemma 3.7. □

Example 3.10. The Azuma-Hoeffding inequality is often applied in the following setting. Let $X_1, \ldots, X_n$ be independent random variables such that $a \le X_i \le b$ for all i. Applying Corollary 3.9 with $\Delta_k = (X_k - \mathbf{E}X_k)/n$ yields

$$\mathbf{P}\left[\frac{1}{n}\sum_{k=1}^{n}\{X_i - \mathbf{E}X_i\} \ge t\right] \le e^{-2nt^2/(b-a)^2}.$$

By the central limit theorem, this bound is of the correct order both in terms of the size of the sum and its Gaussian tail behavior. However, just as for the case of the variance (see the discussion in section 2.1), this bound can be pessimistic in that it does not capture any information on the distribution of the variables X_i: in particular, the variance proxy $(b-a)^2/4$ may be much larger than the actual variance of the random variables X_i. Much of the effort in developing concentration inequalities is to obtain bounds in terms of "good" variance proxies for the purposes of the application at hand.

We motivated the development of tail bounds for martingale differences as a partial replacement of the tensorization inequality for the variance. Let us therefore return to the case of functions $f(X_1, \ldots, X_n)$ of independent random variables $X_1, \ldots, X_n$. Using the Azuma-Hoeffding inequality, we readily obtain our first and simplest subgaussian concentration inequality. Recall that

$$D_i f(x) := \sup_z f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n) - \inf_z f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n)$$

are the discrete derivatives defined in section 2.1.

Theorem 3.11 (McDiarmid). *For $X_1, \ldots, X_n$ independent, $f(X_1, \ldots, X_n)$ is subgaussian with variance proxy $\frac{1}{4}\sum_{k=1}^{n}\|D_k f\|_\infty^2$. In particular,*

$$\mathbf{P}[f(X_1, \ldots, X_n) - \mathbf{E}f(X_1, \ldots, X_n) \ge t] \le e^{-2t^2/\sum_{k=1}^{n}\|D_k f\|_\infty^2}.$$

Proof. As in the proof of the tensorization Theorem 2.3, we write

$$f(X_1, \ldots, X_n) - \mathbf{E}f(X_1, \ldots, X_n) = \sum_{k=1}^{n}\Delta_k,$$

where

$$\Delta_k = \mathbf{E}[f(X_1, \ldots, X_n)|X_1, \ldots, X_k] - \mathbf{E}[f(X_1, \ldots, X_n)|X_1, \ldots, X_{k-1}]$$

are martingale differences. Note that $A_k \le \Delta_k \le B_k$ with

$$A_k = \mathbf{E}[\inf_z f(X_1, \ldots, X_{k-1}, z, X_{k+1}, \ldots, X_n) - f(X_1, \ldots, X_n)|X_1, \ldots, X_{k-1}],$$

$$B_k = \mathbf{E}[\sup_z f(X_1, \ldots, X_{k-1}, z, X_{k+1}, \ldots, X_n) - f(X_1, \ldots, X_n)|X_1, \ldots, X_{k-1}]$$

where we have used the independence of X_k and $X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n$. The result now follows immediately from the Azuma-Hoeffding inequality of Corollary 3.9 once we note that $|B_k - A_k| \le \|D_k f\|_\infty$. □

McDiarmid's inequality should be viewed as a subgaussian form of the bounded difference inequality of Corollary 2.4. In Corollary 2.4, the variance is controlled by the expectation of the "square gradient" of the function f. In contrast, McDiarmid's inequality yields the stronger subgaussian property, but here the variance proxy is controlled by a uniform upper bound on the "square gradient" rather than its expectation. Of course, it makes sense that a stronger property requires a stronger assumption. We will repeatedly encounter this idea in the setting of concentration inequalities: typically the *expectation* of the "square gradient" controls the variance, while a *uniform* bound on the "square gradient" controls the subgaussian variance proxy.

However, from this viewpoint, the result of Theorem 3.11 is not satisfactory: as the appropriate notion of "square gradient" in the bounded difference inequality is $\sum_{k=1}^n |D_k f|^2$, we would expect a variance proxy of order $\|\sum_{k=1}^n |D_k f|^2\|_\infty$; however, Theorem 3.11 only yields control in terms of the larger quantity $\sum_{k=1}^n \|D_k f\|_\infty^2$. The former would constitute a crucial improvement over the latter in many situations (for example, in the setting of the random matrix Example 2.5). Unfortunately, the martingale method is far too crude to capture this idea. In the sequel, we will develop new techniques for proving subgaussian concentration inequalities that will make it possible to prove much more refined bounds in many settings.

Problems

3.6 (Bin packing). For the bin packing Problem 2.3, show that the variance bound $\mathrm{Var}[B_n] \le n/4$ can be strengthened to a Gaussian tail bound

$$\mathbf{P}[|B_n - \mathbf{E}B_n| \ge t] \le 2e^{-2t^2/n}.$$

In view of Problem 2.3, this bound has the correct order.

3.7 (Rademacher processes). Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent symmetric Bernoulli random variables $\mathbf{P}[\varepsilon_i = \pm 1] = \frac{1}{2}$, and let $T \subseteq \mathbb{R}^n$. Define

$$Z = \sup_{t \in T} \sum_{k=1}^n \varepsilon_k t_k.$$

In Problem 2.2, we showed that

$$\mathrm{Var}[Z] \le 4 \sup_{t \in T} \sum_{k=1}^{n} t_k^2.$$

Show that McDiarmid's inequality can give, at best, a bound of the form

$$\mathbf{P}[|Z - \mathbf{E}Z| \ge t] \le 2e^{-t^2/2\sigma^2} \quad \text{with} \quad \sigma^2 = \sum_{k=1}^{n} \sup_{t \in T} t_k^2.$$

Show by means of an example that the variance proxy in McDiarmid's inequality can exhibit a vastly incorrect scaling as a function of dimension n.

3.8 (Empirical frequencies). Let $X_1, \ldots, X_n$ be i.i.d. random variables with any distribution μ on a measurable space E, and let $\mathcal{C}$ be a countable class of measurable subsets of E. By the law of large numbers,

$$\frac{\#\{k \in \{1, \ldots, n\} : X_k \in C\}}{n} \approx \mu(C)$$

when n is large. In order to analyze empirical risk minimization methods in machine learning, it is important to control the deviation between the true probability $\mu(C)$ and its empirical average *uniformly* over the class $\mathcal{C}$. In particular, one would like to guarantee that the uniform deviation

$$Z_n = \sup_{C \in \mathcal{C}} \left| \frac{\#\{k \in \{1, \ldots, n\} : X_k \in C\}}{n} - \mu(C) \right|$$

does not exceeed a certain level with high probability. As a starting point towards proving such a result, show that for every $n \ge 1$ and $t \ge 0$

$$\mathbf{P}[Z_n \ge \mathbf{E}Z_n + t] \le e^{-2nt^2}.$$

To obtain a bound on $\mathbf{P}[Z_n \ge t]$, it therefore remains to control $\mathbf{E}Z_n$ (the techniques for this will be developed in the second part of the course).

3.9 (Sums in Hilbert space). Let $X_1, \ldots, X_n$ be independent random variables with zero mean in a Hilbert space, and suppose that $\|X_k\| \le C$ a.s. for every k. Let us prove a sort of Hilbert-valued analogue of Example 3.10.

a. Show that for all $t \ge 0$

$$\mathbf{P}\left[\left\| \frac{1}{n} \sum_{k=1}^{n} X_k \right\| \ge \mathbf{E} \left\| \frac{1}{n} \sum_{k=1}^{n} X_k \right\| + t \right] \le e^{-nt^2/2C^2}.$$

b. Show that

$$\mathbf{E} \left\| \frac{1}{n} \sum_{k=1}^{n} X_k \right\| \le C n^{-1/2}.$$

c. Conclude that for all $t \geq Cn^{-1/2}$

$$\mathbf{P}\left[\left\|\frac{1}{n}\sum_{k=1}^{n} X_k\right\| \geq t\right] \leq e^{-nt^2/8C^2}.$$

d. Finally, argue that for all $t \geq 0$

$$\mathbf{P}\left[\left\|\frac{1}{n}\sum_{k=1}^{n} X_k\right\| \geq t\right] \leq 2e^{-nt^2/8C^2}.$$

3.10 (Random graphs). An Erdős-Rényi random graph $G(n,p)$ is a graph on n vertices such that for every pair of vertices v, v' there is an edge between them with probability p, independently of the other edges. A coloring of the graph is the assignment of a color to each vertex such that every pair of vertices connected by an edge have a distinct color. The *chromatic number* χ is the minimal number of colors needed to color the graph. Show that

$$\mathbf{P}[|\chi - \mathbf{E}\chi| \geq t\sqrt{n}] \leq 2e^{-t^2}.$$

It can be shown that the chromatic number satisfies $\mathbf{E}\chi \sim n/2\log_b n$ as $n \to \infty$, where $b = 1/(1-p)$. We therefore see that the fluctuations of the chromatic number are of much smaller order than its magnitude.

3.11 (A generalization of Azuma-Hoeffding). Consider the same setting as in Corollary 3.9. The Azuma-Hoeffding inequality provides a Gaussian tail bound in the case that $|B_k - A_k|$ is uniformly bounded, but this may not always hold in practice. Prove the following general form of the Azuma-Hoeffding inequality that does not require boundedness of the increments:

$$\mathbf{P}\left[\sum_{k=1}^{n} \Delta_k \geq t \text{ and } \sum_{k=1}^{n}(B_k - A_k)^2 \leq c^2\right] \leq e^{-2t^2/c^2}.$$

Hint: consider $\lambda\sum_{k=1}^{n}\Delta_k - \frac{\lambda^2}{8}\sum_{k=1}^{n}(B_k - A_k)^2$.

3.3 The entropy method

The martingale method developed in the previous section has many useful applications. Nonetheless, as was explained above, the inequalities derived from this approach are often unsatisfactory in high dimension. In essence, the fundamental problem is that the subgaussian property does not tensorize naturally, and the martingale method can only partially address this issue. In order to obtain sharper results, we must confront the tensorization problem directly. In this section, we will introduce a powerful method to do just that. The key idea is to introduce an alternative formulation of the subgaussian property that behaves naturally under tensorization.

Recall that a random variable X is subgaussian if its log-moment generating function satisfies $\psi(\lambda) := \log \mathbf{E}[e^{\lambda(X-\mathbf{E}X)}] \lesssim \lambda^2$. We have already seen the importance of using calculus to bound moment generating functions in the proof of Hoeffding's Lemma 3.6: the idea used there is that if $\frac{d^2}{d\lambda^2}\psi(\lambda) \lesssim 1$, then the subgaussian property is obtained by integrating twice. The idea behind the following result is very similar: as the subgaussian property is equivalent to $\lambda^{-1}\psi(\lambda) \lesssim \lambda$, it suffices to show that $\frac{d}{d\lambda}\lambda^{-1}\psi(\lambda) \lesssim 1$.

Definition 3.12. *The* entropy *of a nonnegative random variable Z is*

$$\mathrm{Ent}[Z] := \mathbf{E}[Z \log Z] - \mathbf{E}[Z] \log \mathbf{E}[Z].$$

Lemma 3.13 (Herbst). *Suppose that*

$$\mathrm{Ent}[e^{\lambda X}] \le \frac{\lambda^2\sigma^2}{2}\,\mathbf{E}[e^{\lambda X}] \quad \textit{for all } \lambda \ge 0.$$

Then

$$\psi(\lambda) := \log \mathbf{E}[e^{\lambda(X-\mathbf{E}X)}] \le \frac{\lambda^2\sigma^2}{2} \quad \textit{for all } \lambda \ge 0.$$

Proof. As $\psi(\lambda) = \log \mathbf{E}[e^{\lambda X}] - \lambda \mathbf{E}X$, we have

$$\frac{d}{d\lambda}\frac{\psi(\lambda)}{\lambda} = \frac{1}{\lambda}\frac{\mathbf{E}[Xe^{\lambda X}]}{\mathbf{E}[e^{\lambda X}]} - \frac{1}{\lambda^2}\log \mathbf{E}[e^{\lambda X}] = \frac{1}{\lambda^2}\frac{\mathrm{Ent}[e^{\lambda X}]}{\mathbf{E}[e^{\lambda X}]}.$$

Thus the assumption of the lemma yields

$$\frac{\psi(\lambda)}{\lambda} = \int_0^\lambda \frac{1}{u^2}\frac{\mathrm{Ent}[e^{uX}]}{\mathbf{E}[e^{uX}]}\,du \le \frac{\lambda\sigma^2}{2}$$

using the fundamental theorem of calculus and $\lim_{\lambda\downarrow 0}\lambda^{-1}\psi(\lambda) = 0$. □

As an immediate consequence, we see that if a random variable X satisfies

$$\mathrm{Ent}[e^{\lambda X}] \le \frac{\lambda^2\sigma^2}{2}\,\mathbf{E}[e^{\lambda X}] \quad \text{for all } \lambda \in \mathbb{R},$$

then X is σ^2-subgaussian. Thus we have a sufficient condition for the subgaussian property in terms of entropy. In fact, up to a constant factor, the converse is also true: if X is $\frac{\sigma^2}{4}$-subgaussian, then the assumption of Lemma 3.13 holds (Problem 3.12). We may therefore view the above entropy inequality as an alternative formulation of the subgaussian property of a random variable X.

It may not be immediately evident what we have accomplished. Indeed, we have obtained yet another formulation of the subgaussian property, which may appear at first sight no more useful than any other (and perhaps somewhat less intuitive than most). However, the formulation in terms of entropy proves to be a very powerful idea. For example, we will presently show that entropy

obeys an *exact analogue of the tensorization property of the variance*, from which its utility in high dimension will be immediately obvious. In fact, it turns out that entropy behaves in many ways like the variance. Once we are comfortable with this idea, it will become evident that several other notions from Chapter 2 extend naturally to the subgaussian setting.

To formulate the tensorization inequality, let $X_1, \ldots, X_n$ be independent random variables. For each function $f(x_1, \ldots, x_n)$, we define the function

$$\mathrm{Ent}_i f(x_1, \ldots, x_n) := \mathrm{Ent}[f(x_1, \ldots, x_{i-1}, X_i, x_{i+1}, \ldots, x_n)].$$

That is, $\mathrm{Ent}_i f(x)$ is the entropy of $f(X_1, \ldots, X_n)$ with respect to the variable X_i only, the remaining variables being kept fixed.

Theorem 3.14 (Tensorization of entropy). *We have*

$$\mathrm{Ent}[f(X_1, \ldots, X_n)] \le \mathbf{E}\left[\sum_{i=1}^{n} \mathrm{Ent}_i f(X_1, \ldots, X_n)\right]$$

whenever $X_1, \ldots, X_n$ are independent.

To prove Theorem 3.14 we will need a fundamental result that can be viewed as an analogue of Hölder's inequality for entropy.

Lemma 3.15 (Variational formula for entropy). *We have*

$$\mathrm{Ent}[Z] = \sup\{\mathbf{E}[ZX] : X \text{ is a random variable satisfying } \mathbf{E}[e^X] = 1\}.$$

Proof. Let $\mathbf{E}[e^X] = 1$ and define the new probability $d\mathbf{Q} = e^X d\mathbf{P}$. Then

$$\begin{aligned}\mathrm{Ent}[Z] - \mathbf{E}[ZX] &= \mathbf{E}[Z \log Z] - \mathbf{E}[Z \log e^X] - \mathbf{E}[Z] \log \mathbf{E}[Z] \\ &= \mathbf{E}_{\mathbf{Q}}[e^{-X} Z \log(e^{-X} Z)] - \mathbf{E}_{\mathbf{Q}}[e^{-X} Z] \log \mathbf{E}_{\mathbf{Q}}[e^{-X} Z].\end{aligned}$$

As $x \mapsto x \log x$ is convex, it follows from Jensen's inequality that $\mathrm{Ent}[Z] - \mathbf{E}[ZX] \ge 0$ for every random variable X such that $\mathbf{E}[e^X] = 1$. But note that $\mathrm{Ent}[Z] - \mathbf{E}[ZX] = 0$ for $X = \log(Z/\mathbf{E}[Z])$, and thus the proof is complete. □

We can now complete the proof of Theorem 3.14.

Proof (Theorem 3.14). Let $Z = f(X_1, \ldots, X_n)$, and define for $k = 1, \ldots, n$

$$U_k = \log \mathbf{E}[Z|X_1, \ldots, X_k] - \log \mathbf{E}[Z|X_1, \ldots, X_{k-1}].$$

The evidently

$$\mathrm{Ent}[Z] = \mathbf{E}[Z(\log Z - \log \mathbf{E}[Z])] = \sum_{k=1}^{n} \mathbf{E}[ZU_k].$$

On the other hand, note that

$$\begin{aligned}
&\mathbf{E}[e^{U_k}|X_1,\dots,X_{k-1},X_{k+1},\dots,X_n]\\
&=\frac{\mathbf{E}[\mathbf{E}[Z|X_1,\dots,X_k]|X_1,\dots,X_{k-1},X_{k+1},\dots,X_n]}{\mathbf{E}[Z|X_1,\dots,X_{k-1}]}\\
&=\frac{\mathbf{E}[\mathbf{E}[Z|X_1,\dots,X_k]|X_1,\dots,X_{k-1}]}{\mathbf{E}[Z|X_1,\dots,X_{k-1}]}=1,
\end{aligned}$$

where we have used that $X_{k+1},\dots,X_n$ and $X_1,\dots,X_k$ are independent. Therefore, applying Lemma 3.15 conditionally yields

$$\begin{aligned}
&\mathbf{E}[ZU_k|X_1,\dots,X_{k-1},X_{k+1},\dots,X_n]\\
&\le \mathrm{Ent}[Z|X_1,\dots,X_{k-1},X_{k+1},\dots,X_n]\\
&=\mathrm{Ent}_k f(X_1,\dots,X_n),
\end{aligned}$$

where $\mathrm{Ent}[Z|\mathcal{G}]:=\mathbf{E}[Z\log Z|\mathcal{G}]-\mathbf{E}[Z|\mathcal{G}]\log\mathbf{E}[Z|\mathcal{G}]$. In particular,

$$\mathbf{E}[ZU_k]\le\mathbf{E}[\mathrm{Ent}_k f(X_1,\dots,X_n)],$$

by the tower property, and the proof is complete. □

The entropic formulation of the subgaussian property and the tensorization inequality for entropy immediately indicate what type of inequalities we should prove to obtain subgaussian concentration inequalities. Informally, suppose we can prove an inequality in one dimension of the form

$$\text{“ entropy}(e^g)\lesssim\mathbf{E}[\,|\text{gradient}(g)|^2\,e^g\,]\text{. ”}$$

Then we obtain for product measures in any dimension, by tensorization,

$$\text{“ entropy}(e^{\lambda f})\lesssim\mathbf{E}[\,\|\text{gradient}(\lambda f)\|^2\,e^{\lambda f}\,]\text{, ”}$$

and thus f is subgaussian with variance proxy of order $\||\|\text{gradient}(f)\|^2\|_\infty$. This is precisely the subgaussian counterpart of the Poincaré inequalities

$$\text{“ variance}(f)\lesssim\mathbf{E}[\,\|\text{gradient}(f)\|^2\,]\text{ ”}$$

that were obtained in Chapter 2. The entropy inequalities informally described above are one form of a class of inequalities called *log-Sobolev inequalities*. In the next section, we will develop a general framework for understanding and proving log-Sobolev inequalities that is similar to (but less powerful than) the theory developed in Chapter 2 for Poincaré inequalities.

As a first illustration of the entropy method, let us prove a log-Sobolev counterpart of the trivial variance inequality of Lemma 2.1.

Lemma 3.16 (Discrete log-Sobolev). *Let* $D^-f:=f-\inf f$. *Then*

$$\mathrm{Ent}[e^f]\le\mathrm{Cov}[f,e^f]\le\mathbf{E}[|D^-f|^2e^f].$$

Remark 3.17. The constant in this inequality is not optimal. Improved constants will be derived in Problem 3.13 below. The suboptimal result is given here as its simple proof seems the most intuitive and insightful.

Proof. Note that $\log \mathbf{E}[e^f] \geq \mathbf{E}[f]$ by Jensen's inequality. Therefore

$$\mathrm{Ent}[e^f] = \mathbf{E}[fe^f] - \mathbf{E}[e^f] \log \mathbf{E}[e^f] \leq \mathbf{E}[fe^f] - \mathbf{E}[f]\mathbf{E}[e^f] = \mathrm{Cov}[f, e^f].$$

To prove the second part, note that

$$\mathrm{Cov}[f, e^f] = \mathbf{E}[(f - \inf f)(e^f - \mathbf{E}[e^f])] \leq \mathbf{E}[(f - \inf f)(e^f - e^{\inf f})].$$

Since e^x is convex, the first-order condition for convexity implies $e^f - e^{\inf f} \leq e^f(f - \inf f)$. Substituting into the above expression completes the proof. □

We can now obtain Gaussian tail bounds in terms of one-sided differences

$$D_i^- f(x) := f(x_1, \ldots, x_n) - \inf_z f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n),$$
$$D_i^+ f(x) := \sup_z f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_n)$$

by combining the discrete log-Sobolev inequality with tensorization of entropy.

Theorem 3.18 (Bounded difference inequality). *For all* $t \geq 0$

$$\mathbf{P}[f(X_1, \ldots, X_n) \geq \mathbf{E}f(X_1, \ldots, X_n) + t] \leq e^{-t^2/4\| \sum_{i=1}^n |D_i^- f|^2 \|_\infty},$$
$$\mathbf{P}[f(X_1, \ldots, X_n) \leq \mathbf{E}f(X_1, \ldots, X_n) - t] \leq e^{-t^2/4\| \sum_{i=1}^n |D_i^+ f|^2 \|_\infty}$$

whenever $X_1, \ldots, X_n$ *are independent. In particular, the random variable* $f(X_1, \ldots, X_n)$ *is subgaussian with variance proxy* $2\| \sum_{i=1}^n |D_i f|^2 \|_\infty$.

Proof. By Lemma 3.16, we have

$$\mathrm{Ent}_i[e^f] \leq \mathbf{E}[|D_i^- f|^2 e^f | X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n].$$

Thus we have for $\lambda \geq 0$

$$\mathrm{Ent}[e^{\lambda f}] \leq \lambda^2 \mathbf{E}\left[\sum_{i=1}^n |D_i^- f|^2 e^{\lambda f}\right] \leq \lambda^2 \left\| \sum_{i=1}^n |D_i^- f|^2 \right\|_\infty E[e^{\lambda f}]$$

using the tensorization Theorem 3.14, where we used that $D_i^-(\lambda f) = \lambda D_i^- f$ for $\lambda \geq 0$. Thus Lemma 3.13 and the Chernoff bound yields the upper tail bound. The lower tail bound is obtained by applying the upper tail bound to $-f$ and noting that $D_i^-(-f) = -D_i^+ f$. As $|D_i^- f| \leq |D_i f|$ and $|D_i^+ f| \leq |D_i f|$, the subgaussian property is deduced identically from Lemma 3.13. □

The bounds of Theorem 3.18 are a vast improvement over McDiarmid's inequality of Theorem 3.11: here the variance proxy is a genuine upper bound on the square gradient $\|\sum_{i=1}^n |D_i f|^2\|_\infty$, while in McDiarmid's inequality the gradient must be bounded coordinatewise $\sum_{i=1}^n \|D_i f\|_\infty^2$. We also obtain finer bounds in terms of one-sided differences, which is important in many applications. What enables these improved bounds is that the log-Sobolev inequality tensorizes much more efficiently than the subgaussian property itself. Indeed, note how we have kept the gradient inside the expectation throughout the tensorization process, and only took its uniform norm at the end to obtain a subgaussian inequality; had we directly tensorized the subgaussian bound of Lemma 3.13, we would only be able to recover McDiarmid's inequality.

On the other hand, unlike in the previous bounds we have encountered, we see here an important case where the upper and lower tail bounds are not symmetric: the upper tail bound is given in terms of the negative gradient $D_i^- f$, while the lower tail bound is given in terms of the positive gradient $D_i^+ f$. There are applications where only one of these quantities can be controlled.

Example 3.19 (Random matrices). We recall the setting of Example 2.5. Let M be an $n \times n$ symmetric matrix where $\{M_{ij} : i \geq j\}$ are i.i.d. symmetric Bernoulli random variables $\mathbf{P}[M_{ij} = \pm 1] = \frac{1}{2}$. We denote by $\lambda_{\max}(M)$ the largest eigenvalue of M, and by $v_{\max}(M)$ a corresponding eigenvector.

It was shown in Example 2.5 that

$$D_{ij}^- \lambda_{\max}(M) \leq 4|v_{\max}(M)_i||v_{\max}(M)_j|.$$

Thus we can estimate

$$\left\| \sum_{i,j=1}^n |D_{ij}^- \lambda_{\max}(M)|^2 \right\|_\infty \leq 16 \left[\sum_{i=1}^n |v_{\max}(M)_i|^2 \right]^2 = 16,$$

and we therefore obtain by Theorem 3.18 the upper tail bound

$$\mathbf{P}[\lambda_{\max}(M) - \mathbf{E}\lambda_{\max}(M) \geq t] \leq e^{-t^2/64}$$

for all $t \geq 0$. This is a much sharper control of the fluctuations *above* the mean in comparison to the variance bound of Example 2.5.

On the other hand, we cannot use Theorem 3.18 to control the fluctuations *below* the mean. Indeed, for the positive gradient, we can compute

$$D_{ij}^+ \lambda_{\max}(M) \leq 4|v_{\max}(M^{(ij)})_i||v_{\max}(M^{(ij)})_j|$$

as in Example 2.5, where $M^{(ij)}$ is the matrix such that $M_{ij}^{(ij)} = M_{ji}^{(ij)}$ is chosen to maximize $\lambda_{\max}(M)$ while the remaining entries are kept fixed. Now there is no reason to expect that $\sum_{i=1}^n |v_{\max}(M^{(ij)})_i|^2$ is bounded uniformly in the dimension (as a different matrix $M^{(ij)}$ is chosen for every entry i), and thus we cannot obtain a dimension-free lower tail bound in this manner.

It does not seem to be possible to prove a subgaussian lower tail bound in terms of $D_i^- f$ (or, equivalently, an upper tail bound in terms of $D_i^+ f$). It is instructive to attempt to repeat the proof of the discrete log-Sobolev inequality of Lemma 3.16 in terms of the positive gradient: this gives at best

$$\mathrm{Ent}[e^f] \le \mathbf{E}[|D^+ f|^2]\,\mathbf{E}[e^f],$$

which does not behave well under tensorization. Thus the situation is inherently asymmetric. However, in many examples where the negative gradient $D_i^- f$ can be controlled, it turns out that in fact a stronger property holds as well that makes it possible to obtain both upper and lower tail bounds using a result known as Talagrand's concentration inequality. The machinery needed to prove such bounds will be discussed in the next chapter.

Problems

3.12 (Subgaussian variables and entropy). Lemma 3.13 states that if

$$\mathrm{Ent}[e^{\lambda X}] \le \frac{\lambda^2\sigma^2}{2}\,\mathbf{E}[e^{\lambda X}] \quad \text{for all } \lambda \in \mathbb{R},$$

then the random variable X is σ^2-subgaussian. Prove the following converse implication: if X is $\frac{\sigma^2}{4}$-subgaussian, then the above entropy inequality holds. We may therefore view this property as yet another equivalent formulation of the subgaussian property in the spirit of Problem 3.1.

Hint: Note that $\mathrm{Ent}[e^{\lambda X}]/\mathbf{E}[e^{\lambda X}] = \mathbf{E}[Z\log Z]$ for $Z = e^{\lambda X}/\mathbf{E}[e^{\lambda X}]$. Now use concavity of the logarithm and that $\mathbf{E}[e^{\lambda(X-\mathbf{E}X)}] \ge 1$ (why?).

3.13 (Optimal discrete log-Sobolev constants). The discrete log-Sobolev inequality of Lemma 3.16 yields a bounded difference inequality with variance proxy $2\|\sum_{i=1}^n |D_i^- f|^2\|_\infty$. The constant is not optimal: in view of the bounded difference inequality for the variance (Corollary 2.4), we would expect a variance proxy $\|\sum_{i=1}^n |D_i^- f|^2\|_\infty$ without the additional factor 2. Moreover, in terms of the two-sided difference, we would expect $\frac{1}{4}\|\sum_{i=1}^n |D_i f|^2\|_\infty$ which gains an additional factor $\frac{1}{4}$. It turns out that a more refined proof of the discrete log-Sobolev inequality can attain these improved numerical constants.

One place where we lose in the proof of Lemma 3.16 is in estimating the entropy by a covariance. Instead, we can use a variational principle for the entropy to obtain an improved upper bound. Of course, Lemma 3.15 is useless for this purpose as it can only yield lower bounds on the entropy.

a. Prove the following variational principle:

$$\mathrm{Ent}[Z] = \inf_{t>0} \mathbf{E}[Z\log Z - Z\log t - Z + t].$$

b. Use the above variational principle to show that

$$\mathrm{Ent}[e^f] \le \mathbf{E}[\varphi(D^- f)e^f], \qquad \varphi(x) := e^{-x} + x - 1.$$

c. Show that $\varphi(x) \le \frac{x^2}{2}$ for $x \ge 0$, and use it to improve Lemma 3.16 to

$$\mathrm{Ent}[e^f] \le \frac{1}{2}\,\mathbf{E}[|D^- f|^2 e^f].$$

d. We now consider the two-sided gradient $Df = \sup f - \inf f$. Use the bound $\psi''(\lambda) \le (Df)^2/4$ on the log-moment generating function from the proof of Lemma 3.6 and reason as in the proof of Lemma 3.13 to show that

$$\mathrm{Ent}[e^f] \le \frac{1}{8}\,\mathbf{E}[|Df|^2 e^f].$$

Hint: express $\mathrm{Ent}[e^{\lambda f}]/\mathbf{E}[e^{\lambda f}]$ in terms of $\psi(\lambda)$ and its derivative and apply the fundamental theorem of calculus.

3.14 (Rademacher processes). Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent symmetric Bernoulli random variables $\mathbf{P}[\varepsilon_i = \pm 1] = \frac{1}{2}$, and let $T \subseteq \mathbb{R}^n$. Define

$$Z = \sup_{t \in T} \sum_{k=1}^n \varepsilon_k t_k.$$

Show that for $t \ge 0$

$$\mathbf{P}[Z - \mathbf{E}Z \ge t] \le e^{-t^2/4\sigma^2} \quad \text{with} \quad \sigma^2 = 4 \sup_{t \in T} \sum_{k=1}^n t_k^2.$$

This is a crucial improvement over the result obtained in Problem 3.7 using McDiarmid's inequality. However, here we only obtain an upper tail bound: Talagrand's inequality is needed to obtain a matching lower tail.

3.15 (Convex log-Sobolev). Show that for a *convex* function $f : [a, b] \to \mathbb{R}$

$$\mathrm{Ent}[e^f] \le (b-a)^2 \mathbf{E}[|f'|^2 e^f],$$

where f' is the calculus (not discrete) derivative. Conclude that if $f : \mathbb{R}^n \to \mathbb{R}$ is convex and L-Lipschitz, i.e., $|f(x) - f(y)| \le L\|x - y\|$ for all $x, y \in \mathbb{R}^n$, and if $X_1, \ldots, X_n$ are independent with values in $[a, b]$, then for every $t \ge 0$

$$\mathbf{P}[f(X_1, \ldots, X_n) - \mathbf{E}f(X_1, \ldots, X_n) \ge t] \le e^{-t^2/4(b-a)^2 L^2}.$$

Note that this does not yield a lower tail bound: if f is convex, $-f$ is concave.
Hint: Recall Problem 2.5.

3.16 (Exponential Poincaré inequalities). In this problem, we will assume the validity of a general kind of log-Sobolev inequality of the form

$$\mathrm{Ent}[e^{\lambda f}] \le \frac{\lambda^2}{2}\,\mathbf{E}[\Gamma(f) e^{\lambda f}]$$

for $\lambda \geq 0$, where $\Gamma(f)$ is some suitable notion of "$\|\text{gradient}(f)\|^2$." Such an inequality can be used to prove that f is $\|\Gamma(f)\|_\infty^2$-subgaussian using Lemma 3.13. In this problem, we will show that it is in fact possible to obtain more precise control on the moment generating function of f. In fact, we will prove

$$\mathbf{E}[e^{f-\mathbf{E}f}] \leq \mathbf{E}[e^{\Gamma(f)}],$$

which could be viewed as an "exponential Poincaré inequality."

a. Show that

$$\text{Ent}[e^{\lambda f}] \geq \lambda^2 \mathbf{E}[\Gamma(f) e^{\lambda f}] - \mathbf{E}[e^{\lambda f}] \log \mathbf{E}[e^{\lambda^2 \Gamma(f)}].$$

Hint: use the variational formula for entropy.

b. Use the log-Sobolev inequality to show that

$$\text{Ent}[e^{\lambda f}] \leq \lambda^2 \gamma(\lambda^2)\, \mathbf{E}[e^{\lambda f}], \qquad \gamma(s) = \log \|e^{\Gamma(f)}\|_s.$$

c. Prove the exponential Poincaré inequality $\mathbf{E}[e^{f-\mathbf{E}f}] \leq \mathbf{E}[e^{\Gamma(f)}]$.

3.4 Log-Sobolev inequalities

In the previous section, we have seen that one can prove dimension-free subgaussian concentration inequalities by establishing log-Sobolev inequalities. We proved a simple discrete log-Sobolev inequality using elementary methods, and used it to obtain subgaussian analogues of the bounded difference inequalities for the variance of section 2.1. As in the case of the variance, however, we would like to develop machinery to prove log-Sobolev inequalities in different settings and with respect to different notions of gradient.

In this section, we will develop a partial log-Sobolev analogue of the powerful Markov process machinery developed in section 2.3 to prove Poincaré inequalities: we will show that the validity of a log-Sobolev inequality for a measure μ is intimately connected to exponential convergence of a Markov semigroup to its stationary measure μ *in the sense of entropy* (rather than in $L^2(\mu)$, which would only yield a Poincaré inequality as in section 2.3). To be precise, we will prove an entropic analogue of the "easy" implications $3 \Rightarrow 1 \Leftrightarrow 2$ of Theorem 2.18 whose proofs do not require reversibility. It is not too surprising that we cannot reproduce the remaining implications in the entropic setting: exploiting reversibility essentially requires the structure of $L^2(\mu)$, while entropy (unlike the variance) is not an $L^2(\mu)$ notion (in the context of Remark 2.36, note that the entropy is not naturally expressed in terms of the spectrum of the generator). As a consequence, our log-Sobolev analogue of Theorem 2.18 is significantly less powerful than its Poincaré counterpart. Nonetheless, we will see that this approach remains extremely useful, particularly in the setting of continuous distributions.

In the sequel, we define $\text{Ent}_\mu[f] := \mu(f \log f) - \mu f \log \mu f$.

Theorem 3.20 (Log-Sobolev inequality). *Let P_t be a Markov semigroup with stationary measure μ. The following are equivalent:*

1. $\mathrm{Ent}_\mu[f] \le c\mathcal{E}(\log f, f)$ *for all* f (log-Sobolev inequality).
2. $\mathrm{Ent}_\mu[P_t f] \le e^{-t/c}\mathrm{Ent}_\mu[f]$ *for all* f, t (entropic exponential ergodicity).

Moreover, if $\mathrm{Ent}_\mu[P_t f] \to 0$ *as* $t \to \infty$ (entropic ergodicity), *then*

3. $\mathcal{E}(\log P_t f, P_t f) \le e^{-t/c}\mathcal{E}(\log f, f)$ *for all* f, t

implies 1 and 2 above.

Proof. An elementary computation yields

$$\frac{d}{dt}\mathrm{Ent}_\mu[P_t f] = \mu(\mathscr{L}P_t f\ \log P_t f) + \mu(\mathscr{L}P_t f) = -\mathcal{E}(\log P_t f, P_t f),$$

where we have used that $\mu(\mathscr{L}P_t f) = \frac{d}{dt}\mu(P_t f) = \frac{d}{dt}\mu f = 0$. We now prove:

- $3 \Rightarrow 1$: By the fundamental theorem of calculus, 3 implies

$$\mathrm{Ent}_\mu[f] = \int_0^\infty \mathcal{E}(\log P_t f, P_t f)\, dt \le \mathcal{E}(\log f, f) \int_0^\infty e^{-t/c} dt = c\mathcal{E}(\log f, f).$$

- $1 \Rightarrow 2$: Assuming 1, we obtain 2 directly from

$$\frac{d}{dt}\mathrm{Ent}_\mu[P_t f] = -\mathcal{E}(\log P_t f, P_t f) \le -\frac{1}{c}\mathrm{Ent}_\mu[P_t f].$$

- $2 \Rightarrow 1$: Assuming 2, we can compute

$$\mathcal{E}(\log f, f) = \lim_{t \downarrow 0} \frac{\mathrm{Ent}_\mu[f] - \mathrm{Ent}_\mu[P_t f]}{t} \ge \lim_{t \downarrow 0} \frac{1 - e^{-t/c}}{t}\mathrm{Ent}_\mu[f].$$

This completes the proof. □

As in section 2.3, it may not be obvious at first sight why the inequality $\mathrm{Ent}_\mu[f] \le c\mathcal{E}(\log f, f)$ should be viewed as a log-Sobolev inequality in the sense that we introduced in the previous section. Once we consider some illuminating examples, it should become clear that this is indeed the case.

Example 3.21 (Discrete log-Sobolev inequality). Let μ be any probability measure, and define a Markov process X_t as follows:

- Draw $X_0 \sim \mu$.
- Let N_t be a Poisson process with rate 1, independent of X_0. Each time N_t jumps, replace the current value of X_t by an independent sample from μ.

This is nothing other than the case $n = 1$ of the ergodic Markov process defined in Example 2.29. In particular, it is easily seen that μ is the stationary measure of X_t, and that its semigroup and Dirichlet form are given by

$$P_t f = e^{-t} f + (1 - e^{-t})\mu f, \qquad \mathcal{E}(f, g) = \mathrm{Cov}_\mu[f, g].$$

Now note that, by the convexity of $x \mapsto x \log x$,

$$P_t f \log P_t f \le e^{-t} f \log f + (1 - e^{-t})\, \mu f \log \mu f.$$

Thus we have

$$\mathrm{Ent}_\mu[P_t f] = \mu(P_t f \log P_t f) - \mu f \log \mu f \le e^{-t} \mathrm{Ent}_\mu[f],$$

and we conclude by implication $2 \Rightarrow 1$ of Theorem 3.20 that

$$\mathrm{Ent}_\mu[f] \le \mathcal{E}(\log f, f) = \mathrm{Cov}_\mu[\log f, f].$$

Replacing f by e^g, we see that we have obtained the discrete log-Sobolev inequality of Lemma 3.16 as a special case of Theorem 3.20.

Remark 3.22. We have seen in Example 2.29 that the characterization of Poincaré inequalities of Theorem 2.18 is sufficiently powerful to reproduce that tensorization inequality for variance. In contrast, in view of the above example, we see that Theorem 3.20 cannot reproduce the tensorization inequality for entropy. Indeed, extending the above example to the setting of Example 2.29, we can obtain at best an inequality of the form

$$\mathrm{Ent}[f(X_1, \ldots, X_n)] \le \mathbf{E}\left[\sum_{i=1}^n \mathrm{Cov}_i[\log f, f](X_1, \ldots, X_n)\right],$$

which has covariances on the right-hand side instead of entropies (that is, Theorem 3.20 yields a combination of the tensorization Theorem 3.14 and the discrete log-Sobolev inequality of Lemma 3.16). Thus the result of Theorem 3.20 is inherently less complete than that of Theorem 2.18. On the other hand, Theorem 3.20 still provides a powerful tool to prove log-Sobolev inequalities. This is particularly useful in the continuous case, as we will see presently.

Example 3.23 (Gaussian log-Sobolev inequality). Let us prove a log-Sobolev inequality for the standard Gaussian distribution $\mu = N(0, 1)$ in one dimension (we will subsequently use tensorization to extend to higher dimensions). To this end, we will again use the Ornstein-Uhlenbeck process X_t introduced in Example 2.22. In particular, we recall two important properties of the Ornstein-Uhlenbeck process that were proved in Example 2.22:

$$\mathcal{E}(f, g) = \mu(f' g'), \qquad (P_t f)' = e^{-t} P_t f'.$$

Using these properties, we will now proceed to prove a log-Sobolev inequality.

Note that $(\log f)' f' = |f'|^2 / f$. We therefore have

$$(\log P_t f)'(P_t f)' = e^{-2t} \frac{|P_t f'|^2}{P_t f}.$$

By Cauchy-Schwarz, we obtain

$$|P_t f'|^2 = \left| P_t\left(\frac{f'}{\sqrt{f}}\sqrt{f}\right)\right|^2 \le P_t\left(\frac{|f'|^2}{f}\right) P_t f = P_t((\log f)' f')\, P_t f,$$

and consequently

$$(\log P_t f)'(P_t f)' \le e^{-2t}\, P_t((\log f)' f').$$

Integrating with respect to μ on both sides yields

$$\mathcal{E}(\log P_t f, P_t f) \le e^{-2t}\mathcal{E}(\log f, f),$$

and thus the implication $3 \Rightarrow 1$ of Theorem 3.20 yields

$$\mathrm{Ent}_\mu[f] \le \frac{1}{2}\mathcal{E}(\log f, f).$$

This is the log-Sobolev inequality for the Gaussian distribution.

Having proved the Gaussian log-Sobolev inequality in one dimension, we immediately obtain an n-dimensional inequality by tensorization.

Theorem 3.24 (Gaussian log-Sobolev inequality). *Let $X_1, \ldots, X_n$ be independent Gaussian random variables with zero mean and unit variance. Then*

$$\mathrm{Ent}[f(X_1, \ldots, X_n)] \le \frac{1}{2}\,\mathbf{E}[\nabla f(X_1, \ldots, X_n) \cdot \nabla \log f(X_1, \ldots, X_n)]$$

for every $f \ge 0$.

Why is this a log-Sobolev inequality in the sense of the previous section? Note that, by the chain rule, the inequality of Theorem 3.24 is equivalent to

$$\mathrm{Ent}[e^{f(X_1,\ldots,X_n)}] \le \frac{1}{2}\,\mathbf{E}[\|\nabla f(X_1, \ldots, X_n)\|^2\, e^{f(X_1,\ldots,X_n)}]$$

for every f. This is precisely the type of inequality that arises in the previous section. In particular, in this form, it is immediately evident that Theorem 3.24 provides the key ingredient to prove a Gaussian concentration inequality. The following result is one of the most important properties of Gaussian variables.

Theorem 3.25 (Gaussian concentration). *Let $X_1, \ldots, X_n$ be independent Gaussian random variables with zero mean and unit variance. Then*

$$\mathbf{P}[f(X_1, \ldots, X_n) - \mathbf{E}f(X_1, \ldots, X_n) \ge t] \le e^{-t^2/2\sigma^2}$$

for all $t \ge 0$, where $\sigma^2 = \|\|\nabla f\|^2\|_\infty$. In fact, $f(X_1, \ldots, X_n)$ is σ^2-subgaussian.

Proof. By Theorem 3.24 and the chain rule, we can estimate

$$\mathrm{Ent}[e^{\lambda f(X_1,\ldots,X_n)}] \le \frac{\lambda^2 \||\nabla f|^2\|_\infty}{2}\, \mathbf{E}[e^{\lambda f(X_1,\ldots,X_n)}]$$

for all $\lambda \in \mathbb{R}$. The result now follows from Lemma 3.13. □

Remark 3.26. In the Gaussian case, we have seen several different forms of the log-Sobolev inequality. Beside the form as stated in Theorem 3.24

$$\mathrm{Ent}[f] \le \frac{1}{2}\,\mathbf{E}[\nabla f \cdot \nabla \log f] = \frac{1}{2}\mathcal{E}(\log f, f)$$

(which corresponds to the inequality in Theorem 3.20), we can write

$$\mathrm{Ent}[f] \le \frac{1}{2}\,\mathbf{E}\left[\frac{\|\nabla f\|^2}{f}\right]$$

(which is in fact the form that was used in the proof of Theorem 3.24), or

$$\mathrm{Ent}[e^f] \le \frac{1}{2}\,\mathbf{E}[\|\nabla f\|^2 e^f]$$

(which was used in the proof of Theorem 3.25). Another equivalent form is

$$\mathrm{Ent}[f^2] \le 2\mathbf{E}[\|\nabla f\|^2] = 2\mathcal{E}(f, f).$$

The latter form is in fact the "classical" form of the log-Sobolev inequality as it is found in the analysis literature. For the Gaussian case, all these forms of the log-Sobolev inequality are equivalent due to the fact that the Dirichlet form is given in terms of a gradient that satisfies the chain rule. This is not the case in general, however: for many Markov processes (such as in Remark 3.21) the Dirichlet form does not satisfy the chain rule property, and in this case the above inequalities are typically not equivalent to one another. Nonetheless, it is often possible to deduce useful forms of these inequalities even in the absence of the chain rule, as we did, for example, in the proof of Lemma 3.16. We will loosely refer to all inequalities of this kind as "log-Sobolev inequalities."

Remark 3.27. The "classical" form of the log-Sobolev inequality reads

$$\mathbf{E}[f^2 \log f] - \mathbf{E}[f^2] \log \|f\|_2 \le c\|\nabla f\|_2^2,$$

while the Poincaré inequality reads

$$\mathbf{E}[f^2] - \mathbf{E}[f]^2 \le c\|\nabla f\|_2^2.$$

When viewed in this manner, the log-Sobolev inequality looks only slightly stronger than the Poincaré inequality: the latter controls the L^2-norm of a function by the L^2-norm of its gradient, while the former controls the function

in a slightly stronger (by a logarithmic factor) $L^2 \log L$-norm.[1] As we have seen, this apparently minor improvement has far-reaching consequences.

In classical analysis, an important role is played by *Sobolev inequalities* that have the form $\|f - \mathbf{E}f\|_q \le c\|\nabla f\|_2$ for $q > 2$. Such inequalities are even better than log-Sobolev inequalities: they ensure that the L^q-norm of function is controlled by the L^2-norm of its gradient, while log-Sobolev inequalities only improve over L^2 by a logarithmic factor (hence the name). However, unlike log-Sobolev inequalities, classical Sobolev inequalities do not tensorize. It is for this reason that log-Sobolev inequalities are much more important than classical Sobolev inequalities in high-dimensional probability.

In view of the previous remark, it is natural to conclude that log-Sobolev inequalities are strictly stronger than Poincaré inequalities, but this is not entirely obvious. We conclude this section by showing that this is indeed the case, even in the more general setting of Theorem 3.20. This clarifies, in particular, that the methods developed in this chapter to prove concentration inequalities can be viewed in a precise sense as direct extensions of the theory developed in the previous chapter to prove variance bounds.

Lemma 3.28. *The log-Sobolev inequality* $\mathrm{Ent}[f] \le c\mathcal{E}(\log f, f)$ *for all* $f \ge 0$ *implies the Poincaré inequality* $\mathrm{Var}[f] \le 2c\mathcal{E}(f,f)$ *for all* f.

Proof. The log-Sobolev inequality states for $\lambda \ge 0$

$$\mathbf{E}[\lambda f e^{\lambda f}] - \mathbf{E}[e^{\lambda f}] \log \mathbf{E}[e^{\lambda f}] \le c\mathcal{E}(\lambda f, e^{\lambda f}).$$

As $\mathcal{E}(f,1) = 0$, we can estimate

$$\mathcal{E}(\lambda f, e^{\lambda f}) = \lambda^2 \mathcal{E}(f,f) + o(\lambda^2),$$

while we have

$$\mathbf{E}[\lambda f e^{\lambda f}] = \lambda \mathbf{E}[f] + \lambda^2 \mathbf{E}[f^2] + o(\lambda^2),$$

and

$$\mathbf{E}[e^{\lambda f}] \log \mathbf{E}[e^{\lambda f}] = \lambda \mathbf{E}[f] + \lambda^2 \{\mathbf{E}[f^2] + \mathbf{E}[f]^2\}/2 + o(\lambda^2).$$

Thus we obtain the Poincaré inequality $\mathrm{Var}[f] \le 2c\mathcal{E}(f,f)$ by dividing the log-Sobolev inequality $\mathrm{Ent}[e^{\lambda f}] \le c\mathcal{E}(\lambda f, e^{\lambda f})$ by λ^2 and letting $\lambda \downarrow 0$. □

Problems

3.17 (Relative entropy convergence). As Theorem 3.20 does not require P_t to be reversible, the log-Sobolev inequality $\mathrm{Ent}_\mu[f] \le c\mathcal{E}(\log f, f)$ is not necessarily equivalent to the reverse inequality $\mathrm{Ent}_\mu[f] \le c\mathcal{E}(f, \log f)$. There is, however, a dual form of Theorem 3.20 that will yield the latter.

[1] While the idea expressed here is intuitive, it should be noted that entropy is not a norm. However, the statement can be made precise in terms of Orlicz norms.

Define the *relative entropy* between probability measures ν and μ as

$$D(\nu||\mu) := \mathrm{Ent}_\mu\left[\frac{d\nu}{d\mu}\right] \quad \text{for } \nu \ll \mu,$$

and $D(\nu||\mu) := \infty$ otherwise. The relative entropy should be viewed as a notion of "distance" between probability measures: in particular $D(\nu||\mu) \geq 0$ and $D(\nu||\mu) = 0$ if and only of $\mu = \nu$. Note, however, that $D(\nu||\mu)$ is not a metric (it is neither symmetric, nor does it satisfy a triangle inequality). The relative entropy will play an important role in the next chapter.

For every probability measure ν, we can define the probability measure νP_t by setting $(\nu P_t)f = \nu(P_t f)$ for every function f. Note that νP_t is precisely the law of X_t given that the initial condition X_0 is drawn from ν: indeed, if $X_0 \sim \nu$, then $\nu P_t f = \mathbf{E}[P_t f(X_0)] = \mathbf{E}[\mathbf{E}[f(X_t)|X_0]] = \mathbf{E}[f(X_t)]$. In particular, the stationary measure μ satisfies, by its definition, $\mu P_t = \mu$ for all t.

a. Let $h = \frac{d\nu}{d\mu}$. Show that $D(\nu P_t||\mu) = \mathrm{Ent}_\mu[P_t^* h]$, where P_t^* is the adjoint of the semigroup P_t (that is, $\langle f, P_t g\rangle_\mu = \langle P_t^* f, g\rangle_\mu$ for all f, g).

b. Show that the log-Sobolev inequality

$$\mathrm{Ent}_\mu[f] \leq c\mathcal{E}(f, \log f) \quad \text{for all } f$$

holds if and only if P_t is exponentially ergodic in relative entropy:

$$D(\nu P_t||\mu) \leq e^{-t/c} D(\nu||\mu) \quad \text{for all } t, \nu.$$

3.18 (Norms of Gaussian vectors). The goal of this problem is to prove some classical results about norms of Gaussian vectors. We begin with a simple but important consequence of Gaussian concentration.

a. Let $X \sim N(0, \Sigma)$ be an n-dimensional centered Gaussian vector with arbitrary covariance matrix Σ. Prove that (see Problem 2.8 for a hint)

$$\max_{i=1,\ldots,n} X_i \quad \text{is} \quad \tau^2 := \max_{i=1,\ldots,n} \mathrm{Var}[X_i]\text{-subgaussian.}$$

b. Show that the mean and median of $\max_i X_i$ satisfy

$$\mathbf{E}\left[\max_{i=1,\ldots,n} X_i\right] \leq \mathrm{med}\left[\max_{i=1,\ldots,n} X_i\right] + \tau\sqrt{2\log 2}$$

Hint: estimate $\mathbf{P}[\max_i X_i \geq \mathbf{E}[\max_i X_i] - t]$ from below for $t = \tau\sqrt{2\log 2}$.

Let $(B, \|\cdot\|_B)$ be a separable Banach space, and let X be a centered Gaussian vector in B (that is, $X \in B$ and $\langle v, X\rangle$ is a Gaussian random variable for every element $v \in B^*$ in the dual space of B). Recall that the norm satisfies

$$\|x\|_B = \sup_{v \in B^*, \|v\| \leq 1} \langle v, x\rangle$$

by duality. Moreover, as B is separable, the supremum in this expression can be restricted to a countable dense subset $V \subset B^*$ (independent of x). Define

$$\sigma^2 := \sup_{v \in B^*, \|v\| \le 1} \mathbf{E}[\langle v, X \rangle^2].$$

c. Show that $\sigma < \infty$, $\mathbf{E}\|X\|_B < \infty$, and that $\|X\|_B$ is σ^2-subgaussian.
Hint: $\mathrm{med}[|\langle v, X \rangle|] \le \mathrm{med}[\|X\|_B] < \infty$ for all $v \in B^*$, $\|v\| \le 1$.

d. Prove the Landau-Shepp-Marcus-Fernique theorem:

$$\mathbf{E}[e^{\alpha \|X\|_B^2}] < \infty \quad \text{if and only if} \quad \alpha < \frac{1}{2\sigma^2}.$$

Hint: for the only if part, use $\mathbf{E}[e^{\alpha \|X\|_B^2}] \ge \mathbf{E}[e^{\alpha \langle v, X \rangle^2}]$ for $v \in B^*$, $\|v\| \le 1$.

3.19 (Bakry-Émery criterion). In Problems 2.12 and 2.13 (we adopt the notation used there), we showed that the Bakry-Émery criterion $c\Gamma_2(f,f) \ge \Gamma(f,f)$ provides an *algebraic criterion* for the validity of the Poincaré inequality. However, the Bakry-Émery criterion is strictly stronger than the validity of a Poincaré inequality. In the present problem, we will show that if the Markov semigroup is reversible and its carré du champ satisfies a chain rule, then the Bakry-Émery criterion even implies validity of the log-Sobolev inequality. This provides a very useful tool for proving log-Sobolev inequalities for certain classes of continuous distributions.

Let P_t be a reversible and ergodic Markov semigroup with stationary measure μ, and assume that the carré du champ satisfies the chain rule

$$\Gamma(f, \phi \circ g) = \Gamma(f,g)\, \phi' \circ g.$$

For example, this is evidently the case when $\Gamma(f,g) = \nabla f \cdot \nabla g$.

a. Show that

$$\begin{aligned} \mathcal{E}(\log P_t f, P_t f) &= \mu(\Gamma(P_t \log P_t f, f)) \\ &\le \mu(\Gamma(f,f)/f)^{1/2} \mu(f \Gamma(P_t \log P_t f, P_t \log P_t f))^{1/2}. \end{aligned}$$

b. Show that the Bakry-Émery criterion $c\Gamma_2(f,f) \ge \Gamma(f,f)$ for all f implies

$$\mathcal{E}(\log P_t f, P_t f) \le e^{-t/c} \mathcal{E}(\log f, f)^{1/2} \mu(f P_t \Gamma(\log P_t f, \log P_t f))^{1/2}.$$

Hint: use Theorem 2.37 and the chain rule.

c. Show that the above inequality implies

$$\mathcal{E}(\log P_t f, P_t f) \le e^{-t/c} \mathcal{E}(\log f, f)^{1/2} \mathcal{E}(\log P_t f, P_t f)^{1/2},$$

and thus the Bakry-Émery criterion implies the log-Sobolev inequality

$$\mathrm{Ent}_\mu[f] \le \frac{c}{2}\, \mathcal{E}(\log f, f) \quad \text{for all } f.$$

d. Let μ be a ρ-uniformly log-concave probability measure on $\mathbb{R}^n$, that is, $\mu(dx) = e^{-W(x)}dx$ where the potential function W satisfies $\nabla\nabla^* W \succeq \rho\,\mathrm{Id}$. Show that μ satisfies the dimension-free log-Sobolev inequality

$$\mathrm{Ent}_\mu[f^2] \le \frac{2}{\rho}\int \|\nabla f\|^2\, d\mu.$$

Hint: see Problem 2.13.

Remark. In the setting of this problem, it is in fact possible after some further work to show that the Bakry-Émery criterion is equivalent to the validity of a *local* log-Sobolev inequality, which strengthens the result of Theorem 2.37 under the chain rule assumption. We omit the details.

3.20 (Bounded perturbations). Let μ be a probability measure for which we have proved a log-Sobolev inequality. Let ν be a "small perturbation" of μ. It is not entirely obvious that ν will also satisfy a log-Sobolev inequality. In this problem, we will show that log-Sobolev and Poincaré inequalities are stable under bounded perturbations, so that we can deduce an inequality for ν from the corresponding inequality for μ. This can be a useful tool to prove log-Sobolev or Poincaré inequalities in cases for which it is not obvious how to proceed by a direct approach (for example, using Theorem 3.20).

Suppose that μ that satisfies the log-Sobolev inequality

$$\mathrm{Ent}_\mu[f] \le c\,\mu(\Gamma(f, \log f)),$$

where we have expressed the right-hand side in terms of a "square gradient" $\Gamma(f, \log f) \ge 0$. For example, if $\mu \sim N(0, I)$, we choose $\Gamma(f, g) = \nabla f \cdot \nabla \log f$. In the setting of Theorem 3.20, if the Markov semigroup is reversible, we can choose $\Gamma(f, \log f)$ to be the carré du champ of Problem 2.7; however, the present result is not specific to the Markov semigroup setting and can be applied to any log-Sobolev type inequality of the above form.

a. Prove the following identity for $\nu \ll \mu$:

$$\mathrm{Ent}_\nu[X] \le \left\|\frac{d\nu}{d\mu}\right\|_\infty \mathrm{Ent}_\mu[X].$$

Hint: use the variational principle of Problem 3.13.

b. Suppose that ν is a bounded perturbation of μ in the sense that $\varepsilon \le \frac{d\nu}{d\mu} \le \delta$ for some $\delta, \varepsilon > 0$. Show that ν satisfies the log-Sobolev inequality

$$\mathrm{Ent}_\nu[f] \le \frac{c\delta}{\varepsilon}\,\nu(\Gamma(f, \log f)).$$

c. Define the probability measure $\nu(dx) = Z^{-1}e^{-V(x)}dx$ on $\mathbb{R}$, where Z is the normalization factor. Suppose that the potential $V(x)$ is sandwiched between two quadratic functions: $x^2 + a \le V(x) \le x^2 + b$ for all $x \in \mathbb{R}$. Show that ν satisfies the log-Sobolev inequality

$$\mathrm{Ent}_\nu[f^2] \le e^{2(b-a)}\nu(|f'|^2).$$

d. We have shown that the log-Sobolev inequality is stable under bounded perturbations. An analogous result holds for Poincaré inequalities. Indeed, suppose that μ that satisfies the Poincaré inequality

$$\mathrm{Var}_\mu[f] \le c\,\mu(\Gamma(f,f)).$$

Show that if $\varepsilon \le \frac{d\nu}{d\mu} \le \delta$, then

$$\mathrm{Var}_\nu[f] \le \frac{c\delta}{\varepsilon}\,\nu(\Gamma(f,f)).$$

Remark. While bounded perturbation results can be useful, the constant δ/ε can be quite large in practice. In particular, it is typically the case that δ/ε will increase exponentially with dimension, so that the bounded perturbation method does not yield satisfactory results when applied in high dimension. However, one can of course apply the bounded perturbation method in one dimension, and then obtain dimension-free results by tensorization.

Notes

§3.1 and §3.2. Much of this material is classical. See, e.g., [13, 26] for a more systematic treatment of subgaussian inequalities and the martingale method. Theorem 3.11 was popularized by McDiarmid [58] for combinatorial problems.

§3.3 and §3.4. Logarithmic Sobolev inequalities were first systematically studied by Gross [42], together with their connection to Markov semigroups. A comprehensive treatment is given in [44] and in [7] (see also [12] where such connections are developed in the discrete setting). The tensorization property of entropy also appears already in [42]; we followed the proof in [50]. The variational formula for entropy plays a fundamental role in large deviations theory [22]. Lemma 3.13 is due to I. Herbst, but was apparently never published by him. The entropy method was systematically applied to the development of concentration inequalities by Ledoux [48, 50]. A comprehensive treatment of the entropy method for concentration inequalities is given in [13]. Problem 3.16 is from [11], while Problem 3.18 follows the approach in [49].

4

Lipschitz concentration and transportation inequalities

In the previous chapters, we have investigated the concentration phenomenon in the following form: the fluctuations of a function $f(X_1, \ldots, X_n)$ of independent (or weakly dependent) random variables are small if the "gradient" of f is small. In this chapter, we will develop a different perspective on the concentration phenomenon. Rather than measuring the sensitivity of the function f in terms of a *gradient*, we will introduce a *metric* viewpoint that emphasizes the role of Lipschitz functions. This complementary perspective will lead us to new methods to investigate and prove concentration, and to new inequalities that do not have a natural description in terms of gradients. In particular, we will prove Talagrand's inequality, which is important in many applications.

4.1 Concentration in metric spaces

Recall a basic definition.

Definition 4.1 (Lipschitz functions). *Let $(\mathbb{X}, d)$ be a metric space. A function $f : \mathbb{X} \to \mathbb{R}$ is called* L-Lipschitz *if $|f(x)-f(y)| \le L\, d(x,y)$ for all $x, y \in \mathbb{X}$. The family of all* 1*-Lipschitz functions is denoted* $\mathrm{Lip}(\mathbb{X})$.

What do Lipschitz functions have to do with concentration? While we have expressed our concentration results to date in terms of gradient bounds, such results can often be interpreted naturally in terms of Lipschitz properties. To make this point, let us begin by considering two examples.

Example 4.2 (Gaussian concentration). Let $X_1, \ldots, X_n$ be i.i.d. $N(0,1)$ random variables. Gaussian concentration (Theorem 3.25) states that the random variable $f(X_1, \ldots, X_n)$ is $\||\nabla f|^2\|_\infty$-subgaussian. However, the quantity $\||\nabla f|^2\|_\infty$ is naturally expressed in terms of a Lipschitz property.

Lemma 4.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a C^1-function. Then $\||\nabla f|^2\|_\infty \le L^2$ if and only if $|f(x) - f(y)| \le L\|x - y\|$ for all $x, y \in \mathbb{R}^n$.*

Proof. Note that the L-Lipschitz property implies

$$v \cdot \nabla f(x) = \lim_{t \to 0} \frac{f(x+tv) - f(x)}{t} \le L\|v\|.$$

Optimizing over $\|v\| \le 1$ and x yields $\|\|\nabla f\|^2\|_\infty \le L^2$. Conversely,

$$f(x) - f(y) = \int_0^1 \frac{d}{dt} f(tx + (1-t)y)\, dt = \int_0^1 (x-y) \cdot \nabla f(tx + (1-t)y)\, dt$$

by the fundamental theorem of calculus. It therefore follows readily that if $\|\|\nabla f\|^2\|_\infty \le L^2$, then $f(x) - f(y) \le L\|x-y\|$ for all $x, y \in \mathbb{R}^n$. □

In view of this lemma, it follows immediately[1] that Gaussian concentration can be equivalently phrased in terms of Lipschitz functions: *if $X \sim N(0, I)$, then $f(X)$ is 1-subgaussian for every $f \in \mathrm{Lip}(\mathbb{R}^n, \|\cdot\|)$.*

As a second example, let us revisit McDiarmid's inequality.

Example 4.4 (McDiarmid's inequality). Let $X_1, \ldots, X_n$ be independent random variables, where X_i takes values in some measurable space $\mathbb{X}_i$ for $i = 1, \ldots, n$. McDiarmid's inequality (Theorem 3.11) states that the random variable $f(X_1, \ldots, X_n)$ is $\frac{1}{4}\sum_{k=1}^n \|D_k f\|_\infty^2$-subgaussian. Also this inequality can be phrased in terms of a Lipschitz property. To this end, let us introduce the *weighted Hamming distance* $d_c(x, y)$ on $\mathbb{X}_1 \times \cdots \times \mathbb{X}_n$ as

$$d_c(x, y) := \sum_{i=1}^n c_i \mathbf{1}_{x_i \neq y_i}.$$

Lemma 4.5. *Let $f : \mathbb{X}_1 \times \cdots \times \mathbb{X}_n \to \mathbb{R}$. Then $\|D_i f\|_\infty \le c_i$ for all $i = 1, \ldots, n$ if and only if $|f(x) - f(y)| \le d_c(x, y)$ for all $x, y \in \mathbb{X}_1 \times \cdots \times \mathbb{X}_n$.*

Proof. Suppose that f is 1-Lipschitz with respect to d_c. If x, y only differ in the ith coordinate, it follows that $|f(x) - f(y)| \le c_i$. In particular, we conclude that that $\|D_i f\|_\infty \le c_i$ for all i. Conversely, consider the telescoping sum

$$f(x) - f(y) = \sum_{i=1}^n \{f(x_1, \ldots, x_i, y_{i+1}, \ldots, y_n) - f(x_1, \ldots, x_{i-1}, y_i, \ldots, y_n)\}.$$

As the ith term in the sum is the difference between f evaluated at two points that differ only in the ith coordinate, it is bounded by $\|D_i f\|_\infty \mathbf{1}_{x_i \neq y_i}$. Thus if $\|D_i f\|_\infty \le c_i$ for all i, then f is 1-Lipschitz with respect to d_c. □

In view of this simple observation, we obtain the following equivalent formulation of McDiarmid's inequality: *if X is a random vector with independent entries, $f(X)$ is $\frac{1}{4}\|c\|^2$-subgaussian for every $f \in \mathrm{Lip}(\mathbb{X}_1 \times \cdots \times \mathbb{X}_n, d_c)$.*

[1] The claim holds even when f is not C^1 by a simple approximation argument: any Lipschitz function can be approximated uniformly by a smooth Lipschitz function by convolving with a smooth density. The details are left an an exercise.

At an informal level, we have introduced the general concentration principle by stating that a function $f(X_1, \ldots, X_n)$ of independent or weakly dependent random variables is close to its mean if the function f is "not too sensitive" to any of its coordinates. Gradient bounds and Lipschitz properties provide two different ways of making the informal notion of "not too sensitive" precise. In the case of gradient bounds, the sensitivity of the function f is measured locally, while the Lipschitz property quantifies the sensitivity in a global manner. These two points of view are very similar in spirit, however, and are often even equivalent as we have seen above in the case of the Gaussian concentration inequality and McDiarmid's inequality.

Nonetheless, it will prove to be extremely useful to reconsider the concentration principle from the metric perspective. The reasons for this are twofold:

- While in some cases gradient bounds and Lipschitz properties can be shown to be equivalent, there are other cases in which these two notions are distinct. For example, the one-sided difference bound $\| \sum_{i=1}^n |D_i^- f|^2 \|_\infty \le L^2$ is not naturally formulated in terms of a Lipschitz property with respect to some metric. Conversely, there are important Lipschitz-type properties that cannot be naturally formulated in terms of a gradient; we will encounter such a property when we develop Talagrand's concentration inequalities later in this chapter. Thus the complementary viewpoints provided by gradient and metric notions of concentration give rise to genuinely different results that can be of substantial importance in different settings.

- Our emphasis on gradients in the previous chapters was intimately tied to a class of inequalities—Poincaré and log-Sobolev inequalities—that are of fundamental importance in proving and understanding concentration properties. The metric perspective, however, will require us to develop new types of inequalities that exploit the metric structure of the problem. The development of these ideas will significantly enhance our understanding of the concentration principle and will provide us with new tools to prove concentration inequalities that are not easily obtained by other methods.

Having roughly motivated the metric perspective on concentration, we are ready to take some first steps towards a general theory.

We have shown above that Gaussian concentration can be phrased as follows: if $X \sim N(0, I)$, then $f(X)$ is 1-subgaussian for every $f \in \mathrm{Lip}(\mathbb{R}^n, \|\cdot\|)$. Similarly, McDiarmid's inequality states that if X is a random vector with independent entries, $f(X)$ is $\frac{1}{4}\|c\|^2$-subgaussian for $f \in \mathrm{Lip}(\mathbb{X}_1 \times \cdots \times \mathbb{X}_n, d_c)$. Motivated by these examples, we can pose the following basic question.

> *For which probability measures μ on the metric space $(\mathbb{X}, d)$ is it true that if $X \sim \mu$, then $f(X)$ is σ^2-subgaussian for every $f \in \mathrm{Lip}(\mathbb{X})$?*

We presently give a very general answer to this question in terms of a new class of inequalities that will play a fundamental role throughout this chapter.

Definition 4.6 (Wasserstein distance). *The* Wasserstein distance *between probability measures $\mu, \nu \in \mathcal{P}_1(\mathbb{X}) := \{\rho : \int d(x, \cdot)\rho(dx) < \infty\}$ is defined as*[2]

$$W_1(\mu, \nu) := \sup_{f \in \mathrm{Lip}(\mathbb{X})} \left| \int f \, d\mu - \int f \, d\nu \right|.$$

Definition 4.7 (Relative entropy). *The* relative entropy *between probability measures ν and μ on any measurable space is defined as*

$$D(\nu||\mu) := \begin{cases} \mathrm{Ent}_\mu \left[\frac{d\nu}{d\mu} \right] & \text{if } \nu \ll \mu, \\ \infty & \text{otherwise.} \end{cases}$$

Theorem 4.8 (Bobkov-Götze). *Let $\mu \in \mathcal{P}_1(\mathbb{X})$ be a probability measure on a metric space $(\mathbb{X}, d)$. Then the following are equivalent for $X \sim \mu$:*

1. *$f(X)$ is σ^2-subgaussian for every $f \in \mathrm{Lip}(\mathbb{X})$.*
2. *$W_1(\nu, \mu) \le \sqrt{2\sigma^2 D(\nu||\mu)}$ for all ν.*

How should we interpret these concepts? Both the Wasserstein distance and the relative entropy define a form of distance between probability measures. The Wasserstein distance defines a metric in terms of expectations of Lipschitz functions. Relative entropy, on the other hand, is not a metric: it is not even symmetric and does not satisfy a triangle inequality. Nonetheless, it is a natural measure of "closeness" between probability measures (for example, $D(\nu||\mu) \ge 0$ and $D(\nu||\mu) = 0$ if and only of $\mu = \nu$). As we will see in the proof of Theorem 4.8, relative entropy should be viewed as controlling moment generating functions in a suitable sense. As these two notions of distance are of an entirely different nature, there is no *a priori* reason why relative entropy and Wasserstein distance to a given measure μ should be comparable, and this is indeed not necessarily true for arbitrary μ. Theorem 4.8 states that relative entropy and Wasserstein distance are comparable precisely when one can control the moment generating functions of Lipschitz functions. Inequalities such as $W_1(\nu, \mu) \le \sqrt{2\sigma^2 D(\nu||\mu)}$ therefore play a role in the "metric" setting analogous to log-Sobolev inequalities in the "gradient" setting. We can informally view this inequality as a type of dual to the log-Sobolev inequality that is stated in terms of measures rather than functions (cf. Problem 4.1 below).

Before we turn to the proof of Theorem 4.8, let us illustrate how it can be used to prove a well-known inequality for relative entropy.

Example 4.9 (Pinsker's inequality). Let $d(x, y) := \mathbf{1}_{x \ne y}$ be the *trivial metric*. Then $f \in \mathrm{Lip}(\mathbb{X})$ if and only if $\sup f - \inf f \le 1$. Thus the Wasserstein distance in this case is none other than the total variation distance

$$W_1(\mu, \nu) = \sup_{0 \le f \le 1} \left| \int f \, d\mu - \int f \, d\nu \right| =: \|\mu - \nu\|_{\mathrm{TV}}$$

[2] Note that $\rho \in \mathcal{P}_1(\mathbb{X})$ if and only if $\int f \, d\rho < \infty$ for every $f \in \mathrm{Lip}(\mathbb{X})$.

(note that the quantity inside the supremum is invariant under adding a constant to f, so there is no loss in restricting to $0 \le f \le 1$ only).

Now recall from Hoeffding's Lemma 3.6 that $f(X)$ is $\frac{1}{4}\{\sup f - \inf f\}$-subgaussian for every f and μ. Thus Theorem 4.8 implies that

$$\|\mu - \nu\|_{\mathrm{TV}} \le \sqrt{\tfrac{1}{2} D(\nu||\mu)}$$

for every μ, ν. This extremely useful result is known as *Pinsker's inequality* (which also provides additional intuition for the fact that $D(\nu||\mu)$ can be viewed as a form of "closeness" between probability measures). Of course, we could have also gone in the converse direction: if we had given an independent proof of Pinsker's inequality (there are numerous such proofs), then we could have used Theorem 4.8 to provide an alternative proof of Hoeffding's lemma.

Let us now turn to the proof of Theorem 4.8. The key insight that is needed is that relative entropy is intimately related to moment generating functions; once this has been understood, the remainder of the proof of Theorem 4.8 is essentially trivial. The following result, which dates back to the earliest history of statistical mechanics, makes this idea precise.

Lemma 4.10 (Gibbs variational principle).

$$\log \mathbf{E}_\mu[e^f] = \sup_\nu \{\mathbf{E}_\nu[f] - D(\nu\|\mu)\}.$$

Proof. We may assume f is bounded above to avoid integrability problems (if not, apply the result to $f \wedge M$ and then take the supremum over M). Define

$$d\tilde{\mu} = \frac{e^f d\mu}{\mathbf{E}_\mu[e^f]}.$$

We have for $D(\nu||\mu) < \infty$

$$\begin{aligned}
\log \mathbf{E}_\mu[e^f] - D(\nu\|\tilde{\mu}) &= \log \mathbf{E}_\mu[e^f] - \int \left(\log \frac{d\nu}{d\tilde{\mu}}\right) d\nu \\
&= \log \mathbf{E}_\mu[e^f] - \int \left(\log \frac{d\nu}{d\mu}\right) d\nu + \int \left(\log \frac{d\tilde{\mu}}{d\mu}\right) d\nu \\
&= \mathbf{E}_\nu[f] - D(\nu\|\mu).
\end{aligned}$$

Taking the supremum over ν on both sides yields the result. □

Remark 4.11. Note that Lemma 3.15 can be reformulated as

$$D(\nu\|\mu) = \sup\{\mathbf{E}_\nu[f] : \mathbf{E}_\mu[e^f] = 1\} = \sup\{\mathbf{E}_\nu[f] - \log \mathbf{E}_\mu[e^f]\},$$

where the sup is taken over functions f. Thus Lemma 4.10 is precisely the dual convex optimization problem to the variational formula for entropy.

We can now complete the proof of Theorem 4.8.

Proof (Theorem 4.8). By definition, the property 1 can be stated as

$$\log \mathbf{E}_\mu[e^{\lambda\{f-\mathbf{E}_\mu f\}}] \le \frac{\lambda^2\sigma^2}{2} \quad \text{for all } \lambda \in \mathbb{R},\ f \in \mathrm{Lip}(\mathbb{X}).$$

By Lemma 4.10, this is equivalent to

$$\sup_{\lambda\in\mathbb{R}} \sup_{f\in\mathrm{Lip}(\mathbb{X})} \sup_{\nu} \left\{ \lambda\{\mathbf{E}_\nu f - \mathbf{E}_\mu f\} - D(\nu\|\mu) - \frac{\lambda^2\sigma^2}{2} \right\} \le 0.$$

Exchanging the order of the suprema and evaluating explicitly the suprema over f and λ yields that the above expression is equivalent to

$$\sup_{\nu} \left\{ \frac{W_1(\mu,\nu)^2}{2\sigma^2} - D(\nu\|\mu) \right\} \le 0,$$

which is evidently an immediate reformulation of property 2. □

Theorem 4.8 characterizes the subgaussian property of Lipschitz functions on an arbitrary but *fixed* metric space $(\mathbb{X}, d)$. It is important to emphasize that this is not in itself a "high-dimensional" result. As in the previous chapters, the crucial idea that will be needed to work in high dimension is a tensorization principle. In the following section, we will develop a different perspective on the inequality $W_1(\mu,\nu) \le \sqrt{2\sigma^2 D(\nu\|\mu)}$ that will enable us to prove such a tensorization principle. This will provide us with a powerful tool to develop and understand dimension-free Lipschitz concentration inequalities.

Problems

4.1 (Discrete log-Sobolev and Lipschitz concentration). One simple way to gain some insight into the inequality $W_1(\nu,\mu) \le \sqrt{2\sigma^2 D(\nu\|\mu)}$ is to note that it implies a sort of "dual" form of the discrete log-Sobolev inequality $\mathrm{Ent}[e^{\lambda f}] \le \mathrm{Cov}[\lambda f, e^{\lambda f}]$ of Lemma 3.16 for Lipschitz functions.

a. Show that $W_1(\nu,\mu) \le \sqrt{2\sigma^2 D(\nu\|\mu)}$ implies the inequality

$$\mathrm{Cov}[\lambda f, e^{\lambda f}]^2 \le 2\lambda^2\sigma^2 \mathrm{Ent}[e^{\lambda f}]\mathbf{E}[e^{\lambda f}] \quad \text{for } \lambda \in \mathbb{R},\ f \in \mathrm{Lip}(\mathbb{X}).$$

Hint: consider $d\nu = e^{\lambda f}\, d\mu/\mathbf{E}_\mu[e^{\lambda f}]$.

b. Use the above inequality together with the discrete log-Sobolev inequality of Lemma 3.16 to prove that $W_1(\nu,\mu) \le \sqrt{2\sigma^2 D(\nu\|\mu)}$ implies that $f(X)$ is $4\sigma^2$-subgaussian for $X \sim \mu$, $f \in \mathrm{Lip}(\mathbb{X})$ (which agrees precisely with the result of Theorem 4.8 up to the suboptimal constant 4).

4.2 (Isoperimetric inequalities and concentration). There is an entirely different approach to investigating Lipschitz concentration properties that played an important role in the historical development of this area: the isoperimetric method. While we have avoided using this approach in this course, the method remains of fundamental importance in the development and understanding of new concentration phenomena. The goal of this problem is to develop some basic ideas surrounding this approach.

Let $(\mathbb{X}, d)$ be a metric space. The idea behind the isperimetric method is not to investigate the tail behavior of functions directly, but rather to focus attention to the probabilities of sets. For any measurable set $A \subseteq \mathbb{X}$, define its ε-*fattening* as $A^\varepsilon := \{x \in \mathbb{X} : d(x, A) \le \varepsilon\}$. A statement of the form

$$\mu(A^\varepsilon) \ge 1 - Ce^{-\varepsilon^2/2\sigma^2} \quad \text{for all } \varepsilon \ge 0,\ A \subseteq \mathbb{X} \text{ such that } \mu(A) \ge \tfrac{1}{2}$$

is called an *isoperimetric inequality*. It states that almost every point in $\mathbb{X}$ is ε-close to a set of measure $\frac{1}{2}$. One way to interpret this result is geometrically: given any set A with $\mu(A) = \frac{1}{2}$, the measure of its ε-boundary is $\mu(A^\varepsilon \backslash A) \approx \frac{1}{2}$; thus the boundary of the set contains almost as much mass as the interior of the set. Mathematical phenomena relating the size of a set to the size of its boundary are generally referred to as "isoperimetric problems."

a. Suppose that the measure μ satisfies the above isoperimetric inequality. Show that we have the concentration inequality

$$\mathbf{P}_\mu[f - \mathrm{med}(f) \ge t] \le Ce^{-t^2/2\sigma^2} \quad \text{for all } t \ge 0,\ f \in \mathrm{Lip}(\mathbb{X}).$$

Hint: consider the set $A = \{f \le \mathrm{med}(f)\}$. Here $\mathrm{med}(f)$ denotes the median.

b. Conversely, show that the above isoperimetric inequality is implied by

$$\mathbf{P}_\mu[f - \mathrm{med}(f) \ge t] \le Ce^{-t^2/2\sigma^2} \quad \text{for all } t \ge 0,\ f \in \mathrm{Lip}(\mathbb{X}).$$

Hint: consider $f(x) = d(x, A)$.

We have discovered the elementary fact that isoperimetric inequalities are equivalent to tail bounds for Lipschitz functions. However, unlike most of our previous results this course, the deviation here is from the *median* rather than from the *mean*. It turns out that deviation inequalities from the median and the mean are equivalent, however, up to constants. Whether deviation from the median or the mean is more useful depends on the application (see Problem 3.18 for a situation where the median provides useful insight).

c. Suppose that the above isoperimetric inequality holds. Show that

$$\mathrm{med}(f) \le \mathbf{E}_\mu f + C\sigma\sqrt{\pi/2}$$

for all $f \in \mathrm{Lip}(\mathbb{X})$, and conclude that

$$\mathbf{P}_\mu[f - \mathbf{E}_\mu f \ge t] \le e^{C^2\pi/16} e^{-t^2/8\sigma^2} \quad \text{for all } t \ge 0.$$

Hint: estimate $\mathbf{E}_\mu[(\mathrm{med}(f) - f)_+]$ by integrating the tail bound.

d. Conversely, suppose that for $f \in \mathrm{Lip}(\mathbb{X})$

$$\mathbf{P}_\mu[f - \mathbf{E}_\mu f \geq t] \leq Ce^{-t^2/2\sigma^2} \quad \text{for all } t \geq 0.$$

Show that this implies

$$\mathbf{E}_\mu f \leq \mathrm{med}(f) + \sigma\sqrt{2\log 2C}$$

for all $f \in \mathrm{Lip}(\mathbb{X})$, and conclude that

$$\mathbf{P}_\mu[f - \mathrm{med}(f) \geq t] \leq \max\{C, (2C)^{1/4}\}e^{-t^2/8\sigma^2} \quad \text{for all } t \geq 0.$$

Hint: see Problem 3.18.

Finally, we develop a direct connection between Theorem 4.8 and isoperimetry.

e. Suppose that $W_1(\nu,\mu) \leq \sqrt{2\sigma^2 D(\nu||\mu)}$ for all ν. Argue that

$$d(A,B) \leq W_1(\mu(\cdot|A),\mu(\cdot|B)) \leq \sqrt{2\sigma^2\log(1/\mu(A))} + \sqrt{2\sigma^2\log(1/\mu(B))}$$

for any disjoint sets $A, B \subseteq \mathbb{X}$.

f. Applying the above result to $B = \mathbb{X}\backslash A^\varepsilon$, argue that

$$\mu(A^\varepsilon) \geq 1 - 2e^{-\varepsilon^2/8\sigma^2} \quad \text{for all } \varepsilon \geq 0,\ A \subseteq \mathbb{X} \text{ such that } \mu(A) \geq \tfrac{1}{2}.$$

Thus $W_1(\nu,\mu) \leq \sqrt{2\sigma^2 D(\nu||\mu)}$ yields directly an isoperimetric inequality.

4.2 Transportation inequalities and tensorization

In the previous section, we have introduced the fundamental inequality $W_1(\nu,\mu) \leq \sqrt{2\sigma^2 D(\nu||\mu)}$ as a characterization of the Lipschitz concentration property on a *fixed* metric space. However, for this result to be useful in high dimension, we must understand whether it is possible to tensorize inequalities of this type. It turns out that there is indeed a tensorization principle that is extremely useful in this setting, but this is far from obvious from the formulation developed in the previous section. In order to develop this idea, it will prove to be necessary to formulate these inequalities in a different manner in terms of *optimal transportation*. We will presently develop this connection, and the tensorization principle that follows from it.

Optimal transportation is concerned with the classical notion of *coupling*. Recall that a coupling of probability measures of μ, ν is any joint distribution of random variables (X, Y) with marginal distributions $X \sim \mu$ and $Y \sim \nu$. Of course, there exist many different couplings for given μ, ν.

Definition 4.12 (Coupling). *Let μ, ν be two probability measures, and let*

$$\mathcal{C}(\mu,\nu) := \{\mathrm{Law}(X,Y) : X \sim \mu,\ Y \sim \nu\}.$$

Any probability measure $\mathbf{M} \in \mathcal{C}(\mu,\nu)$ is called a coupling *of μ, ν.*

Let $f \in \mathrm{Lip}(\mathbb{X})$. Then for any $\mathbf{M} \in \mathcal{C}(\mu, \nu)$, we have

$$|\mathbf{E}_\mu f - \mathbf{E}_\nu f| = |\mathbf{E}_{\mathbf{M}}[f(X) - f(Y)]| \le \mathbf{E}_{\mathbf{M}}[d(X, Y)].$$

In particular, we obtain the elementary inequality

$$W_1(\mu, \nu) \le \inf_{\mathbf{M} \in \mathcal{C}(\mu,\nu)} \mathbf{E}_{\mathbf{M}}[d(X, Y)].$$

That is, the Wasserstein distance is controlled by the smallest expected distance between random variables X, Y such that $X \sim \mu$ and $Y \sim \mu$. The latter optimization over couplings is called an *optimal transportation problem*. The name derives not from viewing μ, ν as probabilities but rather as distributions of mass, for example, in a sandpile: the optimal transportation problem tells us how to transform one sandpile into another sandpile in a manner that minimizes the total distance we need to transport the grains of sand.

Remarkably, it turns out that nothing is lost in estimating the Wasserstein distance by an optimal transportation cost, under mild technical conditions. This is the statement of the following classical result.

Theorem 4.13 (Monge-Kantorovich duality). *We have*

$$W_1(\mu, \nu) = \sup_{f \in \mathrm{Lip}(\mathbb{X})} |\mathbf{E}_\mu f - \mathbf{E}_\nu f| = \inf_{\mathbf{M} \in \mathcal{C}(\mu,\nu)} \mathbf{E}_{\mathbf{M}}[d(X, Y)]$$

for all probability measures $\mu, \nu \in \mathcal{P}_1(\mathbb{X})$ on a separable metric space $(\mathbb{X}, d)$.

To avoid getting distracted by technicalities, we will prove Theorem 4.13 here in the discrete setting. The full intuition arises here, and the extension to the continuous case is an exercise in approximation (Problem 4.3).

Proof (Discrete case). Let μ, ν be probabilities on the finite set $\mathbb{X} = \{1, \ldots, p\}$. The optimal transportation problem can evidently be phrased as follows:

$$\begin{aligned}
\underset{M}{\text{Minimize:}} \quad & \sum_{i,j=1}^{p} d(i,j) M(i,j) \\
\text{Subject to:} \quad & M(i,j) \ge 0,\ 1 \le i, j \le p \\
& \sum_{j=1}^{p} M(i,j) = \mu(i),\ 1 \le i \le p \\
& \sum_{i=1}^{p} M(i,j) = \nu(j),\ 1 \le j \le p
\end{aligned}$$

This is nothing other than a standard linear programming problem. The dual linear programming problem corresponding to this primary problem is

$$\begin{aligned}
\underset{f,g}{\text{Maximize:}} \quad & \sum_{i=1}^{p} f(i)\mu(i) + \sum_{i=1}^{p} g(j)\nu(j) \\
\text{Subject to:} \quad & f(i) + g(j) \le d(i,j),\ 1 \le i, j \le p.
\end{aligned}$$

By the strong duality theorem of linear programming, the optimal values of these two optimization problems coincide, so we have proved

$$\inf_{\mathbf{M}\in\mathcal{C}(\mu,\nu)} \mathbf{E}_{\mathbf{M}}[d(X,Y)] = \sup\{\mathbf{E}_\mu f + \mathbf{E}_\nu g : f(x)+g(y)\le d(x,y)\ \forall\, x,y\} =: (*).$$

We must now show that the expression $(*)$ on the right-hand side coincides with the Wasserstein distance. Here we need to use the fact that d is a metric (so far, we only used that d is a nonnegative weight function!) To this end, note that f, g satisfy $f(x)+g(y)\le d(x,y)$ for all x, y if and only if

$$f(x)\le \tilde f(x) := \inf_z\{d(x,z)-g(z)\}\le -g(x) \quad \text{for all } x.$$

Moreover, $\tilde f\in \mathrm{Lip}(\mathbb{X})$ as

$$\tilde f(x)-\tilde f(y)\le \sup_z\{d(x,z)-d(y,z)\}\le d(x,y).$$

It follows immediately that

$$\mathbf{E}_\mu f+\mathbf{E}_\nu g\le \mathbf{E}_\mu\tilde f-\mathbf{E}_\nu\tilde f\le W_1(\mu,\nu)$$

whenever $f(x)+g(y)\le d(x,y)$ for all x, y. Thus we have shown $(*)\le W_1(\mu,\nu)$, while $(*)\ge W_1(\mu,\nu)$ holds trivially (restrict the supremum to $g=-f$). □

The separability assumption of Theorem 4.13 is not entirely innocuous. For example, the trivial metric $d(x,y)=\mathbf{1}_{x\neq y}$ considered in Example 4.9 is not separable (unless $\mathbb{X}$ is discrete), yet Monge-Kantorovich duality still holds in this case. As this is both an important example and an interesting illustration, let us provide here a direct proof of Monge-Kantorovich duality for the trivial metric. It is in fact possible to obtain a more general version of Theorem 4.13 that contains both separable metrics and the trivial metric as special cases, but this will not be needed for our purposes.

Example 4.14 (Total variation). Let $d(x,y)=\mathbf{1}_{x\neq y}$ be the trivial metric. We have seen in Example 4.9 that in this case the Wasserstein distance coincides with the total variation distance, so that Monge-Kantorovich duality reads

$$\|\mu-\nu\|_{\mathrm{TV}} = \inf_{\mathbf{M}\in\mathcal{C}(\mu,\nu)} \mathbf{M}[X\neq Y].$$

That is, the total variation distance between μ, ν is the minimal probability that random variables $X\sim\mu$ and $Y\sim\nu$ do not coincide. We will presently give a direct proof of this fundamental result. As

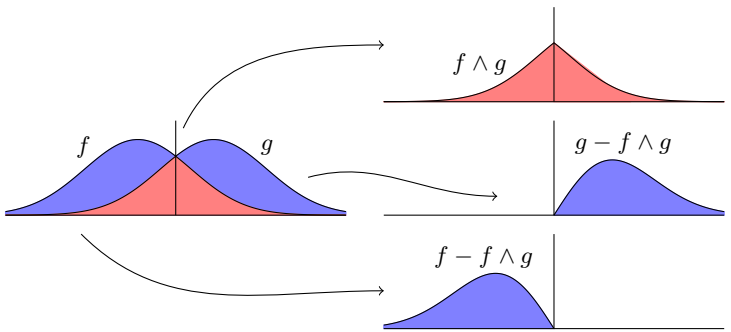
$$\|\mu-\nu\|_{\mathrm{TV}} = \sup_{0\le f\le 1} |\mathbf{E}_{\mathbf{M}}[f(X)-f(Y)]| \le \mathbf{M}[X\neq Y]$$

holds trivially for every $\mathbf{M}\in\mathcal{C}(\mu,\nu)$, it suffices to construct an optimal coupling that attains equality (in contrast, Theorem 4.13 is not constructive).

To construct an optimal coupling, let us assume that we can write $d\mu = f d\rho$ and $d\nu = g d\rho$ for some reference measure ρ and densities f, g (this entails no loss of generality, as we can always choose $\rho = \mu + \nu$). The idea is now to decompose μ and ν into a "common part" and "disjoint parts." We can then construct a coupling by letting either $X = Y$ be drawn from the common part, or drawing X and Y independently from the disjoint parts, with the probabilities chosen appropriately so that this is a coupling. To be precise, let us define the "common part" η and the "disjoint parts" $\tilde\mu, \tilde\nu$ as

$$d\eta := \{f \wedge g\} d\rho, \quad d\tilde\mu := \{f - f \wedge g\} d\rho, \quad d\tilde\nu := \{g - f \wedge g\} d\rho.$$

Then $\eta, \tilde\mu, \tilde\nu$ are all positive measures, $\mu = \tilde\mu + \eta$, $\nu = \tilde\nu + \eta$, and $\tilde\mu, \tilde\nu$ have disjoint supports. This construction is illustrated in the following figure:



We now define the probability measure $\mathbf{M}$ as

$$\mathbf{M}(dx, dy) = \eta(dx)\, \delta_x(dy) + \frac{\tilde\mu(dx)\, \tilde\nu(dy)}{1 - \eta(\mathbb{X})}$$

(here δ_x denotes the point mass at x). It is readily verified that $\mathbf{M} \in \mathcal{C}(\mu, \nu)$ by construction. Moreover, as $\tilde\mu, \tilde\nu$ have disjoint supports, we have

$$\mathbf{M}[X \neq Y] = 1 - \eta(\mathbb{X}) = \int \{f - f \wedge g\}\, d\rho.$$

But note that

$$\int \{f - f \wedge g\}\, d\rho = \int (f - g)_+\, d\rho = \sup_{0 \le h \le 1} \int h \{f - g\}\, d\rho = \|\mu - \nu\|_{\mathrm{TV}}.$$

Thus we have constructed an optimal coupling that attains the infimum in the Monge-Kantorovich duality formula for total variation distance.

We now conclude our detour through the optimal transportation problem and return to the investigation of concentration. By virtue of Monge-Kantorovich duality, it evidently follows from Theorem 4.8 that $f(X)$ is σ^2-subgaussian for every $f \in \mathrm{Lip}(\mathbb{X})$ and $X \sim \mu$ if and only if

$$W_1(\mu,\nu) = \inf_{\mathbf{M}\in\mathcal{C}(\mu,\nu)} \mathbf{E}_{\mathbf{M}}[d(X,Y)] \le \sqrt{2\sigma^2 D(\nu||\mu)} \quad \text{for all } \nu.$$

Inequalities of this type are called *transportation cost inequalities*. While we have previously formulated them without any reference to transportation, it turns out that the formulation in terms of optimal transportation is crucial in order to develop a suitable tensorization principle. This is our next goal.

How might we expect Lipschitz concentration to tensorize? It is not even entirely clear what is meant. Let μ_i be a probability measure on $(\mathbb{X}_i, d_i)$ for $i = 1,\dots,n$, such that each μ_i satisfies the transportation cost inequality

$$W_1(\nu,\mu_i) \le \sqrt{2\sigma^2 D(\nu||\mu_i)} \quad \text{for all } \nu.$$

We would like to deduce that the product measure $\mu_1\otimes\cdots\otimes\mu_n$ on $\mathbb{X}_1\times\cdots\times\mathbb{X}_n$ satisfies a Lipschitz concentration property, that is, that

$$W_1(\nu,\mu_1\otimes\cdots\otimes\mu_n) \le \sqrt{2\sigma^2 D(\nu||\mu_1\otimes\cdots\otimes\mu_n)} \quad \text{for all } \nu.$$

However, to even make sense of this statement, we must first specify a metric d on $\mathbb{X}_1\times\cdots\times\mathbb{X}_n$. For example, one might be interested in working with the ℓ_1-metric $d(x,y) = d_1(x_1,y_1)+\cdots+d(x_n,y_n)$, or with the ℓ_2-metric $d(x,y) = \{d_1(x_1,y_1)^2+\cdots+d_n(x_n,y_n)^2\}^{1/2}$, or with any other suitable combination. Ultimately, however, the appropriate choice of metric will be dictated by whether we are able to prove a tensorization principle. As will become clear in the sequel, we can prove different forms of tensorization in product spaces (i.e., for different definitions of the metric d) by using different types of transportation cost inequalities. It is therefore fruitful, rather than considering one specific setting, to prove a tensorization principle for a rather general class of transportation cost inequalities. The following theorem does precisely that. Once its power has been understood, it will be straightforward to interpret the behavior of different transportation cost inequalities in high dimension.

Theorem 4.15 (Marton). *Let $\varphi:\mathbb{R}_+\to\mathbb{R}_+$ be a convex function, and let $w_i:\mathbb{X}_i\times\mathbb{X}_i\to\mathbb{R}_+$ be positive weight function. Suppose that for $i=1,\dots,n$*

$$\inf_{\mathbf{M}\in\mathcal{C}(\mu_i,\nu)} \varphi(\mathbf{E}_{\mathbf{M}}[w_i(X,Y)]) \le 2\sigma^2 D(\nu||\mu_i) \quad \textit{for all } \nu.$$

Then we have

$$\inf_{\mathbf{M}\in\mathcal{C}(\mu_1\otimes\cdots\otimes\mu_n,\nu)} \sum_{i=1}^n \varphi(\mathbf{E}_{\mathbf{M}}[w_i(X_i,Y_i)]) \le 2\sigma^2 D(\nu||\mu_1\otimes\cdots\otimes\mu_n) \quad \textit{for all } \nu.$$

The transportation cost inequality $W_1(\mu_i,\nu) \le \sqrt{2\sigma^2 D(\nu||\mu_i)}$ corresponds to the assumption of Theorem 4.15 with $\varphi(x) = x^2$ and $w_i(x,y) = d_i(x,y)$. However, the quantity on the left-hand side of the "tensorized" inequality

$$\left[\inf_{\mathbf{M}\in\mathcal{C}(\mu_1\otimes\cdots\otimes\mu_n,\nu)} \sum_{i=1}^n \mathbf{E}_{\mathbf{M}}[d_i(X_i,Y_i)]^2\right]^{1/2} \le \sqrt{2\sigma^2 D(\nu||\mu_1\otimes\cdots\otimes\mu_n)}$$

is not itself a Wasserstein distance. We must therefore take an extra step to use this general tensorization principle. For example, if we define

$$d_c(x,y) := \sum_{i=1}^n c_i d_i(x_i, y_i),$$

the *weighted ℓ_1-metric* on $\mathbb{X}_1 \times \cdots \times \mathbb{X}_n$, we obtain the following.

Corollary 4.16. *Suppose that the transportation cost inequality*

$$W_1(\mu_i, \nu) \le \sqrt{2\sigma^2 D(\nu||\mu_i)} \quad \text{for all } \nu$$

holds for μ_i on $(\mathbb{X}_i, d_i)$ for $i = 1, \ldots, n$. Then the transportation cost inequality

$$W_1(\mu_1 \otimes \cdots \otimes \mu_n, \nu) \le \sqrt{2\sigma^2 D(\nu||\mu_1 \otimes \cdots \otimes \mu_n)} \quad \text{for all } \nu$$

holds for $\mu_1 \otimes \cdots \otimes \mu_n$ on $(\mathbb{X}_1 \times \cdots \mathbb{X}_n, d_c)$ whenever $\sum_{i=1}^n c_i^2 = 1$.

Proof. For probability measures ν, ρ on $(\mathbb{X}_1 \times \cdots \mathbb{X}_n, d_c)$, we have

$$W_1(\nu,\rho) = \inf_{\mathbf{M} \in \mathcal{C}(\nu,\rho)} \sum_{i=1}^n c_i \mathbf{E}_{\mathbf{M}}[d_i(X_i, Y_i)] \le \left[\inf_{\mathbf{M} \in \mathcal{C}(\nu,\rho)} \sum_{i=1}^n \mathbf{E}_{\mathbf{M}}[d_i(X_i, Y_i)]^2 \right]^{1/2}$$

by the Cauchy-Schwarz inequality (as $\sum_{i=1}^n c_i^2 = 1$). The result now follows from Theorem 4.15 with $\varphi(x) = x^2$ and $w_i(x,y) = d_i(x,y)$. □

Corollary 4.16 yields immediately another proof of McDiarmid's inequality.

Example 4.17 (McDiarmid's inequality). The trivial metric $d_i(x,y) = \mathbf{1}_{x \ne y}$ on $\mathbb{X}_i$ satisfies the transportation cost inequality $W_1(\mu,\nu) \le \{\frac{1}{2} D(\nu||\mu)\}^{1/2}$ by Pinsker's inequality (Example 4.9). Therefore, by Corollary 4.16, we have

$$W_1(\mu_1 \otimes \cdots \otimes \mu_n, \nu) \le \sqrt{\tfrac{1}{2} D(\nu||\mu_1 \otimes \cdots \otimes \mu_n)}$$

on $\mathbb{X}_1 \times \cdots \times \mathbb{X}_n$ with respect to the weighted Hamming distance $d_c(x,y) = \sum_{i=1}^n c_i \mathbf{1}_{x_i \ne y_i}$. Thus Theorem 4.8 yields precisely the Lipschitz formulation of McDiarmid's inequality discussed in Example 4.4.

By using the Cauchy-Schwarz inequality as in Corollary 4.16, the tensorization principle of Theorem 4.15 yields dimension-free concentration inequalities in terms of weighted ℓ_1-metrics. In the next section, we will use a more refined version of the argument that led to the transportation proof of McDiarmid's inequality to prove Talagrand's concentration inequality, which is a crucial improvement over McDiarmid's inequality in terms of "one-sided differences" that makes it possible to obtain lower tail bounds in many situations where a direct application of the log-Sobolev machinery fails.

On the other hand, Corollary 4.16 does not capture dimension-free concentration with respect to ℓ_2-metrics, such as we have seen in the case of Gaussian concentration. It turns out that not every probability measure μ tensorizes in an ℓ_2-fashion. Nonetheless, by using Theorem 4.15 in a different manner, we will be able to completely characterize measures μ for which this is the case using transportation cost inequalities. This will be discussed in detail in section 4.4 below, and we postpone further discussion until then.

The remainder of this section is devoted to the proof of Theorem 4.15. The first step in the proof will be based on the following elementary property.

Lemma 4.18 (Chain rule for relative entropy). *Let $\mathbf{M}, \mathbf{N}$ be probability measures that define the joint distribution of random variables X, Y. Then*

$$\begin{aligned}&D(\mathbf{M}\{X, Y \in \cdot\}||\mathbf{N}\{X, Y \in \cdot\}) = \\ &D(\mathbf{M}\{X \in \cdot\}||\mathbf{N}\{X \in \cdot\}) + \mathbf{E}_{\mathbf{M}}[D(\mathbf{M}\{Y \in \cdot|X\}||\mathbf{N}\{Y \in \cdot|X\})].\end{aligned}$$

Proof. It is readily verified for $\mathbf{M} \ll \mathbf{N}$ that

$$\frac{d\mathbf{M}\{X, Y \in \cdot\}}{d\mathbf{N}\{X, Y \in \cdot\}} = \frac{d\mathbf{M}\{X \in \cdot\}}{d\mathbf{N}\{X \in \cdot\}} \frac{d\mathbf{M}\{Y \in \cdot|X\}}{d\mathbf{N}\{Y \in \cdot|X\}}$$

by definition of the Radon-Nikodym density (this is the *Bayes formula*). Thus

$$\begin{aligned}&D(\mathbf{M}\{X, Y \in \cdot\}||\mathbf{N}\{X, Y \in \cdot\}) = \\ &\mathbf{E}_{\mathbf{M}}\left[\log \frac{d\mathbf{M}\{X \in \cdot\}}{d\mathbf{N}\{X \in \cdot\}}\right] + \mathbf{E}_{\mathbf{M}}\left[\mathbf{E}_{\mathbf{M}}\left[\log \frac{d\mathbf{M}\{Y \in \cdot|X\}}{d\mathbf{N}\{Y \in \cdot|X\}}\middle| X\right]\right],\end{aligned}$$

and the conclusion follows from the definition of relative entropy. □

We now complete the proof of Theorem 4.15.

Proof (Theorem 4.15). The case $n = 1$ is trivial as the conclusion coincides with the assumption. We will proceed with the proof by induction on n. That is, let us suppose that the result has been proved for the case $n = k$. We presently show that this implies the result holds also for the case $n = k + 1$.

Fix for the time being a probability measure ν on $\mathbb{X}_1 \times \cdots \times \mathbb{X}_{k+1}$. Let $\nu^{(k)}$ be the marginal of ν on $\mathbb{X}_1 \times \cdots \times \mathbb{X}_k$, and let $\nu_{X_1,\ldots,X_k}$ be a version of the regular conditional probability $\mathbf{P}[X_{k+1} \in \cdot|X_1, \ldots, X_k]$. Then

$$D(\nu||\mu_1 \otimes \cdots \otimes \mu_{k+1}) = D(\nu^{(k)}||\mu_1 \otimes \cdots \otimes \mu_k) + \mathbf{E}_\nu[D(\nu_{X_1,\ldots,X_k}||\mu_{k+1})]$$

by the chain rule for relative entropy. We can now apply the induction hypothesis to the first term on the right and the assumption of the Theorem to the second term on the right. In particular, by the induction hypothesis

$$2\sigma^2 D(\nu^{(k)}||\mu_1 \otimes \cdots \otimes \mu_k) \geq \inf_{\mathbf{M} \in \mathcal{C}(\mu_1 \otimes \cdots \otimes \mu_k, \nu^{(k)})} \sum_{i=1}^{k} \varphi(\mathbf{E}_{\mathbf{M}}[w_i(X_i, Y_i)]),$$

while by the assumption of the Theorem

$$2\sigma^2 D(\nu_{y_1,\ldots,y_k}||\mu_{k+1}) \geq \inf_{\mathbf{M}\in\mathcal{C}(\mu_{k+1},\nu_{y_1,\ldots,y_k})} \varphi(\mathbf{E}_{\mathbf{M}}[w_{k+1}(X,Y)]).$$

Fix $\varepsilon > 0$, and choose an ε-minimizer $\mathbf{M}^{(k)} \in \mathcal{C}(\mu_1 \otimes \cdots \otimes \mu_k, \nu^{(k)})$ in the first inequality and an ε-minimizer $\mathbf{M}_{y_1,\ldots,y_k} \in \mathcal{C}(\mu_{k+1}, \nu_{y_1,\ldots,y_k})$ in the second inequality for every choice of $y_1, \ldots, y_k$. Then we have shown that

$$\begin{aligned}&2\sigma^2 D(\nu||\mu_1 \otimes \cdots \otimes \mu_{k+1}) \geq \\ &\sum_{i=1}^{k} \varphi(\mathbf{E}_{\mathbf{M}^{(k)}}[w_i(X_i,Y_i)]) + \varphi(\mathbf{E}_{\mathbf{M}^{(k)}}[\mathbf{E}_{\mathbf{M}_{Y_1,\ldots,Y_k}}[w_{k+1}(X_{k+1},Y_{k+1})]]) - 2\varepsilon,\end{aligned}$$

where we have used convexity of φ and that $(Y_1, \ldots, Y_k) \sim \nu^{(k)}$ under $\mathbf{M}^{(k)}$.

We now construct a coupling $\mathbf{M} \in \mathcal{C}(\mu_1 \otimes \cdots \otimes \mu_{k+1}, \nu)$ by sticking together the couplings $\mathbf{M}^{(k)}$ and $\mathbf{M}_{y_1,\ldots,y_k}$. To be precise, define $\mathbf{M}$ such that

$$\begin{aligned}&\mathbf{M}[X_1,\ldots,X_k,Y_1,\ldots,Y_k \in \cdot\,] = \mathbf{M}^{(k)},\\ &\mathbf{M}[X_{k+1},Y_{k+1} \in \cdot\,|X_1,\ldots,X_k,Y_1,\ldots,Y_k] = \mathbf{M}_{Y_1,\ldots,Y_k}.\end{aligned}$$

It is readily verified that $\mathbf{M} \in \mathcal{C}(\mu_1 \otimes \cdots \otimes \mu_{k+1}, \nu)$, so by the above inequality

$$2\sigma^2 D(\nu||\mu_1 \otimes \cdots \otimes \mu_{k+1}) \geq \inf_{\mathbf{M}\in\mathcal{C}(\mu_1\otimes\cdots\otimes\mu_{k+1},\nu)} \sum_{i=1}^{k+1} \varphi(\mathbf{E}_{\mathbf{M}}[w_i(X_i,Y_i)]) - 2\varepsilon.$$

As $\varepsilon > 0$ and ν were arbitrary, the proof for the case $n = k+1$ is complete. □

Remark 4.19. There is a minor technical issue that we have ignored in the above proof. We selected an ε-minimizer $\mathbf{M}_{y_1,\ldots,y_k}$ independently for every choice of $y_1, \ldots, y_k$, but in order for the remaining computations to make sense we must ensure that $\mathbf{M}_{y_1,\ldots,y_k}$ depends on $y_1, \ldots, y_k$ in a measurable fashion. However, this purely technical issue can be resolved using standard measurable selection arguments in any standard Borel space.

Problems

4.3 (Monge-Kantorovich duality: continuous case). We have stated Theorem 4.13 in the setting where $(\mathbb{X}, d)$ is a separable metric space. However, we only provided a proof for the case where $\mathbb{X}$ is a finite set. The goal of this problem is to work through the approximations needed to deduce the general result from the discrete case. To avoid confusion, define

$$T_1(\mu,\nu) := \inf_{\mathbf{M}\in\mathcal{C}(\mu,\nu)} \mathbf{E}_{\mathbf{M}}[d(X,Y)].$$

Our aim is to show that $T_1(\mu,\nu) = W_1(\mu,\nu)$.

a. Prove that T_1 is a metric on $\mathcal{P}_1(\mathbb{X})$.

Hint: to prove $T_1(\mu,\nu) \le T_1(\mu,\rho) + T_1(\nu,\rho)$, choose ε-optimal couplings $\mathbf{M}_1, \mathbf{M}_2$ in the definitions of $T_1(\mu,\rho), T_1(\nu,\rho)$ and consider $\mathbf{M}[X,Y,Z \in \cdot\,]$ defined by $\mathbf{M}[X,Y \in \cdot\,] = \mathbf{M}_1$ and $\mathbf{M}[Z \in \cdot\,|X,Y] = \mathbf{M}_2[X \in \cdot\,|Y]$.

b. For every $k \in \mathbb{N}$, construct disjoint sets $B_n^k \subseteq \mathbb{X}$ as follows:

$$B_1^k = \{x \in \mathbb{X} : d(x,x_1) < 2^{-k}\}, \quad B_n^k = \{x \in \mathbb{X} : d(x,x_n) < 2^{-k}\} \setminus \bigcup_{i=1}^{n-1} B_i^k,$$

where $\{x_n : n \in \mathbb{N}\}$ is a countable dense subset of $\mathbb{X}$. Choose an arbitrary point $y_n^k \in B_n^k$ for every n,k. For any $\mu \in \mathcal{P}_1(\mathbb{X})$, we now define

$$\mu_k = \sum_{n=1}^{\infty} \mu(B_n^k)\, \delta_{y_n^k}.$$

Show that we have $W_1(\mu_k,\mu) \le T_1(\mu_k,\mu) \le 2^{-k}$ for all $k \in \mathbb{N}$.

c. Show that the above construction can be modified such that μ_k has finite (rather than countable) support for all $k \in \mathbb{N}$, and $T_1(\mu_k,\mu) \to 0$ as $k \to \infty$.

d. Conclude using the already proved discrete case of Theorem 4.13 that the conclusion extends to the case where $(\mathbb{X},d)$ is any separable metric space.

4.4 (Monge-Kantorovich duality on $\mathbb{R}$). In many cases, explicit computation of the Wasserstein distance is impossible. However, there is an explicit expression for the Wasserstein distance on the real line $(\mathbb{R}, |\cdot|)$:

$$W_1(\mu,\nu) = \int_{-\infty}^{\infty} |F(t) - G(t)|\, dt,$$

where $F(t) = \mathbf{P}_\mu[X \le t]$ and $G(t) = \mathbf{P}_\nu[X \le t]$ denote the cumulative distribution functions of μ and ν, respectively.

a. Show that for smooth functions f with compact support

$$\int f\, d\mu = -\int_{\infty}^{\infty} f'(t) F(t) dt.$$

b. Use the previous part to prove the explicit expression for $W_1(\mu,\nu)$.

c. By Monge-Kantorovich duality, we obtain

$$\inf_{\mathbf{M} \in \mathcal{C}(\mu,\nu)} \mathbf{E}_{\mathbf{M}}[|X - Y|] = \int_{-\infty}^{\infty} |F(t) - G(t)|\, dt.$$

Find an explicit construction for the optimal coupling $\mathbf{M}$.

Hint: let $U \sim \mathrm{Uniform}[0,1]$. Then $F^{-1}(U) \sim \mu$ and $G^{-1}(U) \sim \nu$.

4.5 (Concentration for Markov chains). The transportation method can be useful for obtaining concentration results for dependent random variables. The goal of this problem is to develop the simplest possible example of this kind. Let $X_1, \ldots, X_n$ be a Markov chain with transition kernels

$$\mathbf{P}[X_k \in A | X_1, \ldots, X_{k-1}] = Q_k(X_{k-1}, A).$$

We will assume that the chain satisfies the Doeblin condition

$$\|Q_k(x, \cdot) - Q_k(x', \cdot)\|_{\mathrm{TV}} \le 1 - \alpha \quad \text{for all } x, x'$$

for some $\alpha > 0$. Even though $X_1, \ldots, X_n$ are not independent (we denote their joint distribution as μ), we can still obtain a transportation cost inequality by adapting the proof of the tensorization principle of Theorem 4.15.

a. Let ρ_1, ρ_2, ρ_3 be probability distributions on the same space. Show that there exists a joint distribution $\mathbf{M}$ of random variables X, Y, Z such that

$$\mathbf{M}[X \in \cdot\,] = \rho_1, \quad \mathbf{M}[Y \in \cdot\,] = \rho_2, \quad \mathbf{M}[Z \in \cdot\,] = \rho_3,$$

and such that

$$\mathbf{M}[X \neq Y] = \|\rho_1 - \rho_2\|_{\mathrm{TV}}, \quad \mathbf{M}[Y \neq Z] = \|\rho_2 - \rho_3\|_{\mathrm{TV}}.$$

Hint: this is similar to part a. of Problem 4.3.

b. Let ν be any distribution of random variables $Y_1, \ldots, Y_n$. Construct the probability measure $\mathbf{M}$ such that $Z_k = (X_k, \tilde{X}_k, Y_k)$, $k \le n$ satisfy

$$\begin{aligned}
\mathbf{M}[X_k \in \cdot\,|Z_1, \ldots, Z_{k-1}] &= Q_k(X_{k-1}, A), \\
\mathbf{M}[\tilde{X}_k \in \cdot\,|Z_1, \ldots, Z_{k-1}] &= Q_k(Y_{k-1}, A), \\
\mathbf{M}[Y_k \in \cdot\,|Z_1, \ldots, Z_{k-1}] &= \nu(Y_k \in \cdot\,|Y_1, \ldots, Y_{k-1}),
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{M}[X_k \neq \tilde{X}_k | Z_1, \ldots, Z_{k-1}] &= \|Q_k(X_{k-1}, \cdot) - Q_k(Y_{k-1}, \cdot)\|_{\mathrm{TV}}, \\
\mathbf{M}[\tilde{X}_k \neq Y_k | Z_1, \ldots, Z_{k-1}] &= \|Q_k(Y_{k-1}, \cdot) - \nu(Y_k \in \cdot\,|Y_1, \ldots, Y_{k-1})\|_{\mathrm{TV}}.
\end{aligned}$$

Show that

$$\begin{aligned}
&\mathbf{M}[X_k \neq Y_k | Z_1, \ldots, Z_{k-1}] \\
&\quad \le \sqrt{\tfrac{1}{2} D(\nu(Y_k \in \cdot\,|Y_1, \ldots, Y_{k-1}) || Q_k(Y_{k-1}, \cdot))} + (1 - \alpha) \mathbf{1}_{X_{k-1} \neq Y_{k-1}}.
\end{aligned}$$

c. Now adapt the proof of Theorem 4.15 to show that

$$\inf_{\mathbf{M} \in \mathcal{C}(\mu, \nu)} \frac{\alpha}{\sqrt{n}} \sum_{k=1}^{n} \mathbf{M}[X_k \neq Y_k] \le \sqrt{\tfrac{1}{2} D(\nu || \mu)} \quad \text{for all } \nu,$$

and deduce an extension of McDiarmid's inequality in the present setting (in the case of equal weights). The independent case is recovered if $\alpha = 1$.

4.3 Talagrand's concentration inequality

Up to this point, the metric perspective and the transportation method did not yield any new results beyond a complementary point of view on the concentration phenomenon. In the present section, however, we will see that the metric approach to concentration allows us to prove new concentration results that were not accessible by the methods we have developed so far.

Let $X_1, \ldots, X_n$ be independent. To understand the issue at hand, let us once more consider McDiarmid's inequality. One way to phrase it is as follows:

$$\|D_i f\|_\infty \le c_i \text{ for } 1 \le i \le n \qquad \Longrightarrow$$
$$\mathbf{P}[f(X_1, \ldots, X_n) - \mathbf{E}f(X_1, \ldots, X_n) \ge t] \le e^{-2t^2/\sum_{i=1}^n c_i^2} \text{ for } t \ge 0.$$

We proved this result in three different ways: using the martingale method, the transportation method, and the entropy method. The latter method, however, was able to produce much stronger results in terms of *one-sided* differences. For example, we obtained in Theorem 3.18 the one-sided bound

$$D_i^- f(x) \le c_i(x) \text{ for } 1 \le i \le n \qquad \Longrightarrow$$
$$\mathbf{P}[f(X_1, \ldots, X_n) - \mathbf{E}f(X_1, \ldots, X_n) \ge t] \le e^{-t^2/4\|\sum_{i=1}^n c_i^2\|_\infty} \text{ for } t \ge 0.$$

This is often a crucial improvement over McDiarmid's inequality. Unfortunately, while McDiarmid's inequality is a subgaussian inequality (it gives both an upper and a lower tail bound by applying the bound to f and $-f$), the one-sided result obtained by the entropy method can only give an *upper* tail bound and not a *lower* tail bound in terms of the one-sided differences $D_i^- f$ (as $D_i^-(-f) \ne -D_i^- f$). There are many situations in which one can control $D_i^- f$ only (cf. Example 3.19), and we have not yet developed any tool that can yield the subgaussian property in such cases.

The aim of this section is to investigate the one-sided difference inequality from the perspective of Lipschitz concentration. What type of Lipschitz property does the one-sided bound correspond to? For McDiarmid's inequality, the property $\|D_i\|_\infty \le c_i$ for all i is equivalent to the Lipschitz property

$$f(x) - f(y) \le \sum_{i=1}^n c_i \mathbf{1}_{x_i \ne y_i} \quad \text{for all } x, y.$$

If we relax the assumption to $D_i^-(x) \le c_i(x)$ for all i, x, it is therefore natural to consider the analogous "one-sided Lipschitz property"

$$f(x) - f(y) \le \sum_{i=1}^n c_i(x) \mathbf{1}_{x_i \ne y_i} \quad \text{for all } x, y.$$

It is easily seen that the latter property does indeed imply $D_i^- f(x) \le c_i(x)$. However, the converse is not true: the one-sided Lipschitz property is strictly

stronger than control on the one-sided gradient. While the two assumptions can often be verified in the same manner in applications, the one-sided gradient bound is not naturally expressed as a Lipschitz property, while the one-sided Lipschitz property is not naturally expressed as a gradient.

We have thus arrived at a fork in the road where the perspective of the present chapter diverges from the perspective developed in the previous chapters. To exploit the one-sided Lipschitz property, we will use the transportation method to derive an important concentration inequality due to Talagrand. The remarkable aspect of this result is that it yields the full subgaussian property (i.e., an upper *and* lower tail bound) even though only a one-sided assumption was imposed. This makes it possible to obtain lower tails in many examples that were out of reach of the theory developed in the previous chapter.

Theorem 4.20 (Talagrand). *Let $X_1, \ldots, X_n$ be independent, and suppose*

$$f(x) - f(y) \le \sum_{i=1}^n c_i(x) \mathbf{1}_{x_i \neq y_i} \quad \text{for all } x, y.$$

Then $f(X_1, \ldots, X_n)$ is $\| \sum_{i=1}^n c_i^2 \|_\infty$-subgaussian.

Remark 4.21. As the one-sided Lipschitz assumption implies $D_i^- f(x) \le c_i(x)$, the upper tail bound obtained from Talagrand's inequality is in fact slightly worse than the upper tail bound obtained from the one-sided difference inequality of Theorem 3.18. As was emphasized above, the key improvement over the previous chapter is the lower tail bound. On the other hand, we will see in the proof of Theorem 4.20 that the lower tail bound can be proved with variance proxy $\mathbf{E}[\sum_{i=1}^n c_i^2]$, which is even better than the bound $\| \sum_{i=1}^n c_i^2 \|_\infty$ given in the statement given above (in fact, this variance proxy coincides with the variance bound of Corollary 2.4). Thus the statement of Theorem 4.20 can be somewhat improved both in the upper and lower tails, but the present (already useful) statement is the most compact form of the result.

To illustrate Talagrand's inequality, let us revisit Example 3.19.

Example 4.22 (Random matrices). We recall the setting of Examples 2.5 and 3.19. Let M be an $n \times n$ symmetric matrix where $\{M_{ij} : i \ge j\}$ are i.i.d. symmetric Bernoulli random variables $\mathbf{P}[M_{ij} = \pm 1] = \frac{1}{2}$. We denote by $\lambda_{\max}(M)$ the largest eigenvalue of M, and by $v_{\max}(M)$ a corresponding eigenvector.

In Example 2.5 we computed the one-sided differences $D_{ij}^- \lambda_{\max}(M)$. However, the one-sided Lipschitz property can be verified is precisely the same manner. In particular, repeating the computation of Example 2.5, we obtain

$$\begin{aligned} \lambda_{\max}(M) - \lambda_{\max}(M') &\le 2 \sum_{i \ge j} v_{\max}(M)_i v_{\max}(M)_j (M_{ij} - M'_{ij}) \\ &\le 4 \sum_{i \ge j} |v_{\max}(M)_i| \, |v_{\max}(M)_j| \, \mathbf{1}_{M_{ij} \neq M'_{ij}}. \end{aligned}$$

The function $M \mapsto \lambda_{\max}(M)$ therefore satisfies the one-sided Lipschitz property with weights $c_{ij}(M) = 4|v_{\max}(M)_i||v_{\max}(M)_j|$. It now follows immediately from Talagrand's concentration inequality that the random variable $\lambda_{\max}(M)$ is 16-subgaussian. Thus we have finally obtained a full subgaussian counterpart of the variance bound obtained in Example 2.5.

The one-sided Lipschitz assumption of Talagrand's concentration inequality corresponds to a (local) Lipschitz property with respect to a weighted Hamming distance. When one is dealing with real-valued random variables, it is often most convenient to consider Lipschitz properties with respect to the usual Euclidean distance. While one can obtain such a result for specific distributions (for example, in the Gaussian case), it is not generally true that distributions in $\mathbb{R}^n$ satisfy a concentration proeprty with respect to the Euclidean distance. However, for *convex* functions, such a concentration property turns out to hold for any family of independent bounded random variables, regardless of the specific properties of their distributions. This simple observation is a very useful consequence of Talagrand's inequality.

Corollary 4.23. *Let $X_1, \ldots, X_n$ be independent with values in $[0,1]$. Then $f(X_1, \ldots, X_n)$ is $\|\|\nabla f\|^2\|_\infty$-subgaussian for every* convex *function f.*

Proof. The first-order condition for convexity implies

$$f(x) - f(y) \le \nabla f(x) \cdot (x - y) \quad \text{for all } x, y.$$

As $|x_i - y_i| \le 1$ by assumption, we obtain

$$f(x) - f(y) \le \sum_{i=1}^n \left| \frac{\partial f(x)}{\partial x_i} \right| \mathbf{1}_{x_i \ne y_i}.$$

The result follows immediately from Theorem 4.20. □

We now turn to the proof of Theorem 4.20. We will attempt to follow as closely as possible the transportation proof of McDiarmid's inequality in Example 4.17. Of course, unlike the weighted Hamming distance, the quantity $\sum c_i(x)\mathbf{1}_{x_i \ne y_i}$ that appears in the one-sided Lipschitz property is not a metric: it is not even symmetric in x, y! Remarkably, this turns out to be unimportant: we will prove a transportation cost inequality for an *asymmetric* notion of Wasserstein "distance" that captures the one-sided Lipschitz property.

Theorem 4.24 (Marton). *Define the asymmetric "distance"*

$$d_2(\mu,\nu) := \inf_{\mathbf{M} \in \mathcal{C}(\mu,\nu)} \sup_{\mathbf{E_M}[\sum_{i=1}^n c_i(X)^2] \le 1} \mathbf{E_M}\left[\sum_{i=1}^n c_i(X) \mathbf{1}_{X_i \ne Y_i} \right].$$

between probability measures μ, ν on $\mathbb{X}_1 \times \cdots \times \mathbb{X}_n$. Then

$$d_2(\nu, \mu_1 \otimes \cdots \otimes \mu_n) \le \sqrt{2D(\nu||\mu_1 \otimes \cdots \otimes \mu_n)},$$
$$d_2(\mu_1 \otimes \cdots \otimes \mu_n, \nu) \le \sqrt{2D(\nu||\mu_1 \otimes \cdots \otimes \mu_n)}$$

for any probability measures ν and $\mu_1 \otimes \cdots \otimes \mu_n$ and $\mathbb{X}_1 \times \cdots \times \mathbb{X}_n$.

With this asymmetric transportation cost inequality in hand, the remainder of the proof follows exactly as in the previous sections.

Proof (Theorem 4.20). Suppose f satisfies the one-sided Lipschitz property

$$f(x) - f(y) \le \sum_{i=1}^n c_i(x) \mathbf{1}_{x_i \neq y_i}.$$

Let $\mu := \mu_1 \otimes \cdots \otimes \mu_n$ be a product and let ν be any probability. Then

$$\mathbf{E}_\nu f - \mathbf{E}_\mu f = \inf_{\mathbf{M} \in \mathcal{C}(\nu,\mu)} \mathbf{E}_{\mathbf{M}}[f(X) - f(Y)] \le \mathbf{E}_\nu \big[\textstyle\sum_{i=1}^n c_i^2\big]^{1/2} d_2(\nu, \mu),$$
$$\mathbf{E}_\mu f - \mathbf{E}_\nu f = \inf_{\mathbf{M} \in \mathcal{C}(\mu,\nu)} \mathbf{E}_{\mathbf{M}}[f(X) - f(Y)] \le \mathbf{E}_\mu \big[\textstyle\sum_{i=1}^n c_i^2\big]^{1/2} d_2(\mu, \nu).$$

We therefore have by Theorem 4.24

$$|\mathbf{E}_\nu f - \mathbf{E}_\mu f| \le \sqrt{2 \, \| \textstyle\sum_{i=1}^n c_i^2 \|_\infty \, D(\nu||\mu)},$$

and it follows precisely as in the proof of Theorem 4.8 that $f(X_1, \ldots, X_n)$ is $\| \sum_{i=1}^n c_i^2 \|_\infty$-subgaussian whenever $X \sim \mu_1 \otimes \cdots \otimes \mu_n$. □

Remark 4.25. We have used Theorem 4.8 to deduce the subgaussian property, which by its definition controls both the upper and lower tail probabilities. The proof of Theorem 4.8, however, implies also a one-sided result: given f, μ,

$$\log \mathbf{E}_\mu[e^{\lambda\{f - \mathbf{E}_\mu f\}}] \le \frac{\lambda^2 \sigma^2}{2} \quad \text{for all } \lambda \ge 0$$

if and only if

$$\mathbf{E}_\nu f - \mathbf{E}_\mu f \le \sqrt{2\sigma^2 D(\nu||\mu)} \quad \text{for all } \nu.$$

As $\lambda \ge 0$ here, this characterizes the upper tail; the lower tail is obtained by applying this result to $-f$. Now note that there is an asymmetry in the proof of Theorem 4.20: for the upper tail, the best we can do is

$$\mathbf{E}_\nu f - \mathbf{E}_\mu f \le \sqrt{2 \, \| \textstyle\sum_{i=1}^n c_i^2 \|_\infty \, D(\nu||\mu)} \quad \text{for all } \nu;$$

for the lower tail, however, we have an even better bound

$$\mathbf{E}_\mu f - \mathbf{E}_\nu f \le \sqrt{2 \, \mathbf{E}_\mu[\textstyle\sum_{i=1}^n c_i^2] \, D(\nu||\mu)} \quad \text{for all } \nu.$$

Thus the proof of Theorem 4.20 already yields a sharper conclusion: for $t \ge 0$

$$\mathbf{P}[f(X) \ge \mathbf{E}f(X) + t] \le e^{-t^2/2\| \sum_i c_i^2 \|_\infty},$$
$$\mathbf{P}[f(X) \le \mathbf{E}f(X) - t] \le e^{-t^2/2\mathbf{E}[\sum_i c_i(X)^2]}$$

when $\{X_i\}$ are independent and f satisfies the one-sided Lipschitz property.

The rest of this section is devoted to the proof of Theorem 4.24. Following the logic of the previous section, the proof will consist of two parts. First, we will use a tensorization principle to reduce the problem to the one-dimensional case. Then, we will give a direct proof of Theorem 4.24 in one dimension, that is, we will prove an asymmetric analogue of Pinsker's inequality.

In order to understand how to apply tensorization, let us begin by stating a simple reformulation of the asymmetric distance d_2.

Lemma 4.26. *For any μ, ν on $\mathbb{X}_1 \times \cdots \times \mathbb{X}_n$, we have*

$$d_2(\mu,\nu) = \left[\inf_{\mathbf{M}\in\mathcal{C}(\mu,\nu)} \sum_{i=1}^n \mathbf{E}_{\mathbf{M}}[\mathbf{M}[X_i \neq Y_i|X]^2]\right]^{1/2}.$$

Proof. This follows immediately from

$$\mathbf{E}_{\mathbf{M}}\left[\sum_{i=1}^n c_i(X)\mathbf{1}_{X_i\neq Y_i}\right] = \mathbf{E}_{\mathbf{M}}\left[\sum_{i=1}^n c_i(X)\,\mathbf{M}[X_i \neq Y_i|X]\right]$$

and Cauchy-Schwarz for the inner product $\langle c, \tilde{c}\rangle = \mathbf{E}_{\mathbf{M}}[\sum_{i=1}^n c_i(X)\tilde{c}_i(X)]$. □

This simple reformulation of the definition of d_2 is already very close to the form of the tensorization principle that we proved in Theorem 4.15. In fact, only a minor modification is needed in the proof to establish the following.

Proposition 4.27. *Let μ_i be a probability measure on $\mathbb{X}_i$ such that*

$$\inf_{\mathbf{M}\in\mathcal{C}(\mu_i,\nu)} \mathbf{E}_{\mathbf{M}}[\mathbf{M}[X \neq Y|X]^2] \le 2D(\nu||\mu_i) \quad \textit{for all } \nu$$

holds for every $i = 1, \ldots, n$. Then we have

$$\inf_{\mathbf{M}\in\mathcal{C}(\mu_1\otimes\cdots\otimes\mu_n,\nu)} \sum_{i=1}^n \mathbf{E}_{\mathbf{M}}[\mathbf{M}[X_i \neq Y_i|X]^2] \le 2D(\nu||\mu_1\otimes\cdots\otimes\mu_n) \quad \textit{for all } \nu.$$

The same conclusion follows if the infimum in the first inequality is replaced by $\mathbf{M} \in \mathcal{C}(\nu, \mu_i)$ and in the second inequality by $\mathbf{M} \in \mathcal{C}(\nu, \mu_1 \otimes \cdots \otimes \mu_n)$.

Proof. We follow closely the proof of Theorem 4.15. Suppose the conclusion has been proved for the case $n = k$; it suffices to show that it holds for the case $n = k + 1$. To this end, define probability measures $\nu, \nu^{(k)}, \nu_{y_1,\ldots,y_k}$ as in the proof of Theorem 4.15, and fix $\varepsilon > 0$. By the induction hypothesis, we can find $\mathbf{M}^{(k)} \in \mathcal{C}(\mu_1 \otimes \cdots \otimes \mu_k, \nu^{(k)})$ and $\mathbf{M}_{y_1,\ldots,y_k} \in \mathcal{C}(\mu_{k+1}, \nu_{y_1,\ldots,y_k})$ such that

$$2D(\nu^{(k)}||\mu_1\otimes\cdots\otimes\mu_k) \ge \sum_{i=1}^k \mathbf{E}_{\mathbf{M}^{(k)}}[\mathbf{M}^{(k)}[X_i \neq Y_i|X]^2] - \varepsilon,$$

$$2D(\nu_{y_1,\ldots,y_k}||\mu_{k+1}) \ge \mathbf{E}_{\mathbf{M}_{y_1,\ldots,y_k}}[\mathbf{M}_{y_1,\ldots,y_k}[X \neq Y|X]^2] - \varepsilon.$$

Define $\mathbf{M} \in \mathcal{C}(\mu_1 \otimes \cdots \otimes \mu_{k+1}, \nu)$ as in the proof of Theorem 4.15. Then we obtain using the chain rule of relative entropy and the definition of $\mathbf{M}$

$$2D(\nu||\mu_1 \otimes \cdots \otimes \mu_{k+1}) \geq \sum_{i=1}^{k} \mathbf{E}_{\mathbf{M}}[\mathbf{M}[X_i \neq Y_i|X_1, \ldots, X_k]^2] - 2\varepsilon + \mathbf{E}_{\mathbf{M}}[\mathbf{M}[X_{k+1} \neq Y_{k+1}|Y_1, \ldots, Y_k, X]^2].$$

Now note that as $\mathbf{M}_{Y_1,\ldots,Y_k}[X_{k+1} \in \cdot] = \mu_{k+1}$, evidently X_{k+1} is independent of $\{X_i, Y_i : i \leq k\}$. Thus $\mathbf{M}[X_i \neq Y_i|X_1, \ldots, X_k] = \mathbf{M}[X_i \neq Y_i|X]$, so

$$2D(\nu||\mu_1 \otimes \cdots \otimes \mu_{k+1}) \geq \sum_{i=1}^{k+1} \mathbf{E}_{\mathbf{M}}[\mathbf{M}[X_i \neq Y_i|X]^2] - 2\varepsilon$$

using Jensen. Taking the infimum over $\mathbf{M}$ and letting $\varepsilon \downarrow 0$ yields the claim.

The case where ν and μ are reversed corresponds to reversing the roles of X and Y in the above proof. Thus the only change in the proof is that we must now show $\mathbf{M}[X_i \neq Y_i|Y_1, \ldots, Y_k] = \mathbf{M}[X_i \neq Y_i|Y]$. This follows as Y_{k+1} is conditionally independent of X_i given $Y_1, \ldots, Y_k$ by the definition of $\mathbf{M}$. □

By virtue of Proposition 4.27, it remains only to prove Theorem 4.24 in the case $n = 1$. To this end, we will first prove an analogue of Monge-Kantorovich duality in this setting by adapting the computations in Example 4.14.

Lemma 4.28. *Suppose that $\mu \sim \nu$ are probability measures on $\mathbb{X}$. Then*

$$\inf_{\mathbf{M} \in \mathcal{C}(\mu,\nu)} \mathbf{E}_{\mathbf{M}}[\mathbf{M}[X \neq Y|X]^2]^{\frac{1}{2}} = \sup_{\substack{f \geq 0 \\ \mu(f^2) \leq 1}} \{\mathbf{E}_\mu f - \mathbf{E}_\nu f\} = \left[\int \left(1 - \frac{d\nu}{d\mu}\right)_+^2 d\mu\right]^{\frac{1}{2}}.$$

Proof. It is easily seen by Cauchy-Schwarz that

$$\sup\{\mathbf{E}_\mu f - \mathbf{E}_\nu f\} = \sup \int \left(1 - \frac{d\nu}{d\mu}\right) f \, d\mu = \left[\int \left(1 - \frac{d\nu}{d\mu}\right)_+^2 d\mu\right]^{\frac{1}{2}},$$

where the supremum taken is over $f \geq 0$, $\mu(f^2) \leq 1$. Moreover,

$$\begin{aligned} \sup\{\mathbf{E}_\mu f - \mathbf{E}_\nu f\} &= \inf_{\mathbf{M} \in \mathcal{C}(\mu,\nu)} \sup \mathbf{E}_{\mathbf{M}}[f(X) - f(Y)] \\ &\leq \inf_{\mathbf{M} \in \mathcal{C}(\mu,\nu)} \sup \mathbf{E}_{\mathbf{M}}[f(X)\mathbf{1}_{X \neq Y}] \\ &= \inf_{\mathbf{M} \in \mathcal{C}(\mu,\nu)} \mathbf{E}_{\mathbf{M}}[\mathbf{M}[X \neq Y|X]^2]^{\frac{1}{2}}. \end{aligned}$$

It remains to prove that the inequality is attained. To this end, construct precisely the same coupling $\mathbf{M} \in \mathcal{C}(\mu, \nu)$ as in Example 4.14. Then

$$\mathbf{M}[X \neq Y|X] = \left(1 - \frac{d\nu}{d\mu}(X)\right)_+,$$

and it follows immediately that $\mathbf{E}_{\mathbf{M}}[\mathbf{M}[X \neq Y|X]^2] = \int (1 - \frac{d\nu}{d\mu})_+^2 d\mu$. □

We can now complete the proof of Theorem 4.24.

Proof (Theorem 4.24). By Proposition 4.27, it suffices to consider the case $n = 1$. That is, we must prove for any probability measures μ, ν on $\mathbb{X}$

$$d_2(\nu,\mu) \le \sqrt{2D(\nu||\mu)}, \qquad d_2(\mu,\nu) \le \sqrt{2D(\nu||\mu)}$$

(this is, in essence, an asymmetric analogue of Pinsker's inequality). It suffices to assume $\nu \ll \mu$, as otherwise $D(\nu||\mu) = \infty$ and the result is trivial. By a simple perturbation argument, we can assume that $\mu \sim \nu$ (replace ν by $d\nu_\varepsilon = (1+\varepsilon)^{-1}(\frac{d\nu}{d\mu} + \varepsilon)d\mu$ and let $\varepsilon \downarrow 0$ at the end of the proof).

The proof is ultimately a calculus exercise. It is not difficult to show that

$$x \log x - x + 1 - \frac{(1-x)^2}{2} \ge 0, \qquad -\log x - 1 + x - \frac{(1-x)^2}{2} \ge 0$$

for $0 \le x \le 1$ (note that the inequalities hold for $x = 1$, and the left-hand sides in these inequalities are decreasing functions for $0 \le x \le 1$). Thus

$$\begin{aligned} x \log x - x + 1 &= (x \log x - x + 1)\mathbf{1}_{x\le 1} + x(-\log x^{-1} - 1 + x^{-1})\mathbf{1}_{x>1} \\ &\ge \frac{(1-x)_+^2 + x(1-x^{-1})_+^2}{2} \end{aligned}$$

for all $x \ge 0$. We can therefore estimate

$$\begin{aligned} d_2(\mu,\nu)^2 + d_2(\nu,\mu)^2 &= \int \left(1 - \frac{d\nu}{d\mu}\right)_+^2 d\mu + \int \left(1 - \frac{d\mu}{d\nu}\right)_+^2 \frac{d\nu}{d\mu}\, d\mu \\ &\le 2 \int \left(\frac{d\nu}{d\mu} \log \frac{d\nu}{d\mu} - \frac{d\nu}{d\mu} + 1\right) d\mu = 2D(\nu||\mu). \end{aligned}$$

This evidently implies the claim. □

Problems

4.6 (Rademacher processes). Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent symmetric Bernoulli random variables $\mathbf{P}[\varepsilon_i = \pm 1] = \frac{1}{2}$, and let $T \subseteq \mathbb{R}^n$. Define

$$Z = \sup_{t\in T} \sum_{k=1}^n \varepsilon_k t_k, \qquad \sigma^2 = 4 \sup_{t\in T} \sum_{k=1}^n t_k^2.$$

Show that Z is σ^2-subgaussian (cf. Problems 2.2, 3.7, and 3.14).

4.7 (Balls and bins). Suppose that m balls are thrown independently and uniformly at random into n bins. Let Z be the number of empty bins. What can we say about the magnitude and fluctuations of the random variable Z?

a. Show that $\mathbf{E}[Z] = n(1 - 1/n)^m$.

b. Use McDiarmid's inequality to show that Z is $m/4$-subgaussian.

The bound on the fluctuations obtained by McDiarmid's inequality is counterintuitive: $\mathbf{E}[Z]$ decreases with m but the variance proxy in McDiarmid's inequality increases with m! Using Talagrand's concentration inequality, we can obtain a better bound on the fluctuations of Z.

c. Use Talagrand's inequality to show that Z is $n \wedge m$-subgaussian.

Hint: let $f_m(b_1, \dots, b_m)$ be the number of nonempty bins if we put ball i in bin b_i, and note that $f_m(b_1, \dots, b_m) = \sum_{i=1}^m \mathbf{1}_{b_i \neq b_j \text{ for } j<i}$. Show that $f_m(b) \leq f_{2m}(b'_1, b_1, \dots, b'_m, b_m) \leq f_m(b') + \sum_{i=1}^m \mathbf{1}_{b_i \neq b'_i} \mathbf{1}_{b_i \neq b_j \text{ for } j<i}$.

4.8 (Travelling salesman problem). Let $X_1, \dots, X_n$ be i.i.d. points that are uniformly distributed in the unit square $[0,1]^2$. We think of X_i as the location of city i. The goal of the travelling salesman problem is to find a tour through all n cities with the shortest possible length. We denote by

$$L_n := \min_\sigma \{\|X_{\sigma(1)} - X_{\sigma(2)}\| + \|X_{\sigma(2)} - X_{\sigma(3)}\| + \cdots + \|X_{\sigma(n)} - X_{\sigma(1)}\|\}$$

the length of the shortest tour, where the minimum is taken over all permutations of $\{1, \dots, n\}$. Let us begin by investigating the magnitude of L_n.

a. Show that $\mathbf{E}[L_n] \asymp \sqrt{n}$.

Hint: argue that $L_n \geq \sum_{k=1}^n \min_{l \neq k} \|X_k - X_l\|$ for the lower bound and $L_n \leq L_{n-1} + 2 \min_{k<n} \|X_n - X_k\|$ for the upper bound.

b. Use McDiarmid's inequality to show that L_n is $2n$-subgaussian.

The bound using McDiarmid's inequality is terrible: it yields an upper bound on the magnitude of the fluctuations that is of the same order as the mean. Thus McDiarmid's inequality does not even show that L_n concentrates around its mean. Using Talagrand's inequality, we will be able to obtain a much sharper concentration result. This requires some geometric insight.

c. Let $v = (0, a)$ and $w = (b, 0)$ be corners of a right-angled triangle $T = \text{conv}\{0, v, w\}$. Show that $\|v - x\|^2 + \|x - w\|^2 \leq \|v - w\|^2$ for any $x \in T$.

d. Prove the following: for any $x_1, \dots, x_n \in T$, there is a permutation σ such that $\|v - x_{\sigma(1)}\|^2 + \sum_{i=1}^{n-1} \|x_{\sigma(i)} - x_{\sigma(i+1)}\|^2 + \|x_{\sigma(n)} - w\|^2 \leq \|v - w\|^2$.

Hint: argue by induction. Suppose the result is true for all right-angled triangles S and $x_1, \dots, x_{n-1} \in S$. Divide T into two right-angled triangles by drawing a line from the origin to the hypothenuse. If both triangles contain points, then use the induction hypothesis to conclude. Otherwise, continue subdividing until the induction hypothesis applies.

e. Conclude that for any points $x_1, \dots, x_n \in [0,1]^2$, there exists a permutation σ such that $\|x_{\sigma(1)} - x_{\sigma(2)}\|^2 + \|x_{\sigma(2)} - x_{\sigma(3)}\|^2 + \cdots + \|x_{\sigma(n)} - x_{\sigma(1)}\|^2 \leq 4$.

We are now going to use this geometric insight to analyze the length of travelling salesman tours. Recall that a tour through a set of points $x_1, \ldots, x_n$ is defined by a permutation σ of $\{1, \ldots, n\}$. The length of a given tour will be denoted as $l_n(x, \sigma)$, so we have $L_n := \min_\sigma l_n(X, \sigma)$.

f. Let $x = \{x_1, \ldots, x_n\}$ and $y = \{y_1, \ldots, y_n\}$ be sets of points with $x \cap y \neq \varnothing$. Let σ be a tour of x and τ be a tour of y. Show that there exists a tour ρ of $x \cup y$ such that $l_{2n}(x \cup y, \rho) \leq l_n(y, \tau) + 2\sum_{i=1}^n \mathbf{1}_{x_i \notin y} d_i(x, \sigma)$, where $d_i(x, \sigma)$ is the distance between x_i and the previous point in the tour σ.

Hint: imagine σ and τ are two partially overlapping hiking trails marked red and blue. Your aim is to systematically explore the union of the trails. To this end, perform the following walk: start walking the blue trail; if at any point the red trail diverges from the blue trail, walk down the red trail until just before it hits the blue trail again, then walk back to where you diverged from the blue trail and continue down the blue trail. While this walk is not a tour (as some points are visited twice), you can "straighten it out" into a genuine tour without increasing its length.

g. Fix for every $x_1, \ldots, x_n \in [0,1]^2$ a tour σ_x as in part e. above. Show that $\min_\sigma l_n(x, \sigma) \leq \min_\sigma l_n(y, \sigma) + \sum_{i=1}^n 2 d_i(x, \sigma_x) \mathbf{1}_{x_i \neq y_i}$ for all $x, y \in [0,1]^{2n}$.

h. Conclude that L_n is 16-subgaussian for every $n \geq 1$.

4.9 (Convexity and Euclidean concentration). Corollary 4.23 shows that *convex* Lipschitz functions of bounded independent variables concentrate in the same manner as Lipschitz functions of Gaussian random variables. However, in the Gaussian case, convexity is not needed. The goal of this problem is to show that convexity is in fact essential in the setting of Corollary 4.23.

Let $\{X_k : k \geq 1\}$ be i.i.d. symmetric Bernoulli variables $\mathbf{P}[X_i = \pm 1] = \frac{1}{2}$. Consider for each $n \geq 1$ the function $f^n(x) = d(x, A^n)$ on $\mathbb{R}^n$, where

$$A^n = \left\{ y \in \{-1, 1\}^n : \sum_{i=1}^n y_i \leq 0 \right\}$$

and $d(x, A) := \inf_{y \in A} \|x - y\|$. Note that the function $f^n(x)$ is not convex.

a. Show that f^n is 1-Lipschitz with respect to the Euclidean distance on $\mathbb{R}^n$.

b. Show that $\mathrm{med}[f^n(X_1, \ldots, X_n)] = 0$.

c. Show that if $x \in \{-1, 1\}^n$ satisfies $\sum_{i=1}^n x_i \geq \sqrt{n}$, then

$$\sqrt{n} \leq \sum_{i=1}^n (x_i - y_i) \leq \sum_{i=1}^n |x_i - y_i|^2 \quad \text{for all } y \in A.$$

In particular, this implies $f^n(x) \geq n^{1/4}$.

d. Show that

$$\liminf_{n\to\infty} \mathbf{P}[f^n(X_1,\dots,X_n) \geq n^{1/4}] > 0.$$

Argue that this implies that $f^n(X_1,\dots,X_n)$ cannot be subgaussian with variance proxy independent of the dimension n.

e. Show that if g is convex and 1-Lipschitz with respect to the Euclidean distance on $\mathbb{R}^n$, then $g(X_1,\dots,X_n)$ is 4-subgaussian (independent of dimension n). In view of the above, convexity is evidently essential.

4.4 Dimension-free concentration and the T_2-inequality

In the previous sections we have obtained a complete characterization of the concentration of Lipschitz functions on a fixed metric space in terms of transportation cost inequalities (Theorem 4.8), and we have developed a tensorization principle for such inequalities (Theorem 4.15). Together, these two principles allow us deduce concentration of independent random variables in the following manner. Suppose that $X_i \sim \mu_i$ on $(\mathbb{X}_i, d_i)$ are such that

$$f(X_i) \text{ is 1-subgaussian when } |f(x) - f(y)| \leq d_i(x,y),$$

and that $X_1,\dots,X_n$ are independent. Then we have for any $\sum_{i=1}^n c_i^2 \leq 1$

$$f(X_1,\dots,X_n) \text{ is 1-subgaussian when } |f(x) - f(y)| \leq \sum_{i=1}^n c_i d_i(x_i,y_i).$$

This suffices to recover, for example, McDiarmid's inequality.

However, in the previous chapters, we have seen examples that exhibit substantially better concentration properties than is suggested by this general principle. For example, let $X_i \sim N(0,1)$ on $\mathbb{X}_i = \mathbb{R}$. Then the Gaussian concentration property states not only that each X_i exhibits the Lipschitz concentration property with respect to the metric $d_i(x,y) = |x-y|$, but also

$$f(X_1,\dots,X_n) \text{ is 1-subgaussian when } |f(x) - f(y)| \leq \left[\sum_{i=1}^n d_i(x_i,y_i)^2\right]^{\frac{1}{2}}.$$

Thus we even have dimension-free concentration for independent Gaussian variables with respect to the Euclidean distance $d(x,y) = [\sum_i d_i(x_i,y_i)^2]^{1/2}$ rather than just the weighted ℓ_1-distance $d_c(x,y) = \sum_i c_i d_i(x_i,y_i)$. This is a much stronger conclusion: indeed, any 1-Lipschitz function with respect to d_c is 1-Lipschitz with respect to d, but a function that is 1-Lipschitz with respect to d may not be better than $\sqrt{n}$-Lipschitz with respect to d_c.

At first sight, the fact that we do not capture concentration with respect to the Euclidean distance might appear to be an inefficiency in our approach.

One might hope that the conclusion of Theorem 4.15 can be improved to yield a statement of the following form: if

$$W_1(\mu_i,\nu) \le \sqrt{2\sigma^2 D(\nu||\mu_i)} \quad \text{for all } \nu$$

holds for each μ_i on $(\mathbb{X}_i, d_i)$, then for any $n \ge 1$

$$W_1(\mu_1 \otimes \cdots \otimes \mu_n,\nu) \le \sqrt{2\sigma^2 D(\nu||\mu_1 \otimes \cdots \otimes \mu_n)} \quad \text{for all } \nu$$

holds for $\mu_1 \otimes \cdots \otimes \mu_n$ on $(\mathbb{X}_1 \times \cdots \times \mathbb{X}_n, [\sum_{i=1}^n d_i^2]^{1/2})$. However, this conclusion is false: in general, it is *not* true that a distribution that exhibits the Lipschitz concentration property in one dimension will exhibit dimension-free concentration with respect to the Euclidean distance. For example, we have seen in Problem 4.9 that this conclusion fails already for symmetric Bernoulli variables. Thus dimension-free Euclidean concentration is a strictly stronger property than is guaranteed by Theorem 4.8. In this section, we will show that the latter property can nonetheless be characterized completely by means of a stronger form of the transportation cost inequality.

In order to develop improved concentration results, we must first identify where lies the inefficiency of our previous tensorization argument. Recall that

$$W_1(\mu_i,\nu) \le \sqrt{2\sigma^2 D(\nu||\mu_i)} \quad \text{for all } \nu, i$$

implies, using Theorem 4.15 with $\varphi(x) = x^2$ and $w_i(x,y) = d_i(x,y)$, that

$$\left[\inf_{\mathbf{M}\in\mathcal{C}(\mu_1\otimes\cdots\otimes\mu_n,\nu)} \sum_{i=1}^n \mathbf{E}_{\mathbf{M}}[d_i(X_i,Y_i)]^2\right]^{1/2} \le \sqrt{2\sigma^2 D(\nu||\mu_1 \otimes \cdots \otimes \mu_n)}.$$

The problem with this expression is that the left-hand side is not a Wasserstein distance. We resolved this problem in Corollary 4.16 by applying the Cauchy-Schwarz inequality. Such a brute-force solution can only yield a transportation cost inequality in terms of weighted ℓ_1-distance, however. On the other hand, note that the quantity on the left-hand side is already tantalizingly close to a Euclidean transportation cost inequality: if only $\mathbf{E}_{\mathbf{M}}[d_i(X_i,Y_i)]^2$ could be replaced by $\mathbf{E}_{\mathbf{M}}[d_i(X_i,Y_i)^2]$, we would immediately deduce

$$W_1(\mu_1 \otimes \cdots \otimes \mu_n,\nu) \le \sqrt{2\sigma^2 D(\nu||\mu_1 \otimes \cdots \otimes \mu_n)} \quad \text{for all } \nu$$

on $(\mathbb{X}_1 \times \cdots \times \mathbb{X}_n, [\sum_{i=1}^n d_i^2]^{1/2})$ by Jensen's inequality. Given the technology that we have already developed, can easily engineer this situation by starting from a slighly stronger inequality in one dimension.

Definition 4.29 (Quadratic Wasserstein metric). *The* quadratic Wasserstein metric *for probability measures μ,ν on a metric space $(\mathbb{X},d)$ is*

$$W_2(\mu,\nu) := \inf_{\mathbf{M}\in\mathcal{C}(\mu,\nu)} \sqrt{\mathbf{E}[d(X,Y)^2]}.$$

Corollary 4.30 (T_2-inequality). *Suppose that the probability measures μ_i on $(\mathbb{X}_i, d_i)$ satisfy the* quadratic transportation cost (T_2) inequality

$$W_2(\mu_i, \nu) \le \sqrt{2\sigma^2 D(\nu||\mu_i)} \quad \text{for all } \nu.$$

Then we have

$$W_2(\mu_1 \otimes \cdots \otimes \mu_n, \nu) \le \sqrt{2\sigma^2 D(\nu|\mu_1 \otimes \cdots \otimes \mu_n)} \quad \text{for all } \nu$$

on $(\mathbb{X}_1 \times \cdots \times \mathbb{X}_n, [\sum_{i=1}^n d_i^2]^{1/2})$.

Proof. Apply Theorem 4.15 with $\varphi(x) = x$ and $w_i(x,y) = d_i(x,y)^2$. □

By Jensen's inequality, we evidently have

$$W_1(\mu,\nu) \le \inf_{\mathbf{M}\in\mathcal{C}(\mu,\nu)} \mathbf{E}_{\mathbf{M}}[d(X,Y)] \le \inf_{\mathbf{M}\in\mathcal{C}(\mu,\nu)} \sqrt{\mathbf{E}_{\mathbf{M}}[d(X,Y)^2]} = W_2(\mu,\nu).$$

The T_2-inequality is therefore a stronger assumption than the transportation cost inequalities (or T_1-inequalities) that we have considered so far. On the other hand, combining Corollary 4.30 and Theorem 4.8 shows that if each measure μ_i satisfies a T_2-inequality, then the product measure $\mu_1 \otimes \cdots \otimes \mu_n$ satisfies the Lipschitz concentration property with respect to the Euclidean distance $d = [\sum_i d_i^2]^{1/2}$, which is a much stronger conclusion than could be deduced from the T_1-inequality. We have therefore obtained a *sufficient* condition for dimension-free Euclidean concentration.

We could verify at this point that the Gaussian distribution satisfies the T_2-inequality, so that the improved tensorization principle of Corollary 4.30 is sufficiently strong to capture Gaussian concentration (see Problems 4.10 and 4.11). This explains why the Gaussian distribution exhibits better concentration properties than were predicted by Corollary 4.16. Instead, we will presently prove a remarkable general fact: the T_2-inequality is not only sufficient, but also *necessary* for dimension-free Euclidean concentration to hold!

Theorem 4.31 (Gozlan). *Let μ be a probability measure on a Polish space $(\mathbb{X}, d)$, and let $\{X_i\}$ be i.i.d. $\sim \mu$. Denote by $d_n(x,y) := [\sum_{i=1}^n d(x_i,y_i)^2]^{1/2}$ the Euclidean metric on $\mathbb{X}^n$. Then the following are equivalent:*

1. μ satisfies the T_2-inequality on $(\mathbb{X}, d)$:

$$W_2(\mu,\nu) \le \sqrt{2\sigma^2 D(\nu||\mu)} \quad \text{for all } \nu.$$

2. $\mu^{\otimes n}$ satisfies the T_1-inequality on $(\mathbb{X}^n, d_n)$ for every $n \ge 1$:

$$W_1(\mu^{\otimes n}, \nu) \le \sqrt{2\sigma^2 D(\nu||\mu^{\otimes n})} \quad \text{for all } \nu,\ n \ge 1.$$

3. There is a constant C such that

$$\mathbf{P}[f(X_1,\ldots,X_n) - \mathbf{E}f(X_1,\ldots,X_n) \ge t] \le Ce^{-t^2/2\sigma^2}$$

for every $n \ge 1$, $t \ge 0$ and 1-Lipschitz function f on $(\mathbb{X}^n, d_n)$.

Let us emphasize that this striking result is quite unexpected. While Theorem 4.8 shows that Lipschitz concentration on a fixed metric space is characterized by the T_1-inequality, the necessity in Theorem 4.8 has little bearing on the behavior of the quadratic Wasserstein metric. The necessity of the T_2-inequality in Theorem 4.31 has a different origin: it is a consequence of a classical large deviation result in probability theory.

Theorem 4.32 (Sanov). *Let μ be a probability measure on a Polish space $\mathbb{X}$, and let $\{X_i\}$ be i.i.d. $\sim \mu$. Let O be a set of probability measures on $\mathbb{X}$ that is open for the weak convergence topology. Then*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left[\frac{1}{n}\sum_{k=1}^n \delta_{X_k} \in O\right] \ge - \inf_{\nu\in O} D(\nu||\mu).$$

Remark 4.33. We have only stated half of Sanov's theorem: a matching upper bound can be proved also (see Problem 4.12 below). However, only the lower bound will be needed in the proof of Theorem 4.31.

Proof. Fix $\nu \in O$ such that $D(\nu||\mu) < \infty$. Let $f = d\nu/d\mu$, and let $\mathbf{Q}$ be the probability under which $\{X_i\}$ are i.i.d. $\sim \nu$. As $f > 0$ ν-a.s., we can estimate

$$\begin{aligned}
&\mathbf{P}\left[\frac{1}{n}\sum_{k=1}^n \delta_{X_k} \in O\right] \ge \mathbf{P}\left[\frac{1}{n}\sum_{k=1}^n \delta_{X_k} \in O,\ \prod_{k=1}^n f(X_k) > 0\right] \\
&= \mathbf{E_Q}\left[\mathbf{1}_{\frac{1}{n}\sum_{k=1}^n \delta_{X_k}\in O} \prod_{k=1}^n f(X_k)^{-1}\right] \\
&\ge e^{-n\{\int \log f\, d\nu + \varepsilon\}} \mathbf{Q}\left[\frac{1}{n}\sum_{k=1}^n \delta_{X_k} \in O,\ \frac{1}{n}\sum_{k=1}^n \log f(X_k) \le \int \log f\, d\nu + \varepsilon\right].
\end{aligned}$$

Note that $\int \log f\, d\nu = D(\nu||\mu)$, while we have by the law of large numbers $\frac{1}{n}\sum_{k=1}^n \log f(X_k) \to \int \log f\, d\nu$ and $\frac{1}{n}\sum_{k=1}^n \delta_{X_k} \to \nu$ weakly $\mathbf{Q}$-a.s. Thus the probability in the last line converges to one, and it follows readily that

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left[\frac{1}{n}\sum_{k=1}^n \delta_{X_k} \in O\right] \ge -D(\nu||\mu) - \varepsilon.$$

It remains to let $\varepsilon \downarrow 0$ and take the supremum over all $\nu \in O$. □

We are now ready to prove Theorem 4.31. The proof of a few technical results that will be needed along the way is deferred to the end of this section.

Proof (Theorem 4.31). We already proved $1 \Rightarrow 2$ in Corollary 4.30, while the implication $2 \Rightarrow 3$ with $C = 1$ follows from Theorem 4.8 and the usual Chernoff bound. It therefore remains to prove $3 \Rightarrow 1$.

We will need the following three facts that will be proved below.

1. *Wasserstein law of large numbers:* $\mathbf{E}\big[W_2\big(\frac{1}{n}\sum_{k=1}^n \delta_{X_k}, \mu\big)\big] \to 0$ as $n \to \infty$.

2. *Lower-semicontinuity:* $O_t := \{\nu : W_2(\nu,\mu) > t\}$ is an open set.

3. *Smoothness:* $g_n : (x_1,\ldots,x_n) \mapsto W_2\big(\frac{1}{n}\sum_{k=1}^n \delta_{x_k}, \mu\big)$ is $n^{-1/2}$-Lipschitz.

The first two claims are essentially technical exercises: $\frac{1}{n}\sum_{k=1}^n \delta_{X_k}$ converges weakly to μ by the law of large numbers, so the only difficulty is to verify that the convergence holds in the slightly stronger sense of the quadratic Wasserstein distance; and lower-semicontinuity of W_2 is an elementary technical fact. The third claim is a matter of direct computation, which we will do below. Let us presently take these claims for granted and complete the proof.

As O_t is open, we can apply Sanov's theorem to conclude that

$$-\inf_{\nu\in O_t} D(\nu||\mu) \le \liminf_{n\to\infty} \frac{1}{n}\log \mathbf{P}[g_n(X_1,\ldots,X_n) > t].$$

As the function g_n is $n^{-1/2}$-Lipschitz, however, we have

$$\mathbf{P}[g_n(X_1,\ldots,X_n) > t] \le Ce^{-n(t-\mathbf{E}[g_n(X_1,\ldots,X_n)])^2/2\sigma^2}$$

by the dimension-free concentration assumption. This implies

$$-\inf_{\nu\in O_t} D(\nu||\mu) \le -\limsup_{n\to\infty} \frac{(t-\mathbf{E}[g_n(X_1,\ldots,X_n)])^2}{2\sigma^2} = -\frac{t^2}{2\sigma^2}$$

using the Wasserstein law of large numbers. Thus we have proved

$$\sqrt{2\sigma^2 D(\nu||\mu)} \ge t \quad \text{whenever} \quad W_2(\mu,\nu) > t.$$

The T_2-inequality follows by choosing $t = W_2(\mu,\nu) - \varepsilon$ and letting $\varepsilon \downarrow 0$. □

It remains to establish the three claims used in the proof. We begin with the Lipschitz property of g_n, which follows essentially from the triangle inequality.

Lemma 4.34. *$g_n : x \mapsto W_2\big(\frac{1}{n}\sum_{k=1}^n \delta_{x_k}, \mu\big)$ is $n^{-1/2}$-Lipschitz on $(\mathbb{X}^n, d_n)$.*

Proof. Let $\mathbf{M} \in \mathcal{C}\big(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \mu\big)$. If we define $\mu_i = \mathbf{M}[Y \in \cdot\,|X = x_i]$, then

$$\mathbf{E}_{\mathbf{M}}[f(X,Y)] = \frac{1}{n}\sum_{i=1}^n \int f(x_i,y)\,\mu_i(dy), \qquad \frac{1}{n}\sum_{i=1}^n \mu_i = \mu.$$

Conversely, every family of measures $\mu_1,\ldots,\mu_n$ with $\frac{1}{n}\sum_{i=1}^n \mu_i = \mu$ defines a coupling $\mathbf{M} \in \mathcal{C}\big(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \mu\big)$ in this manner. We can therefore estimate

$$
\begin{aligned}
&W_2\big(\tfrac{1}{n}\textstyle\sum_{i=1}^n \delta_{x_i}, \mu\big) - W_2\big(\tfrac{1}{n}\textstyle\sum_{i=1}^n \delta_{\tilde{x}_i}, \mu\big) \\
&\le \sup_{\frac{1}{n}\sum_{i=1}^n \mu_i = \mu} \left\{ \left[\frac{1}{n} \sum_{i=1}^n \int d(x_i, y)^2 \mu_i(dy) \right]^{\frac{1}{2}} - \left[\frac{1}{n} \sum_{i=1}^n \int d(\tilde{x}_i, y)^2 \mu_i(dy) \right]^{\frac{1}{2}} \right\} \\
&\le \sup_{\frac{1}{n}\sum_{i=1}^n \mu_i = \mu} \left[\frac{1}{n} \sum_{i=1}^n \int \{d(x_i, y) - d(\tilde{x}_i, y)\}^2 \mu_i(dy) \right]^{\frac{1}{2}} \\
&\le \frac{1}{\sqrt{n}} \left[\sum_{i=1}^n d(x_i, \tilde{x}_i)^2 \right]^{\frac{1}{2}},
\end{aligned}
$$

where in the last two lines we used, respectively, the reverse triangle inequality for L^2 norms (that is, $\|X\|_2 - \|Y\|_2 \le \|X - Y\|_2$) and for the metric d. □

Next, we establish lower-semicontinuity of W_2. The proof of this technical lemma is little more than an exercise in using weak convergence.

Lemma 4.35. *$\nu \mapsto W_2(\nu, \mu)$ is lower-semicontinuous in the weak topology.*

Proof. Let $\nu_n \to \nu$ weakly as $n \to \infty$. We must show that

$$\liminf_{n\to\infty} W_2(\nu_n, \mu) \ge W_2(\nu, \mu).$$

Fix $\varepsilon > 0$, and choose for each n a coupling $\mathbf{M}_n \in \mathcal{C}(\nu_n, \mu)$ such that

$$W_2(\nu_n, \mu) \ge \sqrt{\mathbf{E}_{\mathbf{M}_n}[d(X, Y)^2]} - \varepsilon.$$

We claim that the sequence $\{\mathbf{M}_n\}$ is tight. Indeed, the sequence $\{\nu_n\}$ is tight (as it converges) and clearly μ is itself tight. For any $\delta > 0$, choose a compact set K_δ such that $\nu_n(K_\delta) \ge 1 - \delta/2$ for all $n \ge 1$ and $\mu(K_\delta) \ge 1 - \delta/2$. Then evidently $\mathbf{M}_n(K_\delta \times K_\delta) \ge 1 - \delta$, and thus tightness follows.

Using tightness, we can choose a subsequence $n_k \uparrow \infty$ such that $\mathbf{M}_{n_k} \to \mathbf{M}$ weakly for some $\mathbf{M} \in \mathcal{C}(\nu, \mu)$ and $\liminf_n W_2(\nu_n, \mu) = \lim_k W_2(\nu_{n_k}, \mu)$. As the metric d is continuous and nonnegative, we obtain

$$\liminf_{n\to\infty} W_2(\nu_n, \mu) \ge \liminf_{k\to\infty} \sqrt{\mathbf{E}_{\mathbf{M}_{n_k}}[d(X, Y)^2]} - \varepsilon \ge \sqrt{\mathbf{E}_{\mathbf{M}}[d(X, Y)^2]} - \varepsilon.$$

Thus $\liminf_n W_2(\nu_n, \mu) \ge W_2(\nu, \mu) - \varepsilon$, and we conclude by letting $\varepsilon \downarrow 0$. □

Finally, we prove the Wasserstein law of large numbers. As the classical law of large numbers already implies that $\frac{1}{n}\sum_{k=1}^n \delta_{X_k} \to \mu$ weakly, this is almost obvious. The only issue that arises here is that convergence in W_2 is stronger than weak convergence, as it implies convergence of expectations of unbounded functions with up to quadratic growth. Proving that this is indeed the case under the assumption of Theorem 4.31 is an exercise in truncation.

Lemma 4.36. *Suppose that μ satisfies condition 3 of Theorem 4.31. Then we have $\mathbf{E}[W_2(\frac{1}{n}\sum_{k=1}^n \delta_{X_k}, \mu)] \to 0$ as $n \to \infty$ when $\{X_i\}$ are i.i.d. μ.*

Proof. Let $x^* \in \mathbb{X}$ be some arbitrary point. We truncate as follows:

$$\begin{aligned} W_2(\mu,\nu)^2 &= \inf_{\mathbf{M}\in\mathcal{C}(\mu,\nu)} \{\mathbf{E}_{\mathbf{M}}[d(X,Y)^2\mathbf{1}_{d(X,Y)\le a}] + \mathbf{E}_{\mathbf{M}}[d(X,Y)^2\mathbf{1}_{d(X,Y)>a}]\} \\ &\le a \inf_{\mathbf{M}\in\mathcal{C}(\mu,\nu)} \mathbf{E}_{\mathbf{M}}[d(X,Y)\wedge a] + \frac{4\int d(x,x^*)^3\{\mu(dx)+\nu(dx)\}}{a} \end{aligned}$$

using $(b+c)^3 \le 4(b^3+c^3)$ for $b,c \ge 0$. We claim that if $\nu_n \to \mu$ weakly, then

$$\inf_{\mathbf{M}\in\mathcal{C}(\nu_n,\mu)} \mathbf{E}_{\mathbf{M}}[d(X,Y)\wedge a] \xrightarrow{n\to\infty} 0.$$

Indeed, by the Skorokhod representation theorem, we can construct random variables $\{X_n\}$ and X on a common probability space such that $X_n \sim \nu_n$, $X \sim \mu$, and $X_n \to X$ a.s. Thus $\mathbf{E}[d(X_n,X)\wedge a] \to 0$ by bounded convergence, and as the joint law of X_n, X is in $\mathcal{C}(\nu_n,\mu)$ the claim follows. Thus $\nu_n \to \mu$ implies $W_2(\nu_n,\mu) \to 0$ if we can control the second term in the above truncation.

Recall that $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ satisfies $\mu_n \to \mu$ weakly a.s. by the law of large numbers. Therefore, following the above reasoning, we obtain

$$\limsup_{n\to\infty} \mathbf{E}[W_2(\mu_n,\mu)^2] \le \frac{8\int d(x,x^*)^3\mu(dx)}{a}$$

for every $a > 0$. Thus the result follows by letting $a \to \infty$, provided we can show that $\int d(x,x^*)^3\mu(dx) < \infty$. But as $x \mapsto d(x,x^*)$ is 1-Lipschitz, this follows readily from condition 3 of Theorem 4.31. □

We have now proved all the facts that were used above to establish Theorem 4.31. The proof of Theorem 4.31 is therefore complete.

Problems

4.10 (The Gaussian T_2-inequality). As we have already proved the Gaussian concentration property using the entropy method, Theorem 4.31 implies that the standard Gaussian distribution $N(0,1)$ on $\mathbb{R}$ must satisfy the T_2-inequality. It is instructive, however, to give a direct proof of this fact. By Theorem 4.31, this yields an alternative proof of Gaussian concentration.

Fix $X \sim \mu = N(0,1)$ and $\nu \ll \mu$. Denote their cumulative distribution functions as $F(t) = \mathbf{P}_\mu[X \le t]$ and $G(t) = \mathbf{P}_\nu[X \le t]$, and let $\varphi := G^{-1} \circ F$.

a. Show that

$$W_2(\mu,\nu) \le \mathbf{E}[|X-\varphi(X)|^2]^{1/2}, \qquad D(\nu||\mu) = \mathbf{E}\left[\log\frac{d\nu}{d\mu}(\varphi(X))\right].$$

b. Show that

$$e^{-t^2/2} = e^{-\varphi(t)^2/2}\frac{d\nu}{d\mu}(\varphi(t))\varphi'(t).$$

c. Use Gaussian integration by parts (Lemma 2.24) to show that

$$2D(\nu||\mu) = \mathbf{E}[|X - \varphi(X)|^2] + 2\,\mathbf{E}[\varphi'(X) - 1 - \log \varphi'(X)],$$

and conclude that $N(0,1)$ satisfies the T_2-inequality with $\sigma = 1$.

4.11 (Stochastic calculus and the Gaussian T_2-inequality). The goal of this problem is the give an alternative proof of the Gaussian T_2-inequality using stochastic calculus. The method developed here can be extended to prove the T_2-inequality for the laws of diffusion processes. For the purposes of this problem, we assume the reader is already familiar with stochastic calculus.

Fix $\mu = N(0,1)$ and $\nu \ll \mu$. Let $\{W_t\}_{t\in[0,1]}$ be standard Brownian motion under $\mathbf{P}$, and define the probability measure $d\mathbf{Q} = \frac{d\nu}{d\mu}(W_1)d\mathbf{P}$.

a. Show that for some nonanticipating process $\{\beta_t\}_{t\in[0,1]}$

$$\frac{d\nu}{d\mu}(W_1) = \exp\left(\int_0^1 \beta_t\, dW_t - \frac{1}{2}\int_0^1 \beta_t^2\, dt\right).$$

Hint: use the martingale representation theorem and Itô's formula.

b. Show that $\{Y_t\}_{t\in[0,1]}$ is Brownian motion under $\mathbf{Q}$, where

$$Y_t := W_t - \int_0^t \beta_s\, ds.$$

c. Argue that

$$W_2(\mu,\nu)^2 \le \mathbf{E}_{\mathbf{Q}}\left[\int_0^1 \beta_t^2\, dt\right].$$

d. Give a careful proof of the identity

$$D(\nu||\mu) = \mathbf{E}_{\mathbf{Q}}\left[\frac{1}{2}\int_0^1 \beta_t^2\, dt\right].$$

Conclude that $N(0,1)$ satisfies the T_2-inequality with $\sigma = 1$.

4.12 (Sanov's theorem). We proved in Theorem 4.32 half of Sanov's theorem. The other half yields a matching upper bound: if C is a set of probability measures on $\mathbb{X}$ that is compact for the weak convergence topology, then

$$\limsup_{n\to\infty} \frac{1}{n}\log \mathbf{P}\left[\frac{1}{n}\sum_{k=1}^n \delta_{X_k} \in C\right] \le -\inf_{\nu\in C} D(\nu||\mu).$$

Sanov's theorem therefore shows that relative entropy controls the exact asymptotic behavior, on a logarithmic scale, of the probability that empirical measures take values in a (sufficiently regular) unlikely set.

While only the lower bound in Sanov's theorem is needed in the proof of Theorem 4.31, it is instructive to prove the upper bound as well.

a. Show that for any probability measure ν and bounded function f

$$\frac{1}{n}\log \mathbf{P}\left[\frac{1}{n}\sum_{k=1}^{n} f(X_k) > \int f\,d\nu\right] \le \log \int e^f\,d\mu - \int f\,d\nu.$$

b. Fix $\varepsilon > 0$. Use the variational formula for entropy to show that for any probability measure ν, there is a bounded continuous function f_ν such that

$$\frac{1}{n}\log \mathbf{P}\left[\frac{1}{n}\sum_{k=1}^{n} f_\nu(X_k) > \int f_\nu\,d\nu\right] \le -D(\nu||\mu) + \varepsilon.$$

c. Show that if C is compact, then it can be covered by a finite number of sets of the form $\{\rho : \int f_\nu d\rho > \int f_\nu d\nu\}$ with $\nu \in C$.

d. Conclude the proof of the upper bound in Sanov's theorem.

4.13 (T_2-inequality and log-Sobolev inequalities). We have developed two completely different methods to obtain concentration inequalities: the entropy method and the transportation method. The goal of this problem is to develop some connections between the two.

a. Suppose that a probability μ on $\mathbb{R}^d$ satisfies the log-Sobolev inequality

$$\mathrm{Ent}_\mu[e^f] \le \frac{\sigma^2}{2}\,\mathbf{E}_\mu[\|\nabla f\|^2 e^f] \quad \text{for all } f.$$

Show that this implies that μ also satisfies the T_2-inequality.

By Theorem 4.31, the T_2-inequality is equivalent to dimension-free Euclidean concentration. We have just shown that the log-Sobolev inequality implies the T_2-inequality. One might hope that the converse is also true, that is, that T_2 implies log-Sobolev for probability measures on $\mathbb{R}^d$. This proves to be false, however: log-Sobolev is strictly stronger than T_2. It is possible to provide an explicit example that satisfies T_2 but not log-Sobolev (e.g., $\mu(dx) \propto e^{-|x|^3-|x|^{9/4}-3x^2\sin^2 x}dx$ on $\mathbb{R}$), but we omit the tedious verification of this fact.

Remarkably, however, it is easy to show that if μ satisfies the T_2-inequality, then it also satisfies the log-Sobolev inequality for *convex* functions. Moreover, for *concave* functions, the log-Sobolev inequality can even be improved!

a. Show that for any measure μ and function f,

$$\frac{\mathrm{Ent}_\mu[e^f]}{\mathbf{E}_\mu[e^f]} \le \int f\,d\nu - \int f\,d\mu \quad \text{with} \quad d\nu = \frac{e^f}{\mathbf{E}_\mu[e^f]}d\mu.$$

b. Show that

$$\frac{\mathrm{Ent}_\mu[e^f]}{\mathbf{E}_\mu[e^f]} \le \inf_{\mathbf{M}\in\mathcal{C}(\nu,\mu)} \mathbf{E}_{\mathbf{M}}[\nabla f(X)\cdot(X-Y)] \qquad \text{for convex } f,$$

$$\frac{\mathrm{Ent}_\mu[e^f]}{\mathbf{E}_\mu[e^f]} \le \inf_{\mathbf{M}\in\mathcal{C}(\nu,\mu)} \mathbf{E}_{\mathbf{M}}[\nabla f(Y)\cdot(X-Y)] \qquad \text{for concave } f.$$

c. Conclude that if μ satisfies the T_2-inequality, then

$$\begin{aligned} \mathrm{Ent}_\mu[e^f] &\le 2\sigma^2\,\mathbf{E}_\mu[\|\nabla f\|^2 e^f] && \text{for convex } f, \\ \mathrm{Ent}_\mu[e^f] &\le 2\sigma^2\,\mathbf{E}_\mu[\|\nabla f\|^2]\,\mathbf{E}_\mu[e^f] && \text{for concave } f. \end{aligned}$$

d. Deduce a version of the Gaussian concentration property (Theorem 3.25) for concave functions with improved variance proxy.

4.14 (Inf-convolution inequalities). The goal of this problem is to develop an alternative formulation of the T_2-inequality that is particularly useful for analysis of probability measures on $\mathbb{R}^d$. Before we state this alternative formulation, we must develop an analogue of Monge-Kantorovich duality for W_2.

a. Let $(\mathbb{X}, d)$ be a separable metric space. Show that

$$W_2(\mu,\nu)^2 = \sup_{g(x)-f(y)\le d(x,y)^2} \{\mathbf{E}_\nu g - \mathbf{E}_\mu f\}.$$

Hint: emulate the proof of Theorem 4.13 and Problem 4.3.

For any function f, define the *inf-convolution*

$$Q_t f(x) := \inf_{y\in\mathbb{X}} \left\{ f(y) + \frac{1}{2t}\, d(x,y)^2 \right\}.$$

We will show that for any probability μ on a separable metric space $(\mathbb{X}, d)$,

$$W_2(\mu,\nu) \le \sqrt{2\sigma^2 D(\nu||\mu)} \quad \text{for all } \nu \quad \text{iff} \quad \mathbf{E}_\mu[e^{Q_{\sigma^2}\{f-\mathbf{E}_\mu[f]\}}] \le 1 \quad \text{for all } f.$$

The latter inequality is called an *inf-convolution inequality*.

b. Prove the equivalence between the T_2 and inf-convolution inequalities.

Hint: emulate the proof of Theorem 4.8.

Let μ be a probability measure on $\mathbb{R}^d$ that satisfies the T_2-inequality. We have seen above that this does not necessarily imply that μ satisfies a log-Sobolev inequality. However, we will presently show that μ must at least satisfy a Poincaré inequality whenever the T_2-inequality holds.

c. Given any sufficiently smooth function $f : \mathbb{R}^d \to \mathbb{R}$, show that the function $v(t,x) = Q_t f(x)$ is the (Hopf-Lax) solution of the Hamilton-Jacobi equation

$$\frac{\partial v}{\partial t} + \frac{1}{2}\|\nabla v\|^2 = 0, \qquad v(0,\cdot) = f.$$

d. Show that if a probability μ on $\mathbb{R}^d$ satisfies the T_2-inequality, then

$$\mathrm{Var}_\mu[f] \le \sigma^2 \mathbf{E}_\mu[\|\nabla f\|^2] \quad \text{for all } f.$$

Hint: apply the inf-convolution inequality to tf and expand around $t = 0$.

Notes

§4.1. Historically, the metric approach to concentration was the first to be developed. The formulation in terms of Lipschitz functions dates back to the first proof of the Gaussian concentration property due to Tsirelson, Ibragimov, and Sudakov [90] using stochastic calculus, while the fundamental importance of Lipschitz concentration and its connection with with isoperimetric problems (Problem 4.2) was emphasized and systematically exploited by Milman in the context of Banach space theory [62]. A comprehensive treatment of these ideas can be found in [50]. Theorem 4.8 is due to [11]. The Gibbs variational principle dates back to the inception of statistical mechanics [39, Theorem III, p. 131]. Pinsker's inequality is a basic fact in information theory [20].

§4.2. The texts by Villani [97, 98] are a fantastic source on optimal transportation problems and their connections with other areas of mathematics. An elementary introduction to linear programming duality is given in [36] (in fact, linear programming duality was invented by Kantorovich in order to prove Theorem 4.13, see [94] for historical comments). The continuous extension in Problem 4.3 was inspired by the treatment in [31]. The optimal coupling for the trivial metric was constructed in [25].

The transportation method for proving concentration inequalities is due to Marton [54]. Both the tensorization method and Problem 4.5 are from [54]. The general formulation of Theorem 4.15 given here was taken from [13].

§4.3. Talagrand's concentration inequality was developed in [76, 80] in an isoperimetric form in terms of a "convex distance" from a point to a set (an entire family of related inequalities is obtained there as well). A detailed exposition of these results can be found in [84, 50]. It was realized by Marton [55] that Talagrand's inequality can be proved using the transportation method using the asymmetric "distance" d_2, and the proof we give is due to her (with a simplified proof for $n = 1$ due to Samson [71]). The more general inequalities from [80] can also be recovered by the transportation method [21]. Problems 4.7 and 4.8 were inspired by the presentation in [26]. Problem 4.9 is from [76].

It is also possible to prove Talagrand's concentration inequality indirectly (through its isoperimetric form) using log-Sobolev methods; see [13].

§4.4. That the T_2-inequality suffices for dimension-free Euclidean transportation was noted by Talagrand [85]. Problem 4.10 follows the proof in [85] that the Gaussian measure satisfies the T_2-inequality. The stochastic calculus proof of Problem 4.11 is taken from [24]. Theorem 4.31 is due to Gozlan [41]. Sanov's theorem is a classical result in large deviations theory [22]; the proof given here was taken from lecture notes by Varadhan. Problem 4.13 is from [71]. The connection between concentration and inf-convolutions is due to Maurey [57]; Problem 4.14 follows the presentation in [50].

Part II

Suprema

5

Maxima, approximation, and chaining

We have shown in the previous chapters that in many cases a function $f(X_1, \ldots, X_n)$ of i.i.d. random variables is close to its mean $\mathbf{E}[f(X_1, \ldots, X_n)]$. The concentration phenomenon says nothing, however, about the magnitude of the mean $\mathbf{E}[f(X_1, \ldots, X_n)]$ itself. One cannot hope to address such questions at the same level as generality as we investigated concentration: some additional structure is needed in order to develop any meaningful theory.

The type of structure that will be investigated in the sequel are suprema

$$F = \sup_{t \in T} X_t,$$

where $\{X_t\}_{t \in T}$ is a random process that is defined on some index set T. Such problems arise in numerous high-dimensional applications, such as random matrix theory and probability in Banach spaces, control of empirical processes in statistics and machine learning, random optimization problems, etc. It is typically the case that the distribution of individual X_t is well understood, so that the main difficulty lies in understanding the effect of the supremum. To this end, we formulated in Chapter 1 the following informal principle:

> *If $\{X_t\}_{t \in T}$ is "sufficiently continuous," the magnitude of $\sup_{t \in T} X_t$ is controlled by the "complexity" of the index set T.*

In the sequel, we proceed to make this informal idea precise.

5.1 Finite maxima

Before we can develop a general theory to control suprema of random processes, we must understand the simplest possible situation: the maximum of a finite number of random variables, that is, the case where the index set T has finite cardinality $|T| < \infty$. In fact, this special case will form the most basic ingredient of our theory. To develop a more general theory, the fundamental idea in the sequel will be to approximate the supremum over a general

index set by the maximum over a finite set in increasingly sophisticated ways. By appropriately combining these two basic ingredients—finite maxima and approximation—we will develop powerful tools that yield remarkably sharp control over the suprema of many random processes.

How can one bound the maximum of a finite number of random variables? The most naive approach imaginable is to bound the supremum by a sum:

$$\sup_{t\in T} X_t \le \sum_{t\in T} |X_t|.$$

Plugging this trivial fact into an expectation, we obtain

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \le |T|\, \sup_{t\in T} \mathbf{E}|X_t|.$$

Thus if we can control the magnitude of every random variable X_t individually, then we obtain a bound that grows linearly in the cardinality $|T|$.

Of course, bounding a maximum by a sum is an exceedingly crude idea, and it seems unlikely *a priori* that one could draw any remotely accurate conclusions from such a procedure. Nonetheless, this simple idea is not a bad as it may appear on first sight if we use it a bit more carefully. Suppose, for example, that the random variables X_t have bounded pth moment. Then

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \le \mathbf{E}\left[\sup_{t\in T} |X_t|^p\right]^{1/p} \le |T|^{1/p}\, \sup_{t\in T} \mathbf{E}[|X_t|^p]^{1/p},$$

where we have bounded the maximum by a sum after applying Jensen's inequality. This has significantly improved the dependence on the cardinality from $|T|$ to $|T|^{1/p}$. Evidently our control of the maximum of random variables is closely related to the tail behavior of these random variables: the thinner the tails (i.e., the larger p), the better we can control their maximum. Once this idea has been understood, however, there is no need to stop at moments: if the random variables X_t possess a finite moment generating function, we can apply an exponential transformation precisely as in the development of Chernoff bounds in section 3.1 to estimate the maximum.

Lemma 5.1 (Maximal inequality). *Suppose that* $\log \mathbf{E}[e^{\lambda X_t}] \le \psi(\lambda)$ *for all* $\lambda \ge 0$ *and* $t \in T$*, where* ψ *is convex and* $\psi(0) = \psi'(0) = 0$*. Then*

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \le \psi^{*-1}(\log |T|),$$

where $\psi^*(x) = \sup_{\lambda\ge 0}\{\lambda x - \psi(\lambda)\}$ *denotes the Legendre dual of the function* ψ*. In particular, if* X_t *is* σ^2*-subgaussian for every* $t \in T$*, we have*

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \le \sqrt{2\sigma^2 \log |T|}.$$

Proof. By Jensen's inequality, we have for any $\lambda > 0$

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \le \frac{1}{\lambda}\log \mathbf{E}[e^{\lambda \sup_{t\in T} X_t}] \le \frac{1}{\lambda}\log \sum_{t\in T} \mathbf{E}[e^{\lambda X_t}] \le \frac{\log|T| + \psi(\lambda)}{\lambda}.$$

As $\lambda > 0$ is arbitrary, we can now optimize over λ on the right hand side. In the special case that X_t is σ^2-subgaussian (so that $\psi(\lambda) = \lambda^2\sigma^2/2$), we obtain

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \le \inf_{\lambda>0}\left[\frac{\log|T|}{\lambda} + \frac{\sigma^2\lambda}{2}\right] = \sqrt{2\sigma^2\log|T|}.$$

In the general case, the only difficulty is to evaluate the infimum in

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \le \inf_{\lambda>0}\frac{\log|T| + \psi(\lambda)}{\lambda} = \psi^{*-1}(\log|T|).$$

Suppose ψ^* is invertible. Note that $\{\psi^*(z) + \psi(\lambda)\}/\lambda \ge z$ for all $\lambda > 0$ by the definition of ψ^*, and that the inequality is attained if we choose λ to be the optimizer in the definition of ψ^*. Setting $\psi^*(z) = \log|T|$ yields the conclusion.

It remains to show that that ψ^* is invertible. As ψ^* is the supremum of linear functions, $x \mapsto \psi^*(x)$ is convex and strictly increasing except at those values x where the maximum in the definition of ψ^* is attained at $\lambda = 0$, that is, when $\lambda x - \psi(\lambda) \le -\psi(0)$ for all $\lambda \ge 0$. By the first-order condition for convexity, the latter occurs if and only if $x \le \psi'(0) = 0$. Moreover, as $\psi^*(0) = 0$, we conclude that $x \mapsto \psi^*(x)$ is convex, strictly increasing, and nonnegative for $x \ge 0$. Thus the inverse $\psi^{*-1}(x)$ is well defined for $x \ge 0$. □

Lemma 5.1 should be viewed as an analogue of the Chernoff bound of Lemma 3.1 in the setting of maxima of random variables. Recall that the Chernoff bound states that if $\log \mathbf{E}[e^{\lambda X_t}] \le \psi(\lambda)$ for all $\lambda \ge 0$ and $t \in T$, then

$$\mathbf{P}[X_t \ge x] \le e^{-\psi^*(x)} \quad \text{for all } x \ge 0,\ t \in T.$$

Thus our bound on the magnitude of the maximum depends on $|T|$ as the inverse of the tail probability of the individual random variables (as the inverse of the function $e^{\psi^*(x)}$ is $\psi^{*-1}(\log x)$). This is not a coincidence. In fact, we can use the Chernoff bound directly to estimate the tail probabilities of the maximum (rather than the expectation as in Lemma 5.1) as follows.

Lemma 5.2 (Maximal tail inequality). *Suppose that* $\log \mathbf{E}[e^{\lambda X_t}] \le \psi(\lambda)$ *for all* $\lambda \ge 0$ *and* $t \in T$*, where* ψ *is convex and* $\psi(0) = \psi'(0) = 0$*. Then*

$$\mathbf{P}\left[\sup_{t\in T} X_t \ge \psi^{*-1}(\log|T| + u)\right] \le e^{-u} \quad \textit{for all } u \ge 0.$$

In particular, if X_t *is* σ^2*-subgaussian for every* $t \in T$*, we have*

$$\mathbf{P}\left[\sup_{t\in T} X_t \ge \sqrt{2\sigma^2\log|T|} + x\right] \le e^{-x^2/2\sigma^2} \quad \textit{for all } x \ge 0.$$

Proof. We readily estimate using the Chernoff bound

$$\mathbf{P}\left[\sup_{t\in T} X_t \geq x\right] = \mathbf{P}\left[\bigcup_{t\in T}\{X_t \geq x\}\right] \leq \sum_{t\in T}\mathbf{P}[X_t \geq x] \leq e^{\log|T|-\psi^*(x)}.$$

Writing $u = \psi^*(x) - \log|T|$ yields the first inequality (the invertibility of ψ^* was shown in the proof of Lemma 5.1). In the subgaussian case,

$$\psi^{*-1}(\log T + u) = \sqrt{2\sigma^2(\log|T| + u)} \leq \sqrt{2\sigma^2\log|T|} + \sqrt{2\sigma^2 u}$$

yields the second inequality. □

The argument used in the proof of Lemma 5.2 is called a *union bound*: we have estimated the probability of a union of events by the sum of the probabilities $\mathbf{P}[A \cup B] \leq \mathbf{P}[A] + \mathbf{P}[B]$. This crude estimate plays exactly the same role in the proof of Lemma 5.2 as does bounding the maximum of random variables by their sum in the proof of Lemma 5.1.

Remark 5.3. While this may not be evident at the outset, the proofs of Lemmas 5.1 and 5.2 are based on precisely the same idea. Indeed, the union bound is merely another example of bounding a maximum by a sum:

$$\mathbf{P}[A_1 \cup \cdots \cup A_n] = \mathbf{E}[\max\{\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_n}\}] \leq \mathbf{E}[\mathbf{1}_{A_1}] + \cdots + \mathbf{E}[\mathbf{1}_{A_n}].$$

Lemmas 5.1 and 5.2 are therefore ultimately implementing the same bound in a slightly different way. In fact, is not difficult to deduce a form of Lemma 5.1 with a slightly worse constant directly from Lemma 5.2 by integrating the tail bound, that is, using $\mathbf{E}[Z] = \int_0^\infty \mathbf{P}[Z \geq z]\,dz$ for $Z \geq 0$.

We have obtained above some simple bounds on the maximum of a finite number of random variables. How good are these bounds? There are several reasons to be suspicious. On the one hand, we have obtained our estimates in an exceedingly crude fashion by bounding a maximum by a sum. On the other hand, while we made assumptions about the tail behavior of the individual variables X_t, we made no assumptions of any kind about the joint distribution of $\{X_t\}_{t\in T}$. One would expect that dependencies between the random variables X_t to make a significant difference to their maximum. As an extreme example, suppose $\{X_t\}_{t\in T}$ are completely dependent in the sense that $X_t = X_s$ for all $t, s \in T$. Then $\mathbf{E}[\sup_t X_t] = \mathbf{E}[X_s]$ does not depend on $|T|$ at all, whereas the bound in Lemma 5.1 necessarily grows with $|T|$. Of course, there is no contradiction: Lemma 5.1 is correct, but is evidently far from sharp in the presence of strong dependence between the random variables X_t.

Remarkably, however, Lemmas 5.1 and 5.2 prove to be essentially sharp when the random variables $\{X_t\}_{t\in T}$ are *independent*. It is perhaps surprising that a method as crude as bounding a maximum by a sum would lead to a sharp result in any nontrivial situation. However, it turns out that this idea is not as bad as may be expected on first sight in the presence of independence.

For example, consider the union bound $\mathbf{P}[A \cup B] \leq \mathbf{P}[A] + \mathbf{P}[B]$. Equality holds when A and B are disjoint, but this is certainly not the case in the proof of Lemma 5.2. Nonetheless, when A and B are independent, the probability that they occur simultaneously is much smaller than the individual probabilities, so that we still have $\mathbf{P}[A \cup B] \gtrsim \mathbf{P}[A] + \mathbf{P}[B]$. This idea will be exploited in Problem 5.1 below to show that Lemmas 5.1 and 5.2 are essentially sharp in the independent case. When viewed in terms of a sum of random variables, we see that in this setting the sum is dominated by its largest term, so that approximating the maximum by a sum is not such a bad idea after all.

Problems

5.1 (Maxima of independent random variables). The proofs of the maximal inequalities in the present section rely on a very crude device: bounding the maximum of random variable by a sum. Nonetheless, when the random variables are independent, the bounds we obtain above are often sharp. To understand why, we must prove lower bounds of the same order.

It is easiest to consider first the setting of Lemma 5.2. Let us begin by proving matching upper and lower union bounds for independent events.

a. Show that if $A_1, \ldots, A_n$ are independent events, then

$$(1 - e^{-1}) \left\{ 1 \wedge \sum_{k=1}^{n} \mathbf{P}[A_k] \right\} \leq \mathbf{P}\left[\bigcup_{k=1}^{n} A_k \right] \leq 1 \wedge \sum_{k=1}^{n} \mathbf{P}[A_k].$$

Hint: $\prod_{k=1}^{n} \{1 - x_k\} \leq \exp(-\sum_{k=1}^{n} x_k)$ and $1 - e^{-x} \geq (1 - e^{-1})\, 1 \wedge x$.

b. Let η^* be a strictly increasing convex function, and suppose that

$$\mathbf{P}[X_t \geq x] \geq e^{-\eta^*(x)} \quad \text{for all } x \geq 0,\ t \in T.$$

Conclude that for $u \geq 0$

$$\mathbf{P}\left[\sup_{t \in T} X_t \geq \eta^{*-1}(\log |T| + u) \right] \geq (1 - e^{-1})\, e^{-u},$$

and compare with the corresponding upper bound in Lemma 5.2.

Now that we have obtained a lower bound on the tail probability of the maximum (corresponding to the upper bound of Lemma 5.2), we can obtain a lower bound on the expectation of the maximum (corresponding to the upper bound of Lemma 5.1) by integrating the tail bound.

c. Deduce from the previous part that for $x \geq 0$

$$\mathbf{P}\left[\sup_{t \in T} X_t \geq \eta^{*-1}(2 \log |T|)/2 + x \right] \geq (1 - e^{-1})\, e^{-\eta^*(2x)/2}.$$

Hint: use concavity of η^{*-1}.

d. Conclude that if

$$e^{-\eta^*(x)} \le \mathbf{P}[X_t \ge x] \le e^{-\psi^*(x)} \quad \text{for all } x \ge 0,\ t \in T,$$

then we have

$$\frac{1-e^{-1}}{2}\,\eta^{*-1}(2\log|T|) + \sup_{t\in T}\mathbf{E}[0 \wedge X_t] \le \mathbf{E}\left[\sup_{t\in T} X_t\right] \le \psi^{*-1}(\log|T|).$$

Hint: use $\mathbf{E}[0 \vee Z] = \int_0^\infty \mathbf{P}[Z \ge x]dx$.

The upper and lower bound in the previous part are generally of the same order, provided that we start with upper and lower bounds on $\mathbf{P}[X_t \ge x]$ of the same order. For example, let us consider the case of Gaussian variables.

e. For $X \sim N(0,1)$, show that

$$\mathbf{P}[X \ge x] \ge \frac{e^{-x^2}}{2\sqrt{2}} \quad \text{for all } x \ge 0.$$

Hint: write the probability as an integral and use $(v+x)^2 \le 2v^2 + 2x^2$.

f. Let $X_1, \ldots, X_n$ be i.i.d. Gaussian random variables with zero mean and unit variance. Show that the above bound implies

$$\frac{1-e^{-1}}{2}\sqrt{2\log n 2^{-3/4}} - \frac{1}{\sqrt{2\pi}} \le \mathbf{E}\left[\max_{i\le n} X_i\right] \le \sqrt{2\log n}.$$

In particular, $c\sqrt{\log n} \le \mathbf{E}[\max_{i\le n} X_i] \le C\sqrt{\log n}$ for n sufficiently large.

g. If $X_1, X_2, \ldots$ are i.i.d. Gaussian, prove the asymptotic

$$\frac{\max_{i\le n} X_i}{\sqrt{2\log n}} \xrightarrow{n\to\infty} 1 \quad \text{in probability.}$$

Hint: for the upper bound, see Problem 3.5. For the lower bound, proceed analogously using a suitable improvement on the Gaussian tail lower bound obtained above (use $(v+x)^2 \le (1+\varepsilon^{-1})v^2 + (1+\varepsilon)x^2$).

5.2 (Approximating a maximum by a sum). Show that for $\lambda > 0$

$$\max_{t\in T} X_t \le \frac{1}{\lambda}\log\sum_{t\in T} e^{\lambda X_t} \le \max_{t\in T} X_t + \frac{\log|T|}{\lambda}.$$

Thus when λ is large, the sum is increasingly dominated by its largest term. This simple observation is often useful in problems where a smooth approximation of the maximum function $x \mapsto \max_i x_i$ is needed.

5.3 (Johnson-Lindenstrauss lemma). The following functional analysis result has found many applications in computer science and signal processing.

Let $x_1, \ldots, x_n$ be points in a Hilbert space H. Then for every $0 < \varepsilon < 1$ and $k \gtrsim \varepsilon^{-2} \log n$, there exists a linear map $T : H \to \mathbb{R}^k$ such that

$$(1-\varepsilon)\|x_i - x_j\| \le \|Tx_i - Tx_j\| \le (1+\varepsilon)\|x_i - x_j\| \quad \textit{for all } 1 \le i, j \le n.$$

This result should interpreted in terms of compression: if we want to store the distances between n points in a data structure, and if we tolerate a small distortion of order ε, it suffices to store an $n \times k$ matrix of size $\sim n \log n$ rather than the full $n \times n$ distance matrix of size $\sim n^2$.

At first sight, the Johnson-Lindenstrauss lemma has nothing to do with probability: it is a deterministic statement about the geometry of Hilbert spaces. However, the easiest way to find T is to select it randomly!

a. Argue that we can assume without loss of generality that $H = \mathbb{R}^n$.

b. For a $k \times n$ random matrix T such that T_{ij} are i.i.d. $N(0, k^{-1})$, show that

$$\mathbf{P}[|\|Tz\| - \mathbf{E}\|Tz\|| \ge \varepsilon \|z\|] \le 2e^{-k\varepsilon^2/2} \quad \text{for } z \in \mathbb{R}^n.$$

Hint: Gaussian concentration.

c. Show that

$$\sqrt{1 - k^{-1}}\|z\| \le \mathbf{E}\|Tz\| \le \|z\|,$$

and conclude that for $0 < \varepsilon < 1$ and $k \ge \varepsilon^{-1}$

$$\mathbf{P}[(1-\varepsilon)\|z\| < \|Tz\| < (1+\varepsilon)\|z\|] \ge 1 - 2e^{-k\varepsilon^2/8} \quad \text{for } z \in \mathbb{R}^n.$$

Hint: Use $\mathbf{E}\|Tz\| \le \mathbf{E}[\|Tz\|^2]^{1/2}$ for the upper bound. For the lower bound, estimate $\mathrm{Var}\|Tz\|$ from above using the Gaussian Poincaré inequality.

d. Show that if $k > 24\varepsilon^{-2} \log n$, then

$$\mathbf{P}[(1-\varepsilon)\|x_i - x_j\| < \|Tx_i - Tx_j\| < (1+\varepsilon)\|x_i - x_j\| \text{ for all } i, j] > 0.$$

Hint: use a union bound.

5.2 Covering, packing, and approximation

If the set T is infinite, the maximal inequalities of the previous section provide no information. This is, however, not surprising. We have seen that the inequalities for finite maxima work well when the random variables are independent. On the other hand, suppose that T is infinite but that $t \mapsto X_t$ is continuous in a suitable sense. Then $\lim_{t \to s} X_t = X_s$, so X_t and X_s must be strongly dependent when t and s are nearby points! Thus the lack of independence should in fact help us to control the infinite supremum: we should apply the maximal inequalities of the previous section only to a finite number

of well-separated points (at which the process might be expected to be nearly independent), and use continuity to control the fluctuations of the remaining (strongly dependent) degrees of freedom. In this section, we will develop the crudest illustration of this principle, which will be systematically developed in the sequel into a powerful machinery to control suprema.

To implement the above idea, we need to have a quantitative notion of continuity. In this section, we will use the simplest (but, as we will see, often unsatisfactory) such notion for random processes.

Definition 5.4 (Lipschitz process). *The random process $\{X_t\}_{t\in T}$ is called* Lipschitz *for a metric d on T if there exists a random variable C such that*

$$|X_t - X_s| \le Cd(t,s) \quad \text{for all } t,s \in T.$$

Given a Lipschitz process, our aim is to approximate the supremum over T by the maximum over a finite set N, to which we will apply the inequalities of the previous section. To obtain a good bound, we have two competing demands: on the one hand, we would like the set N to be as small as possible (so that the bound on the maximum is small); on the other hand, to control the approximation error, we must make sure that every point in T is close to at least one of the points in N. This leads to the following concept.

Definition 5.5 (ϵ-net and covering number). *A set N is called an* ε-net *for (T,d) if for every $t \in T$, there exists $\pi(t) \in N$ such that $d(t,\pi(t)) \le \varepsilon$. The smallest cardinality of an ε-net for (T,d) is called the* covering number

$$N(T,d,\varepsilon) := \inf\{|N| : N \text{ is an } \varepsilon\text{-net for } (T,d)\}.$$

The covering number $N(T,d,\varepsilon)$ should be viewed as a measure of the complexity of the set T at the scale ε: the more complex T, the more points we will need to approximate its structure up to a fixed precision. Alternatively, we can interpret the covering number as describing the geometry of the metric space (T,d). Indeed, let $B(t,\varepsilon) = \{s : d(t,s) \le \varepsilon\}$ be a ball of radius ε. Then

$$N \text{ is an } \varepsilon\text{-net} \qquad \text{if and only if} \qquad T \subseteq \bigcup_{t\in N} B(t,\varepsilon),$$

so that the covering number $N(T,d,\varepsilon)$ is the smallest number of balls of radius ε needed to cover T (hence the name). We can therefore interpret the covering number as a measure of the degree of (non-)compactness of (T,d).

Remark 5.6. In many applications, we may want to compute the supremum $\sup_{t\in T} X_t$ of a stochastic process $\{X_t\}_{t\in S}$ that is defined on a larger index set $S \supset T$. In this case, even though we are only interested in the process on the set T, it is not necessary to require that the ε-net N is a subset of T: it can be convenient to approximate the set T by points in $S\backslash T$ also. For this reason, we have not insisted in the above definition that $N \subseteq T$.

We are now ready to develop our first bound on the supremum of a random process. We adopt the notation of Definitions 5.4 and 5.5.

Lemma 5.7 (Lipschitz maximal inequality). *Suppose $\{X_t\}_{t\in T}$ is a Lipschitz process (Definition 5.4) and X_t is σ^2-subgaussian for every $t \in T$. Then*

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \le \inf_{\varepsilon>0}\{\varepsilon\mathbf{E}[C] + \sqrt{2\sigma^2 \log N(T,d,\varepsilon)}\}.$$

Note that this result is indeed a simple incarnation of the informal principle formulated in Chapter 1: if the process X_t is "sufficiently continuous," then $\sup_{t\in T} X_t$ is controlled by the "complexity" of the index set T.

Proof. Let $\varepsilon > 0$ and let N be an ε-net. Then

$$\sup_{t\in T} X_t \le \sup_{t\in T}\{X_t - X_{\pi(t)}\} + \sup_{t\in T} X_{\pi(t)} \le C\varepsilon + \max_{t\in N} X_t.$$

Taking the expectation and using Lemma 5.1 yields

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \le \varepsilon\mathbf{E}[C] + \sqrt{2\sigma^2 \log |N|}.$$

Optimizing over ε-nets N and $\varepsilon > 0$ yields the result. □

Remark 5.8. The idea behind Lemma 5.7 is that it allows us to trade off between exploiting independence (better at large scales) and controlling for dependence (worse at large scales). However, note that we never explicitly assume or use independence in the proof: instead, the distance d could be interpreted as a proxy for the degree of independence. While the conclusion of Lemma 5.7 does not depend on this validity of this interpretation, we expect that such bounds (and the more powerful bounds to be developed in the sequel) will be the most effective when the distance d is chosen in such a way that large distance does indeed correspond to more independence. This is often the case in practice. In the case of Gaussian processes, for example, we will see in the next chapter that this idea holds to such a degree that we can obtain matching upper and lower bounds for the supremum of Gaussian processes in terms of the geometry of the index set (T, d), albeit in a much more sophisticated manner than is captured by the trivial Lemma 5.7.

Remark 5.9. When $N(T, d, \varepsilon) = \infty$, the bound of Lemma 5.7 is infinite. However, note that if $X_1, X_2, \ldots$ are i.i.d. unbounded random variables, then we already have $\sup_i X_i = \infty$ a.s. It is therefore to be expected that the supremum of a random process will typically indeed be infinite if it contains infinitely many independent degrees of freedom. Thus the fact that $N(T, d, \varepsilon) = \infty$ (which means there are infinitely many points in T that are well separated) yields an infinite bound is not a shortcoming of Lemma 5.7. To obtain a finite supremum for noncompact index sets T one must often add a penalty inside the supremum; such problems will be investigated in section 5.4 below.

In the remainder of this section, we will illustrate the application of Lemma 5.7 using two illuminating examples. Along the way, we will develop some useful examples of how one can control covering numbers.

Example 5.10 (Random matrices). Let M be an $n \times m$ random matrix such that M_{ij} are independent σ^2-subgaussian random variables. We would like to estimate the magnitude of the operator norm

$$\|M\| := \sup_{v \in B_2^n, w \in B_2^m} \langle v, Mw \rangle = \sup_{(v,w) \in T} X_{v,w},$$

where $B_2^n = \{x \in \mathbb{R}^n : \|x\| \le 1\}$ is the Euclidean unit ball in $\mathbb{R}^n$ and

$$T := B_2^n \times B_2^m, \qquad X_{v,w} := \langle v, Mw \rangle = \sum_{i=1}^n \sum_{j=1}^m v_i M_{ij} w_j.$$

It follows immediately from Azuma's inequality (Lemma 3.7) that $X_{v,w}$ is σ^2-subgaussian for every $(v, w) \in T$. On the other hand, note that

$$\begin{aligned} |X_{v,w} - X_{v',w'}| &= |\langle v, Mw \rangle - \langle v', Mw' \rangle| \\ &\le |\langle v - v', Mw \rangle| + |\langle v', M(w - w') \rangle| \\ &\le \|v - v'\| \|M\| \|w\| + \|v'\| \|M\| \|w - w'\| \\ &\le \|M\| \{\|v - v'\| + \|w - w'\|\} \end{aligned}$$

for $(v, w) \in T$. If we define a metric on T as

$$d((v, w), (v', w')) := \|v - v'\| + \|w - w'\|,$$

we see that the random process $\{X_{v,w}\}_{(v,w) \in T}$ is Lipschitz for the metric d. Note that the random Lipschitz constant happens to be $\|M\|$, which is in fact the quantity we are trying to control in the first place! This is a rather peculiar situation, but we can nonetheless readily apply Lemma 5.7: this yields

$$\mathbf{E}[\|M\|] \le \varepsilon \mathbf{E}[\|M\|] + \sqrt{2\sigma^2 \log N(T, d, \varepsilon)}$$

for every $\varepsilon > 0$, which we can rearrange to obtain

$$\mathbf{E}[\|M\|] \le \inf_{\varepsilon > 0} \frac{\sigma\sqrt{2}}{1 - \varepsilon} \sqrt{\log N(T, d, \varepsilon)}.$$

What remains is to estimate the covering number. To this end, we must introduce an additional idea that will be of significant importance in the sequel.

How can one construct a *small* ε-net N? The defining property of an ε-net is that every point in T is within a distance at most ε of some point in N. We can always achieve this by choosing a very dense set N. However, if we want $|N|$ to be small, we should intuitively choose the points in N to be as far apart as possible. This motivates the following definition.

Definition 5.11 (ε-packing and packing number). *A set $N \subseteq T$ is called an* ε-packing *of (T, d) if $d(t, t') > \varepsilon$ for every $t, t' \in N$, $t \neq t'$. The largest cardinality of an ε-packing of (T, d) is called the* packing number

$$D(T, d, \varepsilon) := \sup\{|N| : N \text{ is an } \varepsilon\text{-packing of } (T, d)\}.$$

The key idea, which was already hinted at above, is that the notion of packing *dual* to the notion of covering, as is made precise by the following result. This means that we can use covering and packing interchangeably (up to constants). In some cases it is easier to estimate packing numbers than covering numbers, as we will see shortly. On the other hand, we will see in the following chapter that packing numbers arise naturally when we aim to prove *lower* bounds for the suprema of random processes (as opposed to *upper* bounds which are considered exclusively in this chapter).

Lemma 5.12 (Duality between covering and packing). *For every $\varepsilon > 0$*

$$D(T, d, 2\varepsilon) \le N(T, d, \varepsilon) \le D(T, d, \varepsilon).$$

Note that this can indeed be viewed as a form of duality (in the sense of optimization): the packing number is defined in terms of a supremum, but the covering number is defined in terms of an infimum.

Proof. Let D be a 2ε-packing and let N be an ε-net. For every $t \in D$, choose $\pi(t) \in N$ such that $d(t, \pi(t)) \le \varepsilon$. Then for $t \neq t'$, we have

$$2\varepsilon < d(t, t') \le d(t, \pi(t)) + d(\pi(t), \pi(t')) + d(\pi(t'), t') \le 2\varepsilon + d(\pi(t), \pi(t')),$$

which implies $\pi(t) \neq \pi(t')$. Thus $\pi : D \to N$ is one-to-one, and therefore $|D| \le |N|$. This yields the first inequality $D(T, d, 2\varepsilon) \le N(T, d, \varepsilon)$.

To obtain the second inequality, let D be a *maximal* ε-packing of (T, d) (that is, $|D| = D(T, d, \varepsilon)$). We claim that D is necessarily an ε-net. Indeed, suppose this is not the case; then there is a point $t \in T$ such that $d(t, t') > \varepsilon$ for every $t' \in D$. But then $D \cup \{t\}$ must be a ε-packing also, which contradicts the maximality of D. Thus we have $D(T, d, \varepsilon) = |D| \ge N(T, d, \varepsilon)$. □

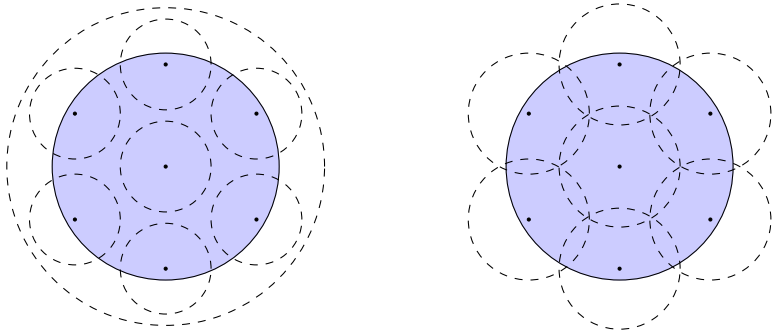
We are now in a position to bound the covering number of the Euclidean ball B_2^n with respect to the Euclidean distance. The proof of this elementary result uses a clever technique known as a *volume argument*.

Lemma 5.13. *We have $N(B_2^n, \|\cdot\|, \varepsilon) = 1$ for $\varepsilon \ge 1$ and*

$$\left(\frac{1}{\varepsilon}\right)^n \le N(B_2^n, \|\cdot\|, \varepsilon) \le \left(\frac{3}{\varepsilon}\right)^n \quad \text{for } 0 < \varepsilon < 1.$$

Proof. That $N(B_2^n, \|\cdot\|, \varepsilon) = 1$ for $\varepsilon \ge 1$ is obvious: by definition, we have $\|t\| = \|t - 0\| \le 1$ for every $t \in B_2^n$, so the singleton $\{0\}$ is an ε-net.

The main part of the proof is illustrated in the following figure:



The colored ball is B_2^n. To obtain an upper bound on the covering number, we choose a 2ε-packing D of B_2^n (black dots in left figure). Then balls of radius ε around $t \in D$ be disjoint, and all these balls are contained in a large ball of size $1+\varepsilon$. As the sum of the volumes of the small balls (of which there are $|D|$) is bounded above by the volume of the large ball, we obtain an upper bound on the size of D (and thus on the covering number by Lemma 5.12). To obtain a lower bound on the covering number, we choose an ε-net N of B_2^n (black dots in right figure). As the balls of radius ε around $t \in N$ cover B_2^n, the sum of the volumes of these balls (of which there are $|N|$) is bounded below by the volume of B_2^n. This yields a lower bound on the size of N.

We now proceed to make this argument precise. Let us begin with the upper bound. Let D be a 2ε-packing of B_2^n. As $d(t,t') > 2\varepsilon$ for all $t \neq t'$ in D, the balls $\{B(t,\varepsilon) : t \in D\}$ must be disjoint. On the other hand, every ball $B(t,\varepsilon)$ for $t \in B_2^n$ must be contained in the larger ball $B(0,1+\varepsilon)$. Thus

$$\sum_{t\in D} \lambda(B(t,\varepsilon)) = \lambda\left(\bigcup_{t\in D} B(t,\varepsilon)\right) \le \lambda(B(0,1+\varepsilon)),$$

where λ denotes the Lebesgue measure on $\mathbb{R}^n$. By homogeneity of the Lebesgue measure, $\lambda(B(t,\alpha)) = \lambda(B(0,\alpha)) = \lambda(\alpha B(0,1)) = \alpha^n \lambda(B(0,1))$. Thus

$$|D| \le \frac{\lambda(B(0,1+\varepsilon))}{\lambda(B(0,\varepsilon))} = \left(\frac{1+\varepsilon}{\varepsilon}\right)^n.$$

As this holds for every 2ε-packing D, we have evidently proved the upper bound $N(T,d,2\varepsilon) \le D(T,d,2\varepsilon) \le (1+1/\varepsilon)^n \le (3/2\varepsilon)^n$ for $2\varepsilon < 1$.

To obtain the lower bound, let N be an ε-net for B_2^n. Then

$$\lambda(B_2^n) \le \lambda\left(\bigcup_{t\in N} B(t,\varepsilon)\right) \le \sum_{t\in N} \lambda(B(t,\varepsilon)),$$

so we obtain

$$|N| \ge \frac{\lambda(B_2^n)}{\lambda(B(0,\varepsilon))} = \left(\frac{1}{\varepsilon}\right)^n.$$

As this holds for every ε-net N, we have proved $N(T,d,\varepsilon) \ge (1/\varepsilon)^n$. □

Remark 5.14. Lemma 5.13 quantifies explicitly the dependence of the covering number on dimension: the number of balls of radius ε needed to cover a ball in $\mathbb{R}^n$ is polynomial in $1/\varepsilon$ of order n. This is not surprising: think of how many cubes of side length ε can fit into the unit cube in $\mathbb{R}^n$. While balls do not pack as nicely as cubes, the ultimate conclusion is the same (in fact, the conclusion of Lemma 5.13 carries over to any norm on $\mathbb{R}^n$, see Problem 5.5). In this manner, the dependence on dimension will enter explicitly into our estimates of the suprema of random processes.

Beyond the concrete result on covering numbers in $\mathbb{R}^n$, Lemma 5.13 provides a good way to think about the notion of dimension in the first place. The classical idea that $\mathbb{R}^n$ is n-dimensional stems from its linear structure: there is a basis of size n such that any vector in $\mathbb{R}^n$ can be written as a linear combination of these basis elements. This linear-algebraic notion of dimension is not very useful in general spaces where one does not need to have any linear structure. Lemma 5.13 motivates a different notion of dimension that makes sense in any metric space: we say that a metric space (T, d) has *metric dimension* n if $N(T, d, \varepsilon) \sim \varepsilon^{-n}$. Lemma 5.13 shows that for (bounded subsets of) $\mathbb{R}^n$, the linear-algebraic and metric notions of dimension coincide; however, the definition of metric dimension is independent of the linear structure of the space. The notion of metric dimension certainly conforms to the intuitive notion that a high-dimensional space has more "room" than a low-dimensional space (the number of balls of fixed radius needed to cover the space increases exponentially in the dimension). Of course, not every metric space has finite metric dimension: we will shortly encounter an infinite-dimensional space (T, d) for which the covering numbers grow exponentially in $1/\varepsilon$.

Having developed some basic estimates, we can now complete the example of random matrices. Here we are not interested in the covering number of B_2^n itself, but rather in the covering number of $T = B_2^n \times B_2^m$ with respect to the metric d. The latter is however easily estimated using Lemma 5.13. Let N be an ε-net for B_2^n and let M be an ε-net for B_2^m. Then $N \times M$ is a 2ε-net for T of cardinality $|N||M|$: indeed, setting $\pi((t, s)) = (\pi(t), \pi(s))$, we have

$$d((t, s), \pi((t, s))) = \|t - \pi(t)\| + \|s - \pi(s)\| \le 2\varepsilon.$$

This evidently implies that

$$N(T, d, 2\varepsilon) \le N(B_2^n, \|\cdot\|, \varepsilon)\, N(B_2^m, \|\cdot\|, \varepsilon) \le \left(\frac{3}{\varepsilon}\right)^{n+m}$$

for $\varepsilon \le 1$. We therefore obtain

$$\mathbf{E}[\|M\|] \le \inf_{\varepsilon > 0} \frac{\sigma\sqrt{2}}{1-\varepsilon} \sqrt{\log N(T, d, \varepsilon)} \lesssim \sigma\sqrt{n+m}.$$

It turns out that this crude bound already captures the correct order of magnitude of the matrix norm! In particular, for square matrices, we obtain $\mathbf{E}[\|M\|] \lesssim \sqrt{n}$ as was already alluded to in Example 2.5.

We now turn to our second example. Unlike in the previous example, where we got a sharp result with little work, we will not be so lucky here: we will derive a nontrivial bound from Lemma 5.7, but the methods we developed so far will prove to be too crude to capture the correct order of magnitude.

Example 5.15 (Wasserstein law of large numbers). Let $X_1, X_2, \ldots$ be i.i.d. random variables with values in the interval $[0,1]$. We denote their distribution as $X_i \sim \mu$. Define the empirical measure of $X_1, \ldots, X_n$ as

$$\mu_n := \frac{1}{n}\sum_{k=1}^n \delta_{X_k}.$$

Then it is easy to estimate

$$\mathbf{E}|\mu_n f - \mu f| \le \mathbf{E}[|\mu_n f - \mu f|^2]^{1/2} \le \frac{\|f\|_\infty}{\sqrt{n}}.$$

In particular, we have $\mu_n f \to \mu f$ in L^1 for every bounded function f: this is none other than the weak law of large numbers with the optimal $n^{-1/2}$ rate.

At what rate the law of large numbers $\mu_n \to \mu$ hold when we consider other notions of distance between probability measures? In this spirit, we will presently attempt to estimate the expected Wasserstein distance $\mathbf{E}[W_1(\mu_n,\mu)]$ between the empirical measure and the underlying distribution. Recall that

$$W_1(\mu_n,\mu) = \sup_{f\in\mathrm{Lip}([0,1])} \{\mu_n f - \mu f\} = \sup_{f\in\mathcal{F}} X_f,$$

where we have defined

$$X_f := \mu_n f - \mu f, \qquad \mathcal{F} := \{f \in \mathrm{Lip}([0,1]) : 0 \le f \le 1\}.$$

Thus this question reduces to controlling the supremum of a random process. (Note that $|f(x)-f(y)| \le |x-y| \le 1$ for $f \in \mathrm{Lip}([0,1])$ and $x,y \in [0,1]$; as X_f is invariant under adding a constant to f, there is no loss of generality in restricting the supremum to functions $0 \le f \le 1$ in the definition of W_1.)

We begin by noting the trivial estimate

$$|X_f - X_g| = |\mu_n(f-g) - \mu(f-g)| \le 2\|f-g\|_\infty.$$

Thus the process $\{X_f\}_{f\in\mathcal{F}}$ is Lipschitz with respect to the uniform distance on $\mathcal{F}$. On the other hand, note that by definition

$$X_f = \sum_{k=1}^n \frac{f(X_k) - \mu f}{n},$$

which is a sum of i.i.d. random variables with values in the interval $[-\frac{1}{n}, \frac{1}{n}]$. Thus X_f is $\frac{1}{n}$-subgaussian for every $f \in \mathcal{F}$ by the Azuma-Hoeffding inequality (Lemma 3.6). We can therefore estimate using Lemma 5.7

$$\mathbf{E}[W_1(\mu_n, \mu)] \le \inf_{\varepsilon > 0} \left\{ 2\varepsilon + \sqrt{\frac{2}{n} \log N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)} \right\}.$$

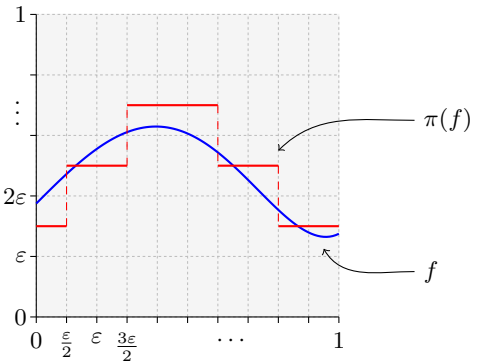
To proceed, we must bound the covering number $N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)$.

Lemma 5.16. *There is a constant $c < \infty$ such that*

$$N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \le e^{c/\varepsilon} \text{ for } \varepsilon < \tfrac{1}{2}, \qquad N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) = 1 \text{ for } \varepsilon \ge \tfrac{1}{2}.$$

Remark 5.17. Note that, unlike in the case of a Euclidean ball where the covering number is polynomial in $1/\varepsilon$, the covering number of the family $\mathcal{F}$ of Lipschitz functions is exponential in $1/\varepsilon$. This indicates that the metric space $(\mathcal{F}, \|\cdot\|_\infty)$ is in fact infinite-dimensional, which is not too surprising.

Proof. Fix $\varepsilon > 0$. For every function $f \in \mathcal{F}$, we will construct a new function $\pi(f)$ in the manner illustrated in the following picture:



To be precise, we approximate $f : [0,1] \to [0,1]$ by $\pi(f) : [0,1] \to [0,1]$ defined as follows. Partition the horizontal axis into consecutive nonoverlapping intervals $I_1, \ldots, I_{\lceil 2/\varepsilon \rceil}$ of size $\varepsilon/2$ and the vertical axis into consecutive nonoverlapping intervals $J_1, \ldots, J_{\lceil 1/\varepsilon \rceil}$ of size ε. We then define

$$\pi(f)(x) = \frac{\max J_\ell + \min J_\ell}{2} \quad \text{whenever} \quad x \in I_k,\ f(\min I_k) \in J_\ell.$$

That is, in each interval on the horizontal axis, we approximate f by its value at the left endpoint of the interval rounded to the center of the interval on the vertical axis to which it belongs. By construction, the set $N = \{\pi(f) : f \in \mathcal{F}\}$ is an ε-net: indeed, note that whenever $x \in I_k$ and $f(\min I_k) \in J_\ell$, we have

$$\begin{aligned} |f(x) - \pi(f)(x)| &\le |f(x) - f(\min I_k)| + \left| f(\min I_k) - \frac{\max J_\ell + \min J_\ell}{2} \right| \\ &\le |x - \min I_k| + \frac{\max J_\ell - \min J_\ell}{2} \le \varepsilon, \end{aligned}$$

where we have used the Lipschitz property of f and the definition of I_k, J_ℓ. (Note that $N \not\subseteq \mathcal{F}$: but this is not a problem, cf. Remark 5.6.)

As we now have an ε-net N, it remains to estimate $|N|$. The most naive bound would be $|N| \le \lceil 1/\varepsilon \rceil^{\lceil 2/\varepsilon \rceil} < \infty$, but we can do somewhat better by taking into account the Lipschitz property of the functions in $\mathcal{F}$. Note that

$$|\pi(f)(\min I_k) - \pi(f)(\min I_{k+1})| \le |f(\min I_k) - f(\min I_{k+1})| + \varepsilon \le \tfrac{3}{2}\varepsilon;$$

As the possible values of $\pi(f)$ can only differ by multiples of ε, this implies that $\pi(f)(\min I_{k+1}) - \pi(f)(\min I_k) \in \{-\varepsilon, 0, \varepsilon\}$. Thus $\pi(f)(0)$ can take any of $\lceil 1/\varepsilon \rceil$ different values, but each subsequent interval can only differ from the previous one in three different ways. This implies the bound

$$N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \le |N| \le \lceil 1/\varepsilon \rceil 3^{\lceil 2/\varepsilon \rceil - 1} \le e^{c/\varepsilon}$$

for some constant c and every $\varepsilon > 0$. On the other hand, as $\|f - \frac{1}{2}\|_\infty \le \frac{1}{2}$ for every $f \in \mathcal{F}$, we clearly have $N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) = 1$ for $\varepsilon \ge \frac{1}{2}$. □

Having estimated the covering numbers of $\mathcal{F}$, we can now readily complete our bound on the convergence rate in the Wasserstein law of large numbers:

$$\mathbf{E}[W_1(\mu_n, \mu)] \le \inf_{\varepsilon > 0} \left\{ 2\varepsilon + \sqrt{\frac{2c}{\varepsilon n}} \right\} \lesssim n^{-1/3}.$$

Recall that the rate of convergence in the law of large numbers for a single function is $\mathbf{E}|\mu_n f - \mu f| \lesssim n^{-1/2}$, but we have obtained a slower rate $n^{-1/3}$ when we consider the convergence uniformly over Lipschitz functions. Is this rate sharp? I turns out that this is not the case: in the present example, we will show in the next section that the optimal rate is actually still $\sim n^{-1/2}$.

Remark 5.18. There is no reason to expect, in general, the the rate of convergence uniformly over a class of functions will be the same as that for a single function. The fact that the rate still turns out to be $n^{-1/2}$ in the present setting is an artefact of the fact that we are working in one dimension: for random variables $X_k \in [0,1]^p$ for $p \ge 2$, the optimal rates turn out to be strictly slower than $n^{-1/2}$. Nonetheless, even in this case, the method we have used in this section does not capture the correct rate of convergence.

The method that we have used in this section to control the suprema or random processes is too crude to obtain sharp results in most examples of interest. While we obtained a sharp result in the random matrix example, this was not the case for the Wasserstein law of large numbers. Unfortunately, the situation encountered in the second example is the norm. It is illuminating to understand in what part of the proof we incurred the loss of precision: this will directly motivate the more powerful approach for bounding the suprema of random processes that will be developed in the next section.

The approach of Lemma 5.7 relies on two steps: the approximation of the supremum by a finite maximum, and the estimation of the finite maximum

using a suitable maximal inequlity. The key problem with this approach is that we have approximated the supremum by a maximum in an extremely inefficient manner by using an *almost sure* Lipschitz property of the process. Let us illustrate this in the second example. Here the Lipschitz property reads

$$|X_f - X_g| \le 2\|f - g\|_\infty \quad \text{a.s.}$$

One cannot substantially improve on this bound if the result is required to hold almost surely. On the other hand, we can easily compute

$$\mathbf{E}|X_f - X_g| \le n^{-1/2}\|f - g\|_\infty.$$

While the almost sure Lipschitz constant of the process X_f is 2, we see that X_f is Lipschitz on average with Lipschitz constant $n^{-1/2} \ll 2$: that is, the *typical* behavior of the increments $|X_f - X_g|$ is much better than their *worst-case* behavior! One can therefore readily understand why using the almost sure Lipschitz property incurs a significant loss in our estimates. If we were to naively substitute the "typical" Lipschitz constant $n^{-1/2}$ rather than the "worst-case" constant 2 in the above computation, we would indeed obtain the correct $n^{-1/2}$ rather than $n^{-1/3}$ rate. However, the almost sure Lipschitz property was crucial in order to control the approximation error in Lemma 5.7, so that such a substitution is certainly unjustified at this point.

Remark 5.19. We can now also understand why the crude approach of Lemma 5.7 proves to be useful in the random matrix example: in this setting, it so happens that the almost sure Lipschitz constant is of the same order as the supremum that we are trying to compute. Therefore, even though our approximation is inefficient, this does not affect the final bound except in the numerical constant. However, this situation is essentially a coincidence. In the Wasserstein law of large numbers example, the almost sure Lipschitz constant is much larger than the supremum of interest, so that the inefficiency in our approximation swamps the final bound that we obtain.

The basic challenge we therefore face at this point in improving the approach of Lemma 5.7 is to devise a method of approximation that only uses an "in probability" version of the Lipschitz property that can capture the typical size of the increments, rather than an a.s. Lipschitz property that captures the worst case. In the next section, we will see that this goal can be accomplished by using a powerful technique known as *chaining*.

Problems

5.4 (Tightness of Johnson-Lindenstrauss). The Johnson-Lindenstrauss lemma proved in Problem 5.3 shows that any n points in a Hilbert space H can be mapped into $\mathbb{R}^k$ with $k \gtrsim \log n$ while distorting the distances between them by at most a constant factor. Show that $k \gtrsim \log n$ is in fact necessary.

Hint: show that the image of n orthonormal vectors $x_1, \ldots, x_n$ in H under a map $T : H \to \mathbb{R}^k$ that nearly preserves distances is a packing of a ball in $\mathbb{R}^k$.

5.5 (Covering norm-balls in $\mathbb{R}^n$). The goal of this problem is to investigate Lemma 5.13 for norms other than the Euclidean norm.

a. Show that the conclusion of Lemma 5.13 holds in any finite-dimensional Banach space: that is, if $|\cdot|$ is any norm on $\mathbb{R}^n$, then we have

$$\left(\frac{1}{\varepsilon}\right)^n \le N(B, |\cdot|, \varepsilon) \le \left(\frac{3}{\varepsilon}\right)^n \quad \text{for } 0 < \varepsilon < 1,$$

where B denotes the unit norm-ball $\{x \in \mathbb{R}^n : |x| \le 1\}$.

b. Show that in the special case $n = 1$, we can compute exactly

$$N(B, |\cdot|, \varepsilon) = \left\lceil \frac{1}{\varepsilon} \right\rceil.$$

5.6 (Proper covering numbers). In our definition of an ε-net N for (T, d), we did not assume that $N \subseteq T$ (cf. Remark 5.6). It can happen quite naturally that the points that we use to approximate the set T are not themselves in T, for example, see the proof of Lemma 5.16. On the other hand, in some applications, it may be convenient to require that $N \subseteq T$. When this is the case, the ε-net is said to be *proper*, and the *proper covering number* $N_{\mathrm{pr}}(T, d, \varepsilon)$ denotes the cardinality of the smallest proper ε-net. Show that

$$N(T, d, \varepsilon) \le N_{\mathrm{pr}}(T, d, \varepsilon) \le N(T, d, \varepsilon/2),$$

which implies that the assumption of properness is harmless in most cases.

5.7 (Parametric classes). Consider a function $f : \Theta \times X \to \mathbb{R}$ such that

$$|f_\theta(x) - f_{\theta'}(x)| \le C d(\theta, \theta') \quad \text{for all } x \in X$$

for some metric d on Θ. We think of $x \mapsto f_\theta(x)$ as a function on X that is parametrized by a parameter $\theta \in \Theta$. Thus it makes sense to define

$$\mathcal{F} = \{f_\theta : \theta \in \Theta\}.$$

Show that

$$N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \le N(\Theta, d, \varepsilon/C).$$

Thus the covering numbers of parametrized classes of functions that are Lipschitz in the parameter can be controlled by the covering numbers of the parameter space. This is often useful, for example, in parametric statistics.

5.8 (Wasserstein LLN in higher dimension). The goal of this problem is to extend Example 5.15 to the multidimensional situation where $X_1, X_2, \ldots$ are i.i.d. random variables with values in the cube $[0, 1]^d$.

a. Let $\mathcal{F}_0 := \{f \in \mathrm{Lip}([0, 1]^d) : f(0) = 0\}$. Show that

$$N(\mathcal{F}_0, \|\cdot\|_\infty, \varepsilon) \le e^{c/\varepsilon^d},$$

where the constant c depends on dimension d only.

b. What upper bound on the rate in the Wasserstein law of large numbers in dimension d does this imply using the crude method of Lemma 5.7?

5.3 The chaining method

In the previous section, we developed a simple method to bound the supremum of a random process that satisfies the Lipschitz property $X_t - X_s \lesssim d(t,s)$ in an *almost sure* sense. However, we have seen that this requirement is very restrictive: in many cases, the typical size of the increments $X_t - X_s$ is much smaller than in the worst case. We therefore aim to develop a method to bound the suprema of random processes that only requires the Lipschitz property $X_t - X_s \lesssim d(t,s)$ to hold *in probability* in a suitable sense.

To understand how one might approach this problem, let us recall the basic idea behind the proof of Lemma 5.7. If N is an ε-net, we can estimate

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \le \mathbf{E}\left[\sup_{t\in T} X_{\pi(t)}\right] + \mathbf{E}\left[\sup_{t\in T}\{X_t - X_{\pi(t)}\}\right].$$

The first term is a finite maximum that can be controlled by the maximal inequality of Lemma 5.1. The second term is a small remainder: each variable inside the supremum has magnitude of order ε by the Lipschitz property of the process. If the Lipschitz property holds in an almost sure sense, the supremum drops out and we can immediately control the remainder term.

However, if the Lipschitz property only holds in probability, we cannot directly control the remainder term. Indeed, in this case each variable inside the supremum has "typical"size ε; however, we have to control the supremum of many such variables, whose magnitude can be much larger than ε (e.g., the maximum of n independent $N(0,\sigma^2)$ variables is of order $\sigma\sqrt{\log n} \gg \sigma$, even though each variable is only of order σ). Therefore, in this case, the problem of controlling the remainder term is essentially of the same type as that of controlling the original supremum of interest. Nonetheless, we expect that the remainder term is smaller than the original supremum, as the size of each variable in the remainder term is now smaller. To shrink the remainder term further, we can approximate it once again by a finite maximum at a smaller scale. For example, if N' is an $\varepsilon/2$-net, then we can estimate

$$\mathbf{E}\left[\sup_{t\in T}\{X_t - X_{\pi(t)}\}\right] \le \mathbf{E}\left[\sup_{t\in T}\{X_{\pi'(t)} - X_{\pi(t)}\}\right] + \mathbf{E}\left[\sup_{t\in T}\{X_t - X_{\pi'(t)}\}\right].$$

The first term on the right is a finite maximum that can be controlled by Lemma 5.1. The remainder term is still an infinite supremum, but now each variable inside the supremum is only of order $\varepsilon/2$: that is, we have cut the remainder term roughly by half. The key idea of this section is that we can repeat this procedure over and over again, each time cutting the size of the remainder term roughly by half. Let us investigate this idea a bit more systematically. For each $k \ge 0$, let N_k be a 2^{-k}-net and choose $\pi_k(t) \in N_k$ such that $d(t,\pi_k(t)) \le 2^{-k}$. Repeating the approximation n times, we obtain

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \le \mathbf{E}\left[\sup_{t\in T} X_{\pi_0(t)}\right] + \sum_{k=1}^{n} \mathbf{E}\left[\sup_{t\in T}\{\overbrace{X_{\pi_k(t)} - X_{\pi_{k-1}(t)}}^{\sim 2^{-k}}\}\right] + \mathbf{E}\left[\sup_{t\in T}\{\overbrace{X_t - X_{\pi_n(t)}}^{\sim 2^{-n}}\}\right].$$

The remainder term is now a supremum of variables of order 2^{-n}. Under mild conditions, the remainder term will disappear if we let $n \to \infty$ *without having to invoke any almost sure Lipschitz property of the process.* Thus we surmount the inefficiency of Lemma 5.7 by approximating the supremum not at a single scale, but at infinitely many scales. The remaining bound is now an infinite sum: the kth term in the sum is a finite maximum of random variables at the scale 2^{-k}. To control these finite maxima, we also do not require an almost sure Lipschitz property: in view of Lemma 5.1, it suffices to assume that the Lipschitz property holds "in probability" in the following sense.

Definition 5.20 (Subgaussian process). *A random process $\{X_t\}_{t\in T}$ on the metric space (T, d) is called* subgaussian *if $\mathbf{E}[X_t] = 0$ and*

$$\mathbf{E}[e^{\lambda\{X_t - X_s\}}] \le e^{\lambda^2 d(t,s)^2/2} \quad \textit{for all } t, s \in T,\ \lambda \ge 0.$$

Remark 5.21. The subgaussian property should indeed be interpreted as an "in probability" form of the Lipschitz property: by Problem 3.1, the subgaussian assumption is equivalent up to constants to an assumption of the form

$$\mathbf{P}[|X_t - X_s| \ge x\, d(t,s)] \le Ce^{-x^2/C}.$$

Note also that the assumption $\mathbf{E}[e^{\lambda\{X_t - X_s\}}] \le e^{\lambda^2 d(t,s)^2/2}$ already implies $\mathbf{E}[X_t - X_s] = 0$ (as $\lim_{\lambda\downarrow 0}\{e^{c\lambda^2/2} - 1\}/\lambda = 0$), so the assumption $\mathbf{E}[X_t] = 0$ merely imposes a convenient normalization. In section 5.4, we will see how to control the suprema of random processes with nontrivial mean $t \mapsto \mathbf{E}[X_t]$.

The technique that we have outlined above is known as *chaining*: the idea is to approximate X_t by a "chain" $X_{\pi_k(t)}$ of increasingly accurate approximations (the "links" in the chain are the increments $X_{\pi_k(t)} - X_{\pi_{k-1}(t)}$). The main remaining difficulty in implementing the method is to show that the remainder term does indeed vanish as $n \to \infty$. To get around this, we will impose a very mild technical assumption that holds in almost all cases of interest.

Definition 5.22 (Separable process). *A random process $\{X_t\}_{t\in T}$ is called* separable *if there is a countable set $T_0 \subseteq T$ such that*

$$X_t \in \lim_{\substack{s\to t\\ s\in T_0}} X_s \quad \textit{for all } t \in T \quad \textit{a.s.}$$

[*Here $x \in \lim_{s\to t} x_s$ means that there is a sequence $s_n \to t$ such that $x_{s_n} \to x$.*]

Remark 5.23. The assumption of separability is technical, and is almost always trivially satisfied. For example, if $t \mapsto X_t$ is continuous a.s., we can take T_0 to be any countable dense subset of T. At the same time, the separability assumption is in some sense intrinsic to the chaining argument. After all, the main idea of the chaining argument is to approximate $X_t = \lim_{k\to\infty} X_{\pi_k(t)}$ for every $t \in T$. If this is in fact valid, however, then the definition of a separable process will hold for the countable set $T_0 = \{\pi_k(t) : k \geq 0,\ t \in T\}$.

For completeness, let us note a somewhat esoteric point that we swept under the rug. If T is uncountable, $\sup_{t\in T} X_t$ is the supremum of an uncountable family of random variables. In general, the supremum of uncountably many measurable functions is not even necessarily measurable. Measurability issues do arise, on occasion, in the control of suprema, but we will shamelessly ignore such problems in these notes. Under the separability assumption, however, $\sup_{t\in T} X_t = \sup_{t\in T_0} X_t$ a.s., and thus no measurability problems arise (as a countable supremum of measurable functions is always measurable).

We now have all the ingredients to implement the chaining argument.

Theorem 5.24 (Dudley). *Let $\{X_t\}_{t\in T}$ be a separable subgaussian process on the metric space (T, d). Then we have the following estimate:*

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \leq 6\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log N(T, d, 2^{-k})}.$$

Proof. We first prove the result in the finite case $|T| < \infty$, which allows us to easily eliminate the remainder term in the chaining argument. We subsequently use the separability assumption to lift this restriction.

Let $|T| < \infty$. Let k_0 be the largest integer such that $2^{-k_0} \geq \mathrm{diam}(T)$. Then any singleton $N_{k_0} = \{t_0\}$ is trivially a 2^{-k_0}-net. We therefore start chaining at the scale 2^{-k_0}. For $k > k_0$, let N_k be a 2^{-k}-net such that $|N_k| = N(T, d, 2^{-k})$. Running the chaining argument up to the scale 2^{-n} yields

$$\begin{aligned}\mathbf{E}\left[\sup_{t\in T} X_t\right] \leq \mathbf{E}[X_{t_0}] &+ \sum_{k=k_0+1}^{n} \mathbf{E}\left[\sup_{t\in T}\{X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\}\right] \\ &+ \mathbf{E}\left[\sup_{t\in T}\{X_t - X_{\pi_n(t)}\}\right].\end{aligned}$$

Let us consider each of the terms. As $\mathbf{E}[X_{t_0}] = 0$ by assumption, the first term disappears. Moreover, as $|T| < \infty$, we can choose n sufficiently large so that $N_n = T$. Then the last term disappears. To control the terms inside the sum, note that the maximum in the kth term contains at most $|N_k||N_{k-1}| \leq |N_k|^2$ terms (as $|N_{k-1}| \leq |N_k|$). Moreover, we can readily estimate

$$d(\pi_k(t), \pi_{k-1}(t)) \leq d(t, \pi_k(t)) + d(t, \pi_{k-1}(t)) \leq 3 \times 2^{-k}.$$

As $X_{\pi_k(t)} - X_{\pi_{k-1}(t)}$ is $d(\pi_k(t), \pi_{k-1}(t))^2$-subgaussian, Lemma 5.1 yields

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \le 6\sum_{k>k_0} 2^{-k}\sqrt{\log|N_k|}.$$

But $|N_k| = N(T,d,2^{-k})$ by construction, so the proof is complete.

In the proof we have used the assumption $|T| < \infty$ to control the remainder term in the chaining argument. We now use separability to show that one can approximate the general case by the finite case. Indeed, by separability, there is a countable subset $T' \subseteq T$ such that $\sup_{t\in T} X_t = \sup_{t\in T'} X_t$ a.s. Denote by T_k the first k elements of T' (in arbitrary order). Then

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] = \mathbf{E}\left[\sup_{t\in T'} X_t\right] = \sup_{k\ge 1}\mathbf{E}\left[\sup_{t\in T_k} X_t\right]$$

by monotone convergence. Applying the chaining inequality to each finite maximum and using $N(T_k,d,\varepsilon) \le N(T,d,\varepsilon)$ yields the general result. □

Very often the result of Theorem 5.24 is written in a slightly different form by noting that the sum can be viewed as a Riemann sum approximation to a certain integral. There is no particular mathmatical significance to this reformulation: it is made for purely aesthetic reasons.

Corollary 5.25 (Entropy integral). *Let $\{X_t\}_{t\in T}$ be a separable subgaussian process on the metric space (T,d). Then we have the following estimate:*

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \le 12\int_0^\infty \sqrt{\log N(T,d,\varepsilon)}\,d\varepsilon.$$

Proof. We can readily estimate

$$\begin{aligned}\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log N(T,d,2^{-k})} &= 2\sum_{k\in\mathbb{Z}}\int_{2^{-k-1}}^{2^{-k}} \sqrt{\log N(T,d,2^{-k})}\,d\varepsilon\\ &\le 2\sum_{k\in\mathbb{Z}}\int_{2^{-k-1}}^{2^{-k}} \sqrt{\log N(T,d,\varepsilon)}\,d\varepsilon\\ &= 2\int_0^\infty \sqrt{\log N(T,d,\varepsilon)}\,d\varepsilon,\end{aligned}$$

where we used that $N(T,d,\varepsilon)$ is decreasing in ε. □

Remark 5.26. It is important to note that we always have $N(T,d,\varepsilon) = 1$ when $\varepsilon \ge \operatorname{diam}(T)$, as in this case any singleton $N = \{t_0\}$ is trivially an ε-net. Thus it suffices to take integral in Corollary 5.25 only up to $\varepsilon = \operatorname{diam}(T)$.

Remark 5.27. The logarithm of the covering number $\log N(T,d,\varepsilon)$ is often called *metric entropy* in analogy with information theory: it measures the number of bits needed to specify an element of T up to precision ε. It is customary to refer to the integral in Corollary 5.25 as the *entropy integral*.

To illustrate Corollary 5.25, let us revisit Example 5.15.

Example 5.28 (Wasserstein law of large numbers revisited). We adopt the same setting and notation as in Example 5.15. Recall that we want to estimate the expected Wasserstein distance between the empirical and true measures

$$W_1(\mu_n, \mu) = \sup_{f \in \mathcal{F}} X_f,$$

where $X_1, X_2, \ldots$ are i.i.d. variables in $[0,1]$ with distribution μ and

$$X_f = \sum_{k=1}^{n} \frac{f(X_k) - \mu f}{n}, \qquad \mathcal{F} = \{f \in \mathrm{Lip}([0,1]) : 0 \le f \le 1\}.$$

By the Azuma-Hoeffding inequality (Corollary 3.9), we have

$$\mathbf{E}[e^{\lambda\{X_f - X_g\}}] \le e^{\lambda^2 \|f-g\|_\infty^2/2n}.$$

The process $\{X_f\}_{f\in\mathcal{F}}$ is therefore subgaussian with respect to the metric $d(f,g) = n^{-1/2}\|f-g\|_\infty$. We can consequently estimate using Corollary 5.25

$$\mathbf{E}[W_1(\mu_n,\mu)] \le 12 \int_0^\infty \sqrt{\log N(\mathcal{F}, n^{-1/2}\|\cdot\|_\infty, \varepsilon)}\, d\varepsilon.$$

But it is easily seen that

$$N(\mathcal{F}, n^{-1/2}\|\cdot\|_\infty, \varepsilon) = N(\mathcal{F}, \|\cdot\|_\infty, n^{1/2}\varepsilon),$$

so that changing variables in the integral and using Lemma 5.16 yields

$$\mathbf{E}[W_1(\mu_n,\mu)] \le \frac{12}{\sqrt{n}} \int_0^\infty \sqrt{\log N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)}\, d\varepsilon \le \frac{12}{\sqrt{n}} \int_0^{\frac{1}{2}} \sqrt{\frac{c}{\varepsilon}}\, d\varepsilon.$$

As $\varepsilon^{-1/2}$ is integrable at the origin, we have proved

$$\mathbf{E}[W_1(\mu_n,\mu)] \lesssim n^{-1/2},$$

which is a huge improvement over the $n^{-1/3}$ rate obtained by the crude method used in Example 5.15. It is evident from the above computations that the crucial improvement is due to the fact that $|X_f - X_g| \lesssim n^{-1/2}\|f-g\|_\infty$ in probability (as is made precise by the subgaussian property), while the best almost sure Lipschitz bound one can hope for is $|X_f - X_g| \lesssim \|f-g\|_\infty$.

In the present example, it is rather easy to obtain a matching lower bound on the Wasserstein distance. Indeed, note that for any function $f \in \mathcal{F}$ that is not constant μ-a.s., we obtain by the central limit theorem

$$\mathbf{E}[W_1(\mu_n,\mu)] \ge \mathbf{E}[X_f \vee X_{1-f}] = \mathbf{E}|X_f| \sim n^{-1/2}.$$

Thus the rate we obtained by chaning is sharp in the present setting.

Now that we understand the chaining principle, we can use it to obtain more sophisticated results. For example, just as we could obtain a tail bound in Lemma 5.2 corresponding to the maximal inequality of Lemma 5.1, we can obtain a tail bound counterpart to Corollary 5.25.

Theorem 5.29 (Chaining tail inequality). *Let $\{X_t\}_{t\in T}$ be a separable subgaussian process on the metric space (T,d). Then for all $t_0 \in T$ and $x \geq 0$*

$$\mathbf{P}\left[\sup_{t\in T}\{X_t - X_{t_0}\} \geq C\int_0^\infty \sqrt{\log N(T,d,\varepsilon)}\,d\varepsilon + x\right] \leq Ce^{-x^2/C\,\mathrm{diam}(T)^2},$$

where $C < \infty$ is a universal constant.

Proof. The beginning of the proof is identical to that of Theorem 5.24, and we adopt the notations used there. As in Theorem 5.24, it is easily seen that it suffices to consider $|T| < \infty$, as we will assume in the remainder of the proof.

The idea here is to run the chaining argument without taking the expectation. As $|T| < \infty$, we have $\pi_n(t) = t$ for n sufficiently large. Thus

$$X_t - X_{t_0} = \sum_{k>k_0}\{X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\}$$

by the telescoping property of the sum. This elementary *chaining identity* lies at the heart of the chaining argument. We immediately obtain

$$\sup_{t\in T}\{X_t - X_{t_0}\} \leq \sum_{k>k_0}\sup_{t\in T}\{X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\}.$$

Rather than bounding the expectation of this quantity, as we did in Theorem 5.24, we will bound the tail behavior of every term in this sum. To this end, note that the subgaussian property of $\{X_t\}_{t\in T}$ and Lemma 5.2 yield

$$\mathbf{P}\left[\sup_{t\in T}\{X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\} \geq 6\times 2^{-k}\sqrt{\log|N_k|} + 3\times 2^{-k}z\right] \leq e^{-z^2/2}.$$

Thus with high probability, every link $X_{\pi_k(t)} - X_{\pi_{k-1}(t)}$ at the scale k is small. We would like to show that all links at *every* scale are small simultaneously, that is, that the probability of the union over all k of the events in the above bound is small. We can use a crude union bound to control the latter probability, but it is clear that we must then choose z to be increasing in such a way that the probabilities of the individual events are summable: that is,

$$\begin{aligned}\mathbf{P}[\Omega] &:= \mathbf{P}\left[\exists\, k > k_0 \text{ s.t. } \sup_{t\in T}\{X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\} \geq 6\,2^{-k}\sqrt{\log|N_k|} + 3\,2^{-k}z_k\right]\\ &\leq \sum_{k>k_0}\mathbf{P}\left[\sup_{t\in T}\{X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\} \geq 6\,2^{-k}\sqrt{\log|N_k|} + 3\,2^{-k}z_k\right]\\ &\leq \sum_{k>k_0} e^{-z_k^2/2}.\end{aligned}$$

How to choose z_k is not so important. An easy choice $z_k = x + \sqrt{k - k_0}$ yields

$$\mathbf{P}[\Omega] \le \sum_{k>k_0} e^{-z_k^2/2} \le e^{-x^2/2} \sum_{k>0} e^{-k/2} \le C e^{-x^2/2}.$$

Now note that on the event Ω^c, we have

$$\begin{aligned}\sup_{t\in T}\{X_t - X_{t_0}\} &\le \sum_{k>k_0} \sup_{t\in T}\{X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\} \\ &\le 6 \sum_{k>k_0} 2^{-k}\sqrt{\log|N_k|} + 3\,2^{-k_0} \sum_{k>0} 2^{-k}\sqrt{k} + 3\,2^{-k_0} \sum_{k>0} 2^{-k}\,x \\ &\le C\int_0^\infty \sqrt{\log N(T,d,\varepsilon)}\,d\varepsilon + C\,\mathrm{diam}(T)\,x,\end{aligned}$$

where we have used that $2^{-k_0} \le 2\,\mathrm{diam}(T)$ and

$$2^{-k_0} \le C 2^{-k_0-1}\sqrt{\log N(T,d,2^{-k_0-1})} \le C \sum_{k>k_0} 2^{-k}\sqrt{\log|N_k|}$$

by the definition of k_0. Thus

$$\mathbf{P}\left[\sup_{t\in T}\{X_t - X_{t_0}\} \ge C\int_0^\infty \sqrt{\log N(T,d,\varepsilon)}\,d\varepsilon + C\,\mathrm{diam}(T)\,x\right] \le \mathbf{P}[\Omega],$$

and the proof is readily completed. □

Remark 5.30. Note that the result of Theorem 5.29 is reminiscent of a concentration inequality. Indeed, if we could establish the concentration inequality

$$\mathbf{P}\left[\sup_{t\in T}\{X_t - X_{t_0}\} \ge \mathbf{E}\left[\sup_{t\in T}\{X_t - X_{t_0}\}\right] + x\right] \le C e^{-x^2/C\,\mathrm{diam}(T)^2},$$

then the conclusion of Theorem 5.29 would follow directly by combining this inequality with the chaining bound of Corollary 5.25 for the expected supremum. Despite the similarities, however, Theorem 5.29 should not be confused with a concentration inequality. Its conclusion is both weaker and stronger: weaker, because Theorem 5.29 cannot establish a deviation inequality from the mean, but only from a particular upper bound on the mean; stronger, because the subgaussian assumption of Theorem 5.29 is much weaker than would be required to establish a concentration inequality.

The proof of Theorem 5.29 suggests that at its core, the chaining method boils down to simultaneously controlling, using a union bound, the magnitude of all the links $X_{\pi_k(t)} - X_{\pi_{k-1}(t)}$ in the chaining identity. We might therefore expect that chaining yields sharp results if the links $\{X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\}_{t\in T, k>k_0}$ are "nearly independent" in some sense. This is not entirely implausible, as two links are either far apart or are at a different scale. It turns out that the

chaining method that we have developed here yields sharp results in many cases, but falls short in others. In the next chapter, we will see that the chaining method can be further improved to adapt to the structure of the set T. The resulting method, called the *generic chaining*, is so efficient that it captures exactly (up to universal constants) the magnitude of the supremum of Gaussian processes! Once this has been understood, we can truly conclude that chaining is the "correct" way to think about the suprema of random processes. Nonetheless, considering that we have ultimately used no idea more sophisticated than the union bound, the remarkably far-reaching power of the chaining method remains somewhat of a miracle to this author.

Problems

5.9 (The entropy integral and sum). Show that

$$\int_0^\infty \sqrt{\log N(T,d,\varepsilon)}\,d\varepsilon \le \sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log N(T,d,2^{-k})} \le 2\int_0^\infty \sqrt{\log N(T,d,\varepsilon)}\,d\varepsilon.$$

Thus nothing is lost in expressing the chaining bound as an integral rather than a sum, as we have done in Corollary 5.25, up to a constant factor.

5.10 (Chaining with arbitrary tails). The chaining method is not restricted to subgaussian processes: it can be developed analogously for processes that are Lipschitz "in probability" in a more general sense.

Let $\{X_t\}_{t\in T}$ be a separable process with $\mathbf{E}[X_t]=0$ and

$$\log \mathbf{E}[e^{\lambda\{X_t-X_s\}/d(t,s)}] \le \psi(\lambda) \quad \text{for all } t,s\in T,\ \lambda\ge 0,$$

where ψ is as in Lemma 5.1. Show that

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \lesssim \int_0^\infty \psi^{*-1}(2\log N(T,d,\varepsilon))\,d\varepsilon.$$

5.11 (An improved chaining bound and Wasserstein LLN). The key improvement of the chaining bound of Corollary 5.25 over the crude approximation of Lemma 5.7 is that the former uses only an *in probability* Lipschitz property, while the latter uses a stronger *almost sure* Lipschitz property. These two ideas are not mutually exclusive, however: when the process $\{X_t\}_{t\in T}$ satisfies both types of Lipschitz property, we can obtain an improved chaining bound that is a sort of hybrid between Corollary 5.25 and Lemma 5.7.

a. Prove the following theorem.

Theorem 5.31 (Improved chaining). *Let $\{X_t\}_{t\in T}$ be a separable process that is both subgaussian (Definition 5.20) and almost surely Lipschitz (Definition 5.4). Then we have the following estimate:*

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \le \inf_{\varepsilon>0}\left\{2\varepsilon\mathbf{E}[C] + 12\int_\varepsilon^\infty \sqrt{\log N(T,d,\varepsilon)}\,d\varepsilon\right\}.$$

Hint: run the chaining argument only up to scale 2^{-n} and use the almost sure Lipschitz property to estimate the remainder term.

To understand the advantage of Theorem 5.31, we first note the following.

b. Show that $N(T,d,\varepsilon)$ diverges as $\varepsilon \downarrow 0$ whenever $|T| = \infty$.

As the covering number diverges, a nontrivial application of Corollary 5.25 requires that this divergence is sufficiently slow that $\sqrt{\log N(T,d,\varepsilon)}$ is integrable at zero. This is not always the case. On the other hand, Lemma 5.7 would give a nontrivial bound even when the covering number is not integrable, but the use of the almost sure Lipschitz property yields a very pessimistic bound. Theorem 5.31 provides the best of both worlds: it uses the "in probability" Lipschitz property as much as possible, while using the almost sure Lipschitz property to cut off the divergent part of the integral.

To illustrate the efficiency of Theorem 5.31, let us revisit once more the Wasserstein law of large numbers. We have resolved completely the rate of convergence in one dimension in Example 5.28. However, in higher dimensions, we have so far only obtained pessimistic rates in Problem 5.8.

c. Show that we cannot obtain any nontrivial bound for the Wasserstein law of large numbers in dimensions $d \geq 2$ from Corollary 5.25.

d. Using Theorem 5.31, show that in the setting of Problem 5.8

$$\mathbf{E}[W_1(\mu_n,\mu)] \lesssim \begin{cases} n^{-1/2} & \text{for } d = 1, \\ n^{-1/2}\log n & \text{for } d = 2, \\ n^{-1/d} & \text{for } d \geq 3. \end{cases}$$

Unlike in the one-dimensional case, a lower bound (and hence the sharpness of the above estimates for the rates) is not immediately obvious in dimensions $d \geq 2$. We must work a little bit harder to obtain some insight.

e. Suppose that $\mu(dx) = \rho(x)dx$ with $\|\rho\|_\infty < \infty$. Show that

$$\mathbf{E}\left[\min_{i=1,\ldots,n} \|x - X_i\|\right] \gtrsim n^{-1/d} \quad \text{for all } x \in [0,1]^d.$$

Hint: use $\mathbf{P}[\min_{i\leq n} \|x - X_i\| \geq t] = \mathbf{P}[\|x - X_1\| \geq t]^n$ and integrate.

f. Conclude that when μ has a bounded density, we have in any dimension d

$$\mathbf{E}[W_1(\mu_n,\mu)] \gtrsim n^{-1/d}.$$

Hint: consider the (random) function $f(x) = -\min_{i\leq n} \|x - X_i\|$.

Taking together all the upper and lower bounds that we have proved for the Wasserstein law of large numbers, we have evidently obtained sharp rates $\sim n^{-1/2}$ in dimension $d = 1$ and $\sim n^{-1/d}$ in dimension $d \geq 3$. The only case

still in question is dimension $d = 2$, where there remains a gap between our lower and upper bounds $n^{-1/2} \lesssim \mathbf{E}[W_1(\mu_n, \mu)] \lesssim n^{-1/2} \log n$. It turns out that neither bound is sharp in this case: the correct rate is $\sim n^{-1/2}(\log n)^{1/2}$. It has been shown by Talagrand that this rather deep result, due to Ajtai, Komlós, and Tusnády, can be derived (in a nontrivial manner) using the more sophisticated generic chaining method that will be developed in Chapter 6.

5.4 Penalization and the slicing method

Up to this point we have considered the suprema of subgaussian processes, which are necessarily centered $\mathbf{E}[X_t] = 0$ (or at least $\mathbf{E}[X_t - X_s] = 0$ for all t, s). It is often of interest, however, to consider random processes that have nontrivial mean behavior $t \mapsto \mathbf{E}[X_t]$. To this end, let us decompose

$$X_t = \mathbf{E}[X_t] + Z_t$$

in terms of its mean $\mathbf{E}[X_t]$ and fluctuations $Z_t = X_t - \mathbf{E}[X_t]$. It is natural to assume that the fluctuations $\{Z_t\}_{t\in T}$ form a subgaussian process. As

$$\sup_{t\in T} X_t = \sup_{t\in T}\{Z_t + \mathbf{E}[X_t]\},$$

the problem of controlling the supremum of $\{X_t\}_{t\in T}$ can evidently be interpreted as the problem of controlling the *penalized* supremum of a subgaussian process, where $\mathbf{E}[X_t]$ plays the role of the penalty. The chaining method is well suited to controlling the fluctuations, but not to controlling the penalty. The aim of this section is to develop a technique, called the *slicing method*, that reduces the problem of controlling a penalized supremum of a subgaussian process to controlling a subgaussian process without penalty. As penalized suprema arise in many settings, the slicing method is an important part of the toolbox needed to control the suprema of random processes.

There is, in fact, nothing special about the specific additive form of the penalty: the slicing method will prove to be useful in other cases as well. For example, in various situations it is of interest to control a *weighted* supremum

$$\sup_{t,s\in T} \frac{X_t - X_s}{\rho(t,s)}$$

of a subgaussian process $\{X_t\}_{t\in T}$ for some suitable function ρ that should be viewed as a multiplicative (rather than additive) penalty. One could of course view $X_{t,s} = \{X_t - X_s\}/\rho(t,s)$ as a new stochastic process whose supremum we wish to compute, but it is generally far from clear that this process is subgaussian with respect to a natural distance. In such situations, the slicing method will once again provide an important tool to handle the penalty.

Let us illustrate the basic idea behind the slicing method in the multiplicative setting (the additive setting works much in the same way). Fix a sequence $\alpha_k \downarrow 0$ such that $\rho(s,t) \le \alpha_0$ for all s, t. Then we can evidently write

$$\mathbf{P}\left[\sup_{s,t\in T}\frac{X_t-X_s}{\rho(t,s)}\ge x\right]=\mathbf{P}\left[\sup_{k\ge1}\sup_{\alpha_k\le\rho(s,t)\le\alpha_{k-1}}\frac{X_t-X_s}{\rho(t,s)}\ge x\right].$$

That is, we decompose the supremum over "slices" $\{(s,t) : \alpha_k \le \rho(s,t) \le \alpha_{k-1}\}$ of the index set $T \times T$. The key point is that on each slice, the penalty is controlled both from above and from below, so that it can be eliminated from the supremum. We can therefore estimate, using a union bound,

$$\begin{aligned}\mathbf{P}\left[\sup_{s,t\in T}\frac{X_t-X_s}{\rho(t,s)}\ge x\right]&\le\sum_{k=1}^{\infty}\mathbf{P}\left[\sup_{\alpha_k\le\rho(s,t)\le\alpha_{k-1}}\frac{X_t-X_s}{\rho(t,s)}\ge x\right]\\&\le\sum_{k=1}^{\infty}\mathbf{P}\left[\sup_{\rho(s,t)\le\alpha_{k-1}}\{X_t-X_s\}\ge\alpha_k x\right].\end{aligned}$$

Each probability inside the sum on the right-hand side is the tail of the supremum of a subgaussian process *without* penalty. However, the penalty still appears implicitly, as it determines the subset of the index set over which the supremum is taken in each term in the sum. This subset is getting smaller as k increases, which will decrease the probability; at the same time, the threshold $\alpha_k x$ also decreases, which will increase the probability. To be able to control the weighted supremum, we must therefore balance these competing forces: that is, the penalty must chosen in such a way that the size of the set $\{\rho(t,s) \le \alpha_{k-1}\}$ shrinks sufficiently rapidly as compared to the level α_k to render the probabilities summable. This basic idea is common to all applications of the slicing method: however, its successful implementation requires a bit of tuning that is specific to the setting in which it is applied. Once the idea has been understood in detail in one representative example, the application of the slicing method in other situations is largely routine; several examples will be encountered in the problems at the end of this chapter.

As a nontrivial illustration of the slicing method, we will presently develop in detail a very useful general result on weighted suprema: we will control the *modulus of continuity* of subgaussian processes. This result is of significant interest in its own right, as it sheds new light on the meaning of the entropy integral that apprears in Corollary 5.25. An increasing function ω such that $\omega(0) = 0$ is called a modulus of continuity for the random process $\{X_t\}_{t\in T}$ on the metric space (T, d) if there is a random variable C such that

$$X_t - X_s \le K\omega(d(t,s)) \quad \text{for all } t,s\in T.$$

Evidently the function ω controls the "degree of smoothness" of $t \mapsto X_t$. To show that ω is a modulus of continuity, it clearly suffices to prove that

$$K=\sup_{t,s\in T}\frac{X_t-X_s}{\omega(d(t,s))}<\infty \quad \text{a.s.}$$

To this end, we will prove the following result.

Theorem 5.32 (Modulus of continuity). *Let $\{X_t\}_{t\in T}$ be a separable subgaussian process on the metric space (T,d). Assume that $N(T,d,\epsilon) \geq (c/\varepsilon)^q$ for some constants $c,q>0$ and all $\varepsilon>0$. Then the function*

$$\omega(\delta) = \int_0^\delta \sqrt{\log N(T,d,\varepsilon)}\, d\varepsilon$$

is a modulus of continuity for $\{X_t\}_{t\in T}$. In particular, we have

$$\mathbf{E}\left[\sup_{t,s\in T} \frac{X_t - X_s}{\omega(d(t,s))}\right] < \infty.$$

Theorem 5.32 provides us with new insight on the relevance of the entropy integral in Corollary 5.25: the latter controls not only the magnitude of the supremum of the process, but in fact even its degree of smoothness!

Remark 5.33. An explicit tail bound on the quantity $\sup_{t,s}\{X_t - X_s\}/\omega(d(t,s))$ can be read off from the proof of Theorem 5.32.

Remark 5.34. The technical condition $N(T,d,\epsilon) \geq (c/\varepsilon)^q$ required by Theorem 5.32 is very mild: it states that the metric dimension of (T,d) is nonzero (cf. Remark 5.14). This is the case in almost all situations of practical interest. Nonetheless, this condition proves to be purely technical, and it can be shown that ω as defined in Theorem 5.32 is still a modulus of continuity for $\{X_t\}_{t\in T}$ even in the absence of the technical condition. The proof of this fact is in the same spirit as that of Theorem 5.32, but requires a more delicate tuning of the slicing and chaining method that does not provide much added insight. We avoid the added complications by imposing the additional technical condition in order to provide a clean illustration of the slicing method.

To control the terms that appear in the slicing method, we need a *local* version of the chaining inequality of Theorem 5.29 where the supremum is taken over $t,s\in T$ such that $\omega(d(t,s)) \leq \alpha_k$. Such a local inequality, which is very useful in its own right, can be derived rather easily from Theorem 5.29.

Proposition 5.35 (Local chaining inequality). *Let $\{X_t\}_{t\in T}$ be a separable subgaussian process on the metric space (T,d). Then for all $x,\delta\geq 0$*

$$\mathbf{P}\left[\sup_{\substack{t,s\in T\\ d(t,s)\leq\delta}} \{X_t - X_s\} \geq C\int_0^\delta \sqrt{\log N(T,d,\varepsilon)}\, d\varepsilon + x\right] \leq Ce^{-x^2/C\delta^2}.$$

Proof. Define the random process $\{\tilde X_{t,s}\}_{(t,s)\in\tilde T}$ as

$$\tilde X_{t,s} = X_t - X_s, \qquad \tilde T = \{(t,s)\in T\times T : d(t,s)\leq\delta\}.$$

Using the subgaussian property of $\{X_t\}_{t\in T}$ and Cauchy-Schwarz, we estimate

$$\begin{aligned}\mathbf{E}[e^{\lambda\{\tilde{X}_{t,s}-\tilde{X}_{u,v}\}}] &= \mathbf{E}[e^{\lambda\{X_t-X_u\}}e^{\lambda\{X_s-X_v\}}] \\ &\le \mathbf{E}[e^{2\lambda\{X_t-X_u\}}]^{1/2}\mathbf{E}[e^{2\lambda\{X_s-X_v\}}]^{1/2} \\ &\le e^{\lambda^2\{d(t,u)^2+d(s,v)^2\}},\end{aligned}$$

and by an entirely analogous argument

$$\mathbf{E}[e^{\lambda\{\tilde{X}_{t,s}-\tilde{X}_{u,v}\}}] \le \mathbf{E}[e^{2\lambda\{X_t-X_s\}}]^{1/2}\mathbf{E}[e^{2\lambda\{X_u-X_v\}}]^{1/2} \le e^{2\lambda^2\delta^2}.$$

If we define the metric $\tilde{d}$ on $\tilde{T}$ as

$$\tilde{d}((t,s),(u,v)) = 2^{1/2}\sqrt{d(t,u)^2+d(s,v)^2} \wedge 2\delta,$$

we see that $\{\tilde{X}_{t,s}\}_{(t,s)\in\tilde{T}}$ is a subgaussian process on the metric space $(\tilde{T},\tilde{d})$. As $\mathrm{diam}(\tilde{T}) \le 2\delta$ (and thus $N(\tilde{T},\tilde{d},\varepsilon) = 1$ for $\varepsilon > 2\delta$), we obtain

$$\mathbf{P}\left[\sup_{(t,s)\in\tilde{T}} \tilde{X}_{t,s} \ge C\int_0^{2\delta}\sqrt{\log N(\tilde{T},\tilde{d},\varepsilon)}\,d\varepsilon + x\right] \le Ce^{-x^2/C\delta^2}$$

by Theorem 5.29. Note that if N is an ε-net for (T,d), then $N\times N$ is a 2ε-net for $(\tilde{T},\tilde{d})$. As $|N\times N| = |N|^2$, we obtain $N(\tilde{T},\tilde{d},2\varepsilon) \le N(T,d,\varepsilon)^2$. Thus

$$\int_0^{2\delta}\sqrt{\log N(\tilde{T},\tilde{d},\varepsilon)}\,d\varepsilon \le 2\sqrt{2}\int_0^{\delta}\sqrt{\log N(T,d,\varepsilon)}\,d\varepsilon,$$

and the proof is readily completed. □

We can now complete the proof of Theorem 5.32.

Proof (Theorem 5.32). The slicing argument with $\alpha_k = \omega(\Delta 2^{-k})$ yields

$$\mathbf{P}\left[\sup_{s,t\in T}\frac{X_t-X_s}{\omega(d(t,s))} \ge x\right] \le \sum_{k=1}^{\infty}\mathbf{P}\left[\sup_{d(s,t)\le\Delta 2^{-k+1}}\{X_t-X_s\} \ge \omega(\Delta 2^{-k})x\right],$$

where we define $\Delta = \mathrm{diam}(T)$ for simplicity. We would like to apply Proposition 5.35 to each term in the sum. The problem is that here the integral $\omega(\Delta 2^{-k})$ goes only up to the scale $\Delta 2^{-k}$, while the supremum is taken up to a larger scale $\Delta 2^{-k+1}$; in Proposition 5.35, the two scales must be the same. To resolve this issue, note that as $\varepsilon \mapsto N(T,d,\varepsilon)$ is a decreasing function

$$\int_{\delta}^{2\delta}\sqrt{\log N(T,d,\varepsilon)}\,d\varepsilon \le \int_0^{\delta}\sqrt{\log N(T,d,\varepsilon)}\,d\varepsilon$$

for every $\delta > 0$, so that in particular $\omega(2\delta) \le 2\omega(\delta)$. Therefore

$$\begin{aligned}
&\mathbf{P}\left[\sup_{s,t\in T}\frac{X_t-X_s}{\omega(d(t,s))}\ge 2(C+x)\right]\\
&\le\sum_{k=1}^{\infty}\mathbf{P}\left[\sup_{d(s,t)\le\Delta 2^{-k+1}}\{X_t-X_s\}\ge(C+x)\int_0^{\Delta 2^{-k+1}}\sqrt{\log N(T,d,\varepsilon)}\,d\varepsilon\right]\\
&\le\sum_{k=1}^{\infty}Ce^{-\frac{x^2}{C}\left(\frac{1}{\Delta 2^{-k+1}}\int_0^{\Delta 2^{-k+1}}\sqrt{\log N(T,d,\varepsilon)}\,d\varepsilon\right)^2}\\
&\le\sum_{k=1}^{\infty}Ce^{-x^2\log N(T,d,\Delta 2^{-k+1})/C},
\end{aligned}$$

where we have used Proposition 5.35 and that $\varepsilon\mapsto N(T,d,\varepsilon)$ is decreasing. We now note that the technical assumption $N(T,d,\varepsilon)\ge(c/\varepsilon)^q$ implies that $\log N(T,d,\Delta 2^{-k+1})$ grows at least linearly in k. Thus the above sum is a geometric series, and we readily obtain an estimate of the form

$$\mathbf{P}\left[\sup_{s,t\in T}\frac{X_t-X_s}{\omega(d(t,s))}\ge 2C+x\right]\le Ae^{-x^2/A}\quad\text{for all } x\ge 1,$$

where C is the universal constant from Proposition 5.35 and A is a constant that depends on c,q only. Integrating the tail bound yields the conclusion. □

Remark 5.36. The proof of Theorem 5.32 highlights the competing demands on our choice of slicing sequence α_k. On the one hand, we want α_{k-1} and α_k to be sufficiently close together that the scales at which the supremum and the tail probability are evaluated are of the same order in each term in the slicing argument. This requires that the sequence α_k converges not too quickly. On the other hand, we want α_{k-1} and α_k to be sufficiently far apart that the probabilities in the slicing bound are summable. This requires that the sequence α_k converges not too slowly. In the proof of Theorem 5.32, we initially chose a geometric sequence $\alpha_k=\omega(\Delta 2^{-k})$ to ensure that $\alpha_k\le\alpha_{k-1}\le 2\alpha_k$ are not too far apart; we subsequently imposed the technical condition on the covering numbers to ensure that the probabilities are summable.

To illustrate Theorem 5.32, let us prove a classical result in stochastic analysis due to P. Lévy on the modulus of continuity of Brownian motion.

Example 5.37 (Modulus of continuity of Brownian motion). Let $\{B_t\}_{t\in[0,1]}$ be standard Brownian motion. As B_t-B_s is Gaussian, we compute exactly

$$\mathbf{E}[e^{\lambda\{B_t-B_s\}}]=e^{\lambda^2|t-s|/2},$$

Thus $\{B_t\}_{t\in[0,1]}$ is subgaussian on $([0,1],d)$ with the metric $d(t,s)=\sqrt{|t-s|}$. Moreover, by Lemma 5.13, we readily obtain the estimates

$$\frac{1}{\varepsilon^2}\le N([0,1],d,\varepsilon)=N([0,1],|\cdot|,\varepsilon^2)\le\frac{3}{\varepsilon^2}$$

for $\varepsilon \le 1$. Thus Theorem 5.32 states that

$$|B_t - B_s| \lesssim \omega(\sqrt{|t-s|}) \quad \text{for all } t,s \in [0,1] \quad \text{a.s.},$$

where

$$\omega(\delta) = \int_0^\delta \sqrt{\log \frac{3}{\varepsilon^2}}\, d\varepsilon \lesssim \delta \sqrt{\log \frac{1}{\delta}}.$$

That is, the sample paths of Brownian motion are slightly less smooth than Hölder-$\frac{1}{2}$ by a logarithmic factor. It is easy to see that this result is sharp! Indeed, note that as Brownian motion has independent increments,

$$\sup_{|t-s|\le\varepsilon} \frac{|B_t - B_s|}{\omega(\sqrt{|t-s|})} \ge \max_{n\le\varepsilon^{-1}} \frac{B_{n\varepsilon} - B_{(n-1)\varepsilon}}{\omega(\sqrt{\varepsilon})} \gtrsim \frac{\max_{n\le N} X_n}{\sqrt{\log N}},$$

where $N = \varepsilon^{-1}$ and $X_n = \varepsilon^{-1/2}\{B_{n\varepsilon} - B_{(n-1)\varepsilon}\}$ are i.i.d. $\sim N(0,1)$. Thus

$$\mathbf{E}\left[\limsup_{|t-s|\downarrow 0} \frac{|B_t - B_s|}{\omega(\sqrt{|t-s|})}\right] \gtrsim \limsup_{N\to\infty} \frac{\mathbf{E}[\max_{n\le N} X_n]}{\sqrt{\log N}} > 0$$

by Problem 5.1, so the modulus of continuity $\omega(\sqrt{|t-s|})$ is evidently sharp.

Problems

5.12 (Empirical risk minimization I: slicing). Empirical risk minimization is a simple but fundamental idea that arises throughout machine learning, statistics (where it is often called M-estimation), and stochastic programming (where it is called sample average approximation). The basic problem can be phrased as follows. Let (T,d) be a metric space, and consider a given family of functions $\{f_t : t \in T\}$ on some probability space $(\mathbb{X},\mu)$. We define the *risk* of $t \in T$ as $R(t) := \mu f_t$. Our goal is to select $t^* \in T$ that minimizes the risk:

$$t^* := \arg\min_{t\in T} R(t) := \arg\min_{t\in T} \mu f_t.$$

However, it may be impossible to do this directly: either because the measure μ is unknown (in machine learning and statistics), or because computing integrals with respect to μ is intractable (in stochastic programming). Instead, we assume that we have access to n i.i.d. samples $X_1,\ldots,X_n \sim \mu$. By the law of large numbers, the risk should be well approximated by the *empirical risk*

$$R(t) \approx \mu_n f_t := \frac{1}{n}\sum_{k=1}^n f_t(X_k)$$

when the sample size n is large. The *empirical risk minimizer*

$$\hat{t}_n := \arg\min_{t\in T} \mu_n f_t$$

should therefore be a good approximation of the optimum t^*. We would like to find out how good of an approximation this is: that is, we would like to bound the *excess risk* $R(\hat{t}_n) - R(t^*)$ of the empirical risk minimizer.

a. Argue that

$$\mathbf{P}[R(\hat{t}_n) - R(t^*) \geq \delta] \leq \mathbf{P}\left[\sup_{\substack{t \in T \\ R(t) - R(t^*) \geq \delta}} \mu_n(f_{t^*} - f_t) \geq 0\right].$$

Hint: use that $\mu_n(f_{t^*} - f_{\hat{t}_n}) \geq 0$ by construction.

b. Define the random process $X_t := \mu_n(f_{t^*} - f_t)$. Note that X_t is not centered, so that we cannot apply chaining directly. However, show that

$$Z_t := n^{1/2}\{X_t + R(t) - R(t^*)\}$$

is subgaussian on (T, d) with the metric $d(t, s) := \|f_t - f_s\|_\infty$.

c. Use the slicing argument to show that

$$\mathbf{P}[R(\hat{t}_n) - R(t^*) \geq \delta] \leq \sum_{k=1}^{\infty} \mathbf{P}\left[\sup_{R(t) - R(t^*) \leq \delta 2^k} Z_t \geq \delta 2^{k-1} n^{1/2}\right].$$

d. The bound we have obtained already suffices to obtain a crude upper bound on the magnitude of the excess risk: show that if

$$\int_0^\infty \sqrt{\log N(T, d, \varepsilon)}\, d\varepsilon < \infty,$$

then we have

$$R(\hat{t}_n) - R(t^*) = O_P(n^{-1/2}).$$

Hint: set $\delta = n^{-1/2}(K + x)$ for a sufficiently large constant K, and replace the supremum in the slicing bound by the supremum over the entire set T.

The above bound on the excess risk is exceedingly pessimistic. Indeed, if we set $\delta = Kn^{-1/2}$, then the suprema in the slicing bound are taken over the sets $T_{k,n} = \{t \in T : R(t) - R(t^*) \leq K2^k n^{-1/2}\}$ which shrink rapidly as n increases. Thus these suprema should be much smaller than is captured by our crude estimate on the excess risk, where we have entirely ignored this effect. However, we cannot obtain more precise rates unless we are able to control the sizes of the sets T_k, and this requires to impose a suitable assumption on the risk $R(t)$. To this end, it is common to assume that a *margin condition*

$$R(t) - R(t^*) \geq (d(t, t^*)/c_1)^\alpha \quad \text{for all } t \in T$$

holds for some constants $c_1 > 0$ and $\alpha > 1$.

e. Assume that the margin condition holds and that

$$\int_0^\delta \sqrt{\log N(T, d, \varepsilon)}\, d\varepsilon \leq c_2 \delta^\beta$$

for some $c_2 > 0$ and $0 < \beta < 1$. Show that

$$R(\hat{t}_n) - R(t^*) = o_P(n^{-\alpha/2(\alpha-\beta)}).$$

Hint: choose $\delta = c_3 n^{-\alpha/2(\alpha-\beta)}$ in the slicing bound for a sufficiently large constant c_3 (depending on c_1, c_2, α, β). Then we can estimate

$$C \int_0^{c_1 \delta^{1/\alpha} 2^{k/\alpha}} \sqrt{\log N(T, d, \varepsilon)}\, d\varepsilon \le \delta 2^{k-2} n^{1/2},$$

and thus it is possible to apply Proposition 5.35.

Remark 5.38. The bounds obtained in the previous problem are often unsatisfactory in practice. The reason is that we have endowed T with the uniform norm $d(t,s) := \|f_t - f_s\|_\infty$, which is too stringent in most applications: it difficult both to satisfy the margin condition and to control the covering numbers for such a strong norm. The uniform norm is the best we can hope for, however, if we use only the subgaussian property of $\{Z_t\}_{t\in T}$ (Azuma-Hoeffding). Later in this course, we will develop new tools from empirical process theory that make it possible to obtain uniform bounds on the supremum of empirical averages $\mu_n f - \mu f$ under much weaker norms. With this machinery in place, however, the slicing argument will go through precisely as we used it above.

5.13 (Empirical risk minimization II: modulus of continuity). The goal of this problem is to outline an alternative proof of the results obtained in the previous problem: rather than employing the slicing argument directly, we will deduce the bound on the excess risk from the modulus of continuity of the process $\{Z_t\}_{t\in T}$. This is not really different, of course, as one must still use slicing (in the form of Theorem 5.32) to control the modulus of continuity. The main point of the present problem, however, is to emphasize that the modulus of continuity arises naturally in the empirical risk minimization problems.

In the sequel, we work in the same setting as in the previous problem.

a. Show that

$$R(\hat{t}_n) - R(t^*) \le \mu_n(f_{t^*} - f_{\hat{t}_n}) - \mu(f_{t^*} - f_{\hat{t}_n}) = n^{-1/2} Z_{\hat{t}_n}.$$

Hint: use that $\mu_n(f_{t^*} - f_{\hat{t}_n}) \ge 0$ by construction.

b. Show directly (without slicing) that if

$$\int_0^\infty \sqrt{\log N(T, d, \varepsilon)}\, d\varepsilon < \infty,$$

then we have

$$\mathbf{E}[R(\hat{t}_n) - R(t^*)] \lesssim n^{-1/2}.$$

c. The reason that the above bound is pessimistic is that $\hat{t}_n \to t^*$, so we expect that $Z_{\hat{t}_n} - Z_{t^*} \ll \sup_{t\in T}\{Z_t - Z_{t^*}\}$. To capture this behavior, suppose that $\omega(\delta) = \delta^\beta$ is a modulus of continuity for $\{Z_t\}_{t\in T}$, so $Z_{\hat{t}_n} - Z_{t^*} \lesssim d(\hat{t}_n, t^*)^\beta$ a.s. If in addition the margin condition holds, show that this implies

$$R(\hat{t}_n) - R(t^*) \lesssim n^{-\alpha/2(\alpha-\beta)} \quad \text{a.s.}$$

d. Deduce the conclusion of the previous problem from the off-the-shelf modulus of continuity result obtained in Theorem 5.32.

5.14 (Law of iterated logarithm). A classical application of the slicing method in probability theory is the proof of the law of iterated logarithm. In this problem, we will prove the simplest form of such a result.

Let $X_1, X_2, \ldots$ be i.i.d. Gaussian random variables with zero mean and unit variance. We aim to show the law of iterated logarithm

$$\limsup_{n\to\infty} \frac{1}{\sqrt{2n\log\log n}} \sum_{k=1}^n X_k \le 1 \quad \text{a.s.}$$

(in fact, with a bit more work one can prove that equality holds a.s.)

a. Use the slicing method to show that for $\beta > 1$ and $m \in \mathbb{N}$

$$\mathbf{P}\left[\sup_{n\ge\beta^m} \frac{1}{\sqrt{2n\log\log n}} \sum_{k=1}^n X_k \ge x\right] \le \sum_{\ell=m}^\infty \mathbf{P}\left[\max_{n\le\beta^{\ell+1}} \sum_{k=1}^n X_k \ge x\sqrt{2\beta^\ell\{\log\ell + \log\log\beta\}}\right].$$

b. Prove the following maximal inequality:

$$\mathbf{P}\left[\sup_{n\le N} \sum_{k=1}^n X_k \ge x\right] \le e^{-x^2/2N}.$$

Hint: without the sup, this is the Chernoff bound for Gaussian variables. Now note that $M_n = \sum_{k=1}^n X_k$ is a martingale, so $e^{\lambda M_n}$ is a submartingale. Improve the Chernoff bound using Doob's submartingale inequality.

c. Show that whenever $x^2 > \beta$

$$\lim_{m\to\infty} \mathbf{P}\left[\sup_{n\ge\beta^m} \frac{1}{\sqrt{2n\log\log n}} \sum_{k=1}^n X_k \ge x\right] = 0,$$

and conclude the form of the law of iterated logarithm stated above.

5.15 (Maxima of independent Gaussians). Let $\{X_n\}_{n\geq 0}$ be i.i.d. $N(0,1)$ random variables. Of course, it is trivially seen that $\sup_n X_n = \infty$ a.s., so there is nothing interesting to be said about the supremum of the process $\{X_n\}_{n\geq 0}$ itself. However, even when the supremum of a process is infinite, the *penalized* supremum can still be finite if the penalty is chosen appropriately.

a. Let $a_n \uparrow \infty$. Show that $\sup_n X_n/a_n < \infty$ if and only if $a_n \gtrsim \sqrt{\log n}$.

b. Let $b_n \uparrow \infty$. Show that $\sup_n\{X_n - b_n\} < \infty$ if and only if $b_n \gtrsim \sqrt{\log n}$.

Notes

§5.1. The use of union bounds to estimate maxima of independent random variables is classical. The proof of Lemma 5.1 arises naturally from the development of maximal inequalities in terms of Orlicz norms, cf. [66]; the present formulation is taken from [13]. Orlicz norms make it possible to define bona fide Banach spaces of random variables with given tail behavior, and are therefore particularly useful in a functional-analytic setting. The Johnson-Lindenstrauss lemma (Problem 5.3) can be found, for example, in [56].

§5.2. Covering and packing numbers were first studied systematically in the beautiful paper of Kolmogorov and Tikhomirov [47], which remains surprisingly modern. The covering number estimates of finite-dimensional balls and of Lipschitz functions are already obtained there. The application of Lemma 5.7 is often referred to as "an ε-net argument"; it is the simplest and most classical method to bound the supremum of a random process. Much more on estimating the norm of a random matrix can be found in [95].

§5.3. The chaining method appears in any first course on stochastic processes in the form of the Kolmogorov continuity theorem [46, Theorem 2.2.8]. It was developed by Kolmogorov in 1934 but apparently never published by him (see [19]). The general formulation for (sub)gaussian processes in terms of covering numbers is due to Dudley [27]. A method of chaining using Orlicz norms due to Pisier [66] has become popular as it yields tail bounds without any additional effort. The tail bound of Theorem 5.29 (whose proof was inspired by [96]) is much sharper, however, and we have therefore avoided chaining with Orlicz norms. A different approach to deriving sharp chaining tail bounds can be found in [51, section 11.1]. The sharp rates of convergence for the Wasserstein LLN stated in Problem 5.11 can be found in [3] (see also [88]).

§5.4. The idea behind the slicing (also known as peeling or stratification) method already arises in the classical proof of the law of iterated logarithm (Problem 5.14) and has a long history of applications to empirical processes. Theorem 5.32 appears, without the additional technical condition, in [28]. Problems 5.12 and 5.13 only give a flavor of numerous applications of these ideas in mathematical statistics; see [38, 37] for much more on this topic.

6

Gaussian processes

In the previous chapter, we developed the chaining method to bound the suprema of subgaussian processes. This provides a powerful tool that is useful in many applications. However, at this point in the course, it is not entirely clear why this method is so effective: at first sight the method appears quite crude, being at its core little more than a conveniently organized union bound. It is therefore a remarkable fact that some form of the chaining method suffices in many situations (in some cases in a more sophisticated form than was developed in the previous chapter) to obtain sharp results.

To understand when the chaining method is sharp, we must supplement our chaining *upper* bounds in terms of corresponding *lower* bounds. It is clear that we cannot expect to obtain sharp lower bounds at the level of generality of subgaussian processes; even in the case of finite maxima, we have seen that we need the additional assumption of independence to obtain lower bounds. In the case of general suprema, a more specific structure is needed. In this chapter we will investigate the case of *Gaussian* processes, for which a very precise understanding of these questions can be obtained.

Definition 6.1 (Gaussian process). *The random process $\{X_t\}_{t\in T}$ is called a (centered)* Gaussian process *if the random variables $\{X_{t_1}, \cdots, X_{t_n}\}$ are centered and jointly Gaussian for all $n \geq 1$, $t_1, \cdots, t_n \in T$.*

There are several reasons to concentrate on Gaussian processes:

1. Gaussian processes arise naturally in many important applications, both explicitly and implicitly as a mathematical tool in proofs.
2. Gaussian processes provide us with the simplest prototypical setting in which to investigate and understand chaining lower bounds.
3. Our investigation of Gaussian processes will give rise to new ideas and methods that are applicable far beyond the Gaussian setting.

Remark 6.2. In the sequel, all Gaussian processes will be assumed to be centered (that is, $\mathbf{E}[X_t] = 0$) unless stated otherwise. Some methods to deal with non-centered processes were discussed in section 5.4.

Let us remark at the outset that for a Gaussian process $\{X_t\}_{t\in T}$, we have

$$\mathbf{E}[e^{\lambda\{X_t-X_s\}}] = e^{\lambda^2\mathbf{E}[|X_t-X_s|^2]/2}.$$

Thus a Gaussian process determines a canonical metric on the index set T.

Definition 6.3 (Natural distance). *A Gaussian process $\{X_t\}_{t\in T}$ is subgaussian on (T,d) for the* natural distance $d(t,s) := \mathbf{E}[|X_t - X_s|^2]^{1/2}$.

Gaussian processes $\{X_t\}_{t\in T}$ will always be considered as being defined on (T,d) endowed with the natural distance d. As we will see in the sequel, the magnitude of the suprema of Gaussian processes can be understood completely (up to universal constants) in terms of chaining under the natural distance. Once this has been understood, we can truly conclude that chaining is the "right" way to think about the suprema of random processes.

6.1 Comparison inequalities

How can we obtain a lower bound on the expected supremum of a Gaussian processes? The simplest possible situation is one that was already developed in Problem 5.1: if $X_1,\ldots,X_n$ are i.i.d. Gaussians, the maximal inequalities of section 5.1 are sharp. As this elementary fact will form the basis for all further developments, let us begin by giving a complete proof.

Lemma 6.4. *If $X_1,\ldots,X_n$ are i.i.d. $\mathcal{N}(0,\sigma^2)$ random variables, then*

$$c\sigma\sqrt{\log n} \le \mathbf{E}\left[\max_{i\le n} X_i\right] \le \sigma\sqrt{2\log n}$$

for a universal constant c.

Proof. The upper bound follows immediately from Lemma 5.1 (and does not require independence). To prove the lower bound, note that for any $\delta > 0$

$$\begin{aligned}\mathbf{E}\left[\max_{i\le n} X_i\right] &= \int_0^\infty \mathbf{P}\left[\max_{i\le n} X_i \ge t\right] dt + \mathbf{E}\left[\max_{i\le n} X_i \wedge 0\right] \\ &\ge \delta\,\mathbf{P}\left[\max_{i\le n} X_i \ge \delta\right] + \mathbf{E}[X_1\wedge 0] \\ &= \delta\{1-(1-\mathbf{P}[X_1\ge\delta])^n\} + \mathbf{E}[X_1\wedge 0],\end{aligned}$$

as $\mathbf{P}[\max_{i\le n} X_i \ge t]$ is decreasing in t and as $\{X_i\}$ are i.i.d. Now note that

$$\mathbf{P}[X_1\ge\delta] = \int_\delta^\infty \frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}\,dx \ge \frac{e^{-\delta^2/\sigma^2}}{c_1}$$

for a universal constant c_1, where we used $x^2 = (x-\delta+\delta)^2 \le 2(x-\delta)^2 + 2\delta^2$. Thus if we choose the parameter δ as

$$\delta = \sigma\sqrt{\log n} - \sigma\sqrt{\log c_1},$$

we have $\mathbf{P}[X_1 \ge \delta] \ge 1/n$. This implies

$$\mathbf{E}\left[\max_{i\le n} X_i\right] \ge (1-e^{-1})\sigma\sqrt{\log n} - c_2\sigma$$

for a universal constant c_2. Thus the result follows when $n \ge e^{4c_2^2/(1-e^{-1})^2}$. On the other hand, as there are only a finite number of values $n < e^{4c_2^2/(1-e^{-1})^2}$, the lower bound trivially holds with some universal constant in this case. □

Let $\{X_t\}_{t\in T}$ be a random process on a general index set T. The intuition behind the upper bounds developed in the previous chapter was that while X_t and X_s will be strongly dependent when t and s are close together, X_t and X_s can be nearly independent when t and s are far apart. This motivated the approximation of the supremum by finite maxima over well separated points, for which the result of Lemma 5.1 might reasonably be expected to be sharp. However, we never actually used any form of independence in the proofs: our upper bounds still work even if the intuition fails. On the other hand, we can only expect these bounds to be sharp if the intuition does in fact hold. The first challenge that we face in proving lower bounds is therefore to make mathematical sense of the above intuition that was only used as a guiding heuristic for obtaining upper bounds in the previous chapter. This is precisely what will be done in this section in the setting of Gaussian processes.

What should such a result look like? Let N be a maximal ε-packing of T. If $\{X_t : t \in N\}$ behave in some sense like independent Gaussians, then we would expect by Lemma 6.4 that $\mathbf{E}[\sup_{t\in T} X_t] \ge \mathbf{E}[\max_{t\in N} X_t] \gtrsim \sqrt{\log|N|}$. In view of the duality between packing and covering numbers (Lemma 5.12), this is precisely the content of the following result.

Theorem 6.5 (Sudakov). *For a Gaussian process $\{X_t\}_{t\in T}$, we have*

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \ge c\sup_{\varepsilon>0}\varepsilon\sqrt{\log N(T,d,\varepsilon)}$$

for a universal constant c.

Remark 6.6. Combining Sudakov's lower bound with the upper bound obtained in the previous chapter by chaining, we have evidently shown that

$$\sup_{\varepsilon>0}\varepsilon\sqrt{\log N(T,d,\varepsilon)} \lesssim \mathbf{E}\left[\sup_{t\in T} X_t\right] \lesssim \int_0^\infty \sqrt{\log N(T,d,\varepsilon)}\,d\varepsilon,$$

or, equivalently up to universal constants,

$$\sup_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log N(T,d,2^{-k})} \lesssim \mathbf{E}\left[\sup_{t\in T} X_t\right] \lesssim \sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log N(T,d,2^{-k})}.$$

Thus the upper bound and the lower bound we have obtained contain precisely the same terms at every scale; however, the upper bound is a multiscale bound (a sum over all scales), while the lower bound is a single scale bound (a maximum over all scales). These two bounds are not as far apart as may appear at first sight: in many situations the terms $2^{-k}\sqrt{\log N(T,d,2^{-k})}$ behave like a geometric series, so that their sum is of the same order as the largest term. There are also many cases, however, where there is indeed a gap between these two bounds. The main objective in the remainder of this chapter will be to close the gap between these upper and lower bounds.

Remark 6.7. We have phrased Theorem 6.5 in terms of the covering numbers $N(T,d,\varepsilon)$ to bring out the similarity between the upper and lower bounds. It should be emphasized, however, that upper and lower bounds require in principle fundamentally different ingredients. Upper bounds, which require approximation of every point in the index set T, are naturally obtained in terms of a covering of T. On the other hand, lower bounds, which require a subset of T that is well separated, are naturally obtained in terms of a packing of T (indeed, it is in fact the packing number $D(T,d,\varepsilon)$ and not the covering number that arises in the proof of Theorem 6.5). The duality of packing and covering, while somewhat hidden in the statement of our results, therefore lies at the heart of the development of matching upper and lower bounds. While the duality between packing an covering numbers (Lemma 5.12) is elementary, the development of a more sophisticated form of this duality will prove to be one of the challenges that we must surmount in our quest to develop matching chaining upper and lower bounds for Gaussian processes.

We now turn to the proof of Theorem 6.5. The key idea that we aim to make precise is that if N is an ε-packing, then the Gaussian vector $\{X_t\}_{t\in N}$ behaves in some sense like a collection $\{Y_t\}_{t\in N}$ of i.i.d. Gaussians, so that we can apply Lemma 6.4. We therefore need a tool that allows us to compare the maxima of two different Gaussian vectors. To this end, we will use the following classical *comparison inequality* for Gaussian vectors.

Theorem 6.8 (Slepian-Fernique). *Let $X \sim N(0,\Sigma^X)$ and $Y \sim N(0,\Sigma^Y)$ be n-dimensional Gaussian vectors. Suppose that we have*

$$\mathbf{E}|X_i - X_j|^2 \ge \mathbf{E}|Y_i - Y_j|^2 \quad \textit{for all } i,j = 1,\ldots,n.$$

Then

$$\mathbf{E}\left[\max_{i\le n} X_i\right] \ge \mathbf{E}\left[\max_{i\le n} Y_i\right].$$

Using this comparison inequality, we can now easily complete the proof of Sudakov's inequality by comparing with the independent case.

Proof (Theorem 6.5). Fix $\varepsilon > 0$ and an ε-packing N of T for the time being. Define $X = \{X_t\}_{t\in N}$, and let $Y = \{Y_t\}_{t\in N}$ be i.i.d. $N(0, \varepsilon^2/2)$ variables. Then

$$\mathbf{E}|X_t - X_s|^2 = d(t,s)^2 \geq \varepsilon^2 = \mathbf{E}|Y_t - Y_s|^2 \quad \text{for all } t, s \in N,\ t \neq s.$$

Therefore, we obtain using Theorem 6.8 and Lemma 6.4

$$\mathbf{E}\left[\max_{t\in T} X_t\right] \geq \mathbf{E}\left[\max_{t\in N} X_t\right] \geq \mathbf{E}\left[\max_{t\in N} Y_t\right] \geq c\varepsilon\sqrt{\log|N|}.$$

We now optimize over $\varepsilon > 0$ and ε-packings N to obtain

$$\mathbf{E}\left[\max_{t\in T} X_t\right] \geq c\sup_{\varepsilon>0}\varepsilon\sqrt{\log D(T,d,\varepsilon)} \geq c\sup_{\varepsilon>0}\varepsilon\sqrt{\log N(T,d,\varepsilon)},$$

where we have used Lemma 5.12 in the last inequality. □

We now turn to the proof of Theorem 6.8. Let us note that up to this point, we have not used any properties that are particularly specific to Gaussian processes. Indeed, in Lemma 6.4 we used only a subgaussian-type lower bound on the tail probabilities, and the conclusions of Theorems 6.5 and 6.8 can certainly hold also for other types of processes. In the proof of Theorem 6.8, however, we will perform computations that exploit the specific form of the Gaussian distribution. This is the only point in this chapter we will use the full strength of the Gaussian assumption. The Gaussian interpolation technique that will be used in the proof is of interest in its own right, and proves to be useful in many other interesting problems involving Gaussian variables.

The idea behind the proof of Theorem 6.8 is as follows. We would like to prove that the expected maximum of the vector Y is smaller than that of the vector X. Rather than proving this directly, we will define a family of Gaussian vectors $\{Z(t)\}_{t\in[0,1]}$ that *interpolate* between $Z(0) = Y$ and $Z(1) = X$. To establish Theorem 6.8, it then suffices to show that the expected maximum of $Z(t)$ is increasing in t. The beauty of this approach is that the latter problem can be investigated "locally" by considering the derivative with respect to t.

Lemma 6.9 (Interpolation). *Let $X \sim N(0, \Sigma^X)$ and $Y \sim N(0, \Sigma^Y)$ be independent n-dimensional Gaussian vectors, and define*

$$Z(t) = \sqrt{t}\,X + \sqrt{1-t}\,Y, \qquad t \in [0,1].$$

Then we have for every smooth function f

$$\frac{d}{dt}\mathbf{E}[f(Z(t))] = \frac{1}{2}\sum_{i,j=1}^{n}(\Sigma^X_{ij} - \Sigma^Y_{ij})\,\mathbf{E}\left[\frac{\partial^2 f}{\partial x_i \partial x_j}(Z(t))\right].$$

The result of Lemma 6.9 is very closely related to the computations that we performed to prove the Gaussian Poincaré inequality in Example 2.22: the second derivative appears here for precisely the same reason as it does in the generator of the Ornstein-Uhlenbeck process. To prove Lemma 6.9, we require a multidimensional version of the Gaussian integration by parts Lemma 2.24.

Lemma 6.10 (Gaussian integration by parts). *Let $X \sim N(0, \Sigma)$. Then*

$$\mathbf{E}[X_i f(X)] = \sum_{j=1}^{n} \Sigma_{ij}\, \mathbf{E}\left[\frac{\partial f}{\partial x_j}(X)\right].$$

Proof. Let $Z \sim N(0, I)$. Then X has the same distribution as $\Sigma^{1/2} Z$ Thus

$$\mathbf{E}[X_i f(X)] = \sum_{k=1}^{n} \Sigma_{ik}^{1/2}\, \mathbf{E}[Z_k f(\Sigma^{1/2} Z)] = \sum_{k=1}^{n} \Sigma_{ik}^{1/2}\, \mathbf{E}[Z_k g(Z)],$$

where $g(z) = f(\Sigma^{1/2} z)$. As $\{Z_k\}$ are independent, we can apply the integration by parts Lemma 2.24 conditionally on $\{Z_j\}_{j \neq k}$ to obtain

$$\mathbf{E}[Z_k g(Z)] = \mathbf{E}\left[\frac{\partial g}{\partial z_k}(Z)\right] = \sum_{j=1}^{n} \Sigma_{jk}^{1/2}\, \mathbf{E}\left[\frac{\partial f}{\partial x_j}(\Sigma^{1/2} Z)\right].$$

The proof is easily completed as $\sum_k \Sigma_{ik}^{1/2} \Sigma_{jk}^{1/2} = \Sigma_{ij}$. □

Using the Gaussian integration by parts property, the proof of the interpolation Lemma 6.9 is now a matter of straightforward computation.

Proof (Lemma 6.9). We readily compute

$$\begin{aligned}\frac{d}{dt}\mathbf{E}[f(Z(t))] &= \sum_{i=1}^{n} \mathbf{E}\left[\frac{\partial f}{\partial x_i}(Z(t))\, \frac{dZ_i(t)}{dt}\right] \\ &= \frac{1}{2} \sum_{i=1}^{n} \mathbf{E}\left[\frac{\partial f}{\partial x_i}(Z(t)) \left\{\frac{X_i}{\sqrt{t}} - \frac{Y_i}{\sqrt{1-t}}\right\}\right].\end{aligned}$$

As X and Y are independent, we can apply Lemma 6.10 to the $2n$-dimensional Gaussian vector (X, Y) to compute the first term on the right as

$$\mathbf{E}\left[\frac{\partial f}{\partial x_i}(Z(t))\, \frac{X_i}{\sqrt{t}}\right] = \sum_{j=1}^{n} \Sigma_{ij}^{X}\, \mathbf{E}\left[\frac{\partial^2 f}{\partial x_i \partial x_j}(Z(t))\right].$$

An identical computation for the second term completes the proof. □

We are now ready to complete the proof of Theorem 6.8. Ideally, we would like the proof to work as follows. First, we define $f(x) = \max_{i \leq n} x_i$. We then use Lemma 6.9 to establish that under the assumptions of Theorem 6.8

$$\frac{d}{dt}\mathbf{E}[f(Z(t))] \geq 0.$$

Then the proof is complete, as this evidently implies

$$\mathbf{E}\left[\max_{i\le n} X_i\right] = \mathbf{E}[f(Z(1))] \ge \mathbf{E}[f(Z(0))] = \mathbf{E}\left[\max_{i\le n} Y_i\right].$$

The problem with this idea is that the function f is not twice differentiable, so that we cannot apply Lemma 6.9 directly. We can nonetheless make the proof work by working with a convenient smooth approximation of the function f.

Proof (Theorem 6.8). Define for $\beta > 0$ the function

$$f_\beta(x) = \frac{1}{\beta}\log\sum_{i=1}^n e^{\beta x_i}.$$

Then evidently (cf. Problem 5.2)

$$\max_{i\le n} x_i = \frac{1}{\beta}\log\left(\max_{i\le n} e^{\beta x_i}\right) \le f_\beta(x) \le \frac{1}{\beta}\log\left(n\max_{i\le n} e^{\beta x_i}\right) = \max_{i\le n} x_i + \frac{\log n}{\beta}.$$

Thus $f_\beta(x) \to \max_{i\le n} x_i$ as $\beta\to\infty$. Moreover,

$$\frac{\partial f_\beta(x)}{\partial x_i} = \frac{e^{\beta x_i}}{\sum_{j=1}^n e^{\beta x_j}} =: p_i(x), \qquad \frac{\partial^2 f_\beta(x)}{\partial x_i\partial x_j} = \beta\{\delta_{ij}p_i(x) - p_i(x)p_j(x)\}.$$

Lemma 6.9 therefore yields

$$\begin{aligned}\frac{d}{dt}\mathbf{E}[f_\beta(Z(t))] = {}& \frac{\beta}{2}\sum_{i=1}^n(\Sigma_{ii}^X - \Sigma_{ii}^Y)\,\mathbf{E}[p_i(Z(t))\{1-p_i(Z(t))\}] \\ & - \frac{\beta}{2}\sum_{i\ne j}(\Sigma_{ij}^X - \Sigma_{ij}^Y)\,\mathbf{E}[p_i(Z(t))p_j(Z(t))].\end{aligned}$$

But noting that $1 - p_i(x) = \sum_{j\ne i} p_j(x)$, we can write

$$\sum_{i=1}^n a_i\,p_i(x)\{1-p_i(x)\} = \sum_{i\ne j} a_i\,p_i(x)p_j(x) = \sum_{i\ne j} a_j\,p_i(x)p_j(x),$$

where we exchanged the roles of the variables i and j. Averaging the two expressions on the right hand side and plugging into the above identity yields

$$\frac{d}{dt}\mathbf{E}[f_\beta(Z(t))] = \frac{\beta}{4}\sum_{i\ne j}\{\mathbf{E}|X_i - X_j|^2 - \mathbf{E}|Y_i - Y_j|^2\}\,\mathbf{E}[p_i(Z(t))p_j(Z(t))]$$

using $\mathbf{E}|X_i - X_j|^2 = \Sigma_{ii}^X - 2\Sigma_{ij}^X + \Sigma_{jj}^X$ and $\mathbf{E}|Y_i - Y_j|^2 = \Sigma_{ii}^Y - 2\Sigma_{ij}^Y + \Sigma_{jj}^Y$. It follows immediately from our assumptions that the right hand side of this expression is nonnegative, so that $\mathbf{E}[f_\beta(Z(t))]$ is increasing in t. Thus

$$\mathbf{E}[f_\beta(X)] = \mathbf{E}[f_\beta(Z(1))] \ge \mathbf{E}[f_\beta(Z(0))] = \mathbf{E}[f_\beta(Y)].$$

Letting $\beta\to\infty$ in this expression completes the proof. □

The conclusion of the proof of Theorem 6.8 marks the last time in this chapter that we will make explicit use of the Gaussian property of the underlying process. In the rest of this chapter, we will only make use of two facts about Gaussian processes: the validity of Sudakov's inequality (Theorem 6.5), and Gaussian concentration (Theorem 3.25). While both these properties are stronger than the subgaussian property used in the previous chapter, such properties or their variants do continue to hold in many situations where the underlying process is not actually Gaussian. For this reason, while we will concentrate our attention here on the classical setting of Gaussian processes for concreteness, the methods that we are about to develop prove to be very useful in a variety of problems that go far beyond the Gaussian setting.

Problems

Problem 6.11 (Norm of a random matrix). Let M be an $n \times m$ random matrix such that M_{ij} are independent $N(0,1)$ random variables. In Example 5.10, we used an ε-net argument to show that $\mathbf{E}\|M\| \le C\sqrt{n+m}$ for some universal constant C (this conclusion holds even in the case where the entries M_{ij} are only subgaussian). The goal of this problem is to obtain some further insight on the norm of a random matrix in the Gaussian case.

a. The ε-net argument only yields an upper bound $\mathbf{E}\|M\| \le C\sqrt{n+m}$. It is far from clear, *a priori*, whether this bound is sharp. Use Sudakov's inequality to show that in the Gaussian case, we have in fact a matching lower bound $\mathbf{E}\|M\| \ge C'\sqrt{n+m}$ for some universal constant C'.

   Hint: consider the Gaussian process $X_{v,w} = \langle v, Mw\rangle$ on $S^{n-1} \times S^{m-1}$ (where S^{n-1} is the unit sphere in $\mathbb{R}^n$), and show that the corresponding natural distance satisfies $d((v,w),(v',w')) \ge \|v-v'\| \vee \|w-w'\|$.

While upper bounds using ε-net arguments or chaining often give sharp results up to universal constants, there is little hope to obtain realistic values of the constants in this manner. If one cares about the best values of the constants, one must typically resort to other techniqes. In the Gaussian setting of this problem, we can use the Slepian-Fernique inequality as a replacement for the ε-net argument to prove the much sharper inequality $\mathbf{E}\|M\| \le \sqrt{n}+\sqrt{m}$. In fact, it is known from random matrix theory that this result is sharp asymptotically as $n \to \infty$ with $m \propto n$ (note that this improved estimate does not contradict our earlier bounds as $2^{-1/2}\{\sqrt{n}+\sqrt{m}\} \le \sqrt{n+m} \le \sqrt{n}+\sqrt{m}$).

b. Let $Z \sim N(0, I_n)$ and $Z' \sim N(0, I_m)$ be independent standard Gaussian vectors of dimensions n and m, and define for $(v,w) \in S^{n-1} \times S^{m-1}$

$$X_{v,w} = \langle v, Mw\rangle, \qquad Y_{v,w} = \langle v, Z\rangle + \langle w, Z'\rangle.$$

   Show that $\mathbf{E}|Y_{v,w} - Y_{v',w'}|^2 \ge \mathbf{E}|X_{v,w} - X_{v',w'}|^2$ for all v, v', w, w'.

c. Conclude by the Slepian-Fernique inequality that $\mathbf{E}\|M\| \le \sqrt{n}+\sqrt{m}$.

Problem 6.12 (Gordon's inequality and the smallest singular value). The Slepian-Fernique inequality is only one of a family of Gaussian comparison inequalities. There is nothing terribly special about the maximum function—the only important property needed to apply the interpolation Lemma 6.9 is that the second derivatives of the function have the appropriate sign.

In this problem, we will develop another Gaussian comparison inequality due to Gordon. To this end, let X and Y be $n \times m$ matrices with centered and jointly Gaussian (but not necessarily independent) entries. To obtain a comparison, we will assume the following inequalities between the covariances:

$$\begin{aligned} \mathbf{E}[X_{ij}X_{il}] &\le \mathbf{E}[Y_{ij}Y_{il}] && \text{for all } i,j,l, \\ \mathbf{E}[X_{ij}X_{kl}] &\ge \mathbf{E}[Y_{ij}Y_{kl}] && \text{for all } i \ne k \text{ and } j,l, \\ \mathbf{E}[X_{ij}^2] &= \mathbf{E}[Y_{ij}^2] && \text{for all } i,j. \end{aligned}$$

a. Show that for all $x \in \mathbb{R}$

$$\mathbf{P}\left[\min_{i\le n}\max_{j\le m} X_{ij} \ge x\right] \ge \mathbf{P}\left[\min_{i\le n}\max_{j\le m} Y_{ij} \ge x\right].$$

Hint: let $\alpha_k : \mathbb{R} \to [0,1]$ be smooth and decreasing in x such that $\alpha_k(x) \to \mathbf{1}_{x<0}$ as $k \to \infty$. Apply Lemma 6.9 to $f_k(x) = \prod_{i=1}^n \{1 - \prod_{j=1}^m \alpha_k(x_{ij} - x)\}$.

b. Conclude that

$$\mathbf{E}\left[\min_{i\le n}\max_{j\le m} X_{ij}\right] \ge \mathbf{E}\left[\min_{i\le n}\max_{j\le m} Y_{ij}\right].$$

Let M be an $n \times m$ random matrix with $n > m$, such that M_{ij} are independent $N(0,1)$ random variables. The minimal and maximal singular values of M are defined as the optimal constants $s_{\min}(M)$, $s_{\max}(M)$ in the inequality

$$s_{\min}(M)\|x\| \le \|Mx\| \le s_{\max}(M)\|x\| \quad \text{for all } x \in \mathbb{R}^m.$$

Evidently $s_{\max}(M) = \|M\|$, and thus we obtained a sharp upper bound for $s_{\max}(M)$ using Slepian's inequality in the previous problem. Using Gordon's inequality, we can obtain a sharp lower bound for $s_{\min}(M)$.

c. Use Gordon's inequality to show that $\mathbf{E}[s_{\min}(M)] \ge \sqrt{n} - \sqrt{m}$.

Hint: If $Z_n \sim N(0, I_n)$ is n-dimensional standard normal, it can be verified by tedious explicit computation that $\mathbf{E}\|Z_n\| - \sqrt{n}$ is increasing in n.

Problem 6.13 (Sudakov's inequality and convex geometry). The proof of Sudakov's inequality that we have given is certainly the most intuitive. However, it relies on the Slepian-Fernique inequality, whose proof is based on explicit Gaussian computations. The goal of this problem is to give a completely different proof of Sudakov's inequality using ideas from convex geometry. The fact that Sudakov's inequality can be proved by such drastically

different means suggests that this result is more robust and less closely tied to the precise form of the Gaussian distribution than might appear from the proof using Slepian-Fernique. In any case, the connection between Sudakov's inequality and convex geometry is of significant interest in its own right.

We begin by reducing the problem to a convenient special case. Let $G = \{g_1, \ldots, g_n\}$ be independent $N(0,1)$ variables, and define

$$X_t = \sum_{k=1}^{n} g_k t_k, \qquad t \in \mathbb{R}^n.$$

Let $T \subseteq \mathbb{R}^n$, and consider the Gaussian process $\{X_t\}_{t\in T}$. The natural distance for this process is simply the Euclidean distance $d(x,y) = \|x - y\|$.

a. Argue that to prove Theorem 6.5 in full generality, it suffices to consider the special Gaussian processes $\{X_t\}_{t\in T}$ as defined above.

Hint: for any Gaussian process $\{Z_u\}_{u\in U}$ and points $u_1, \ldots, u_n \in U$, find points $t_1, \ldots, t_n \in \mathbb{R}^n$ such that $\{Z_{u_i}\}_{i\le n}$ has the same law as $\{X_{t_i}\}_{i\le n}$.

b. Argue further that it suffices to consider only *convex* sets $T \subseteq \mathbb{R}^n$.

c. Show that for any $t_0 \in T$

$$\mathbf{E}\left[\sup_{t\in T} |X_t - X_{t_0}|\right] \le 2\,\mathbf{E}\left[\sup_{t\in T} X_t\right].$$

Conclude that it suffices to consider only *symmetric* convex sets $T \subseteq \mathbb{R}^n$.

We now take a rather surprising detour by proving an apparently quite different result. Given two convex sets A and B in $\mathbb{R}^n$, let $N(B,A)$ be the smallest number of translates of A needed to cover B: that is,

$$N(B,A) := \min\left\{k : \exists\, x_1, \ldots, x_k \in \mathbb{R}^n \text{ such that } B \subseteq \bigcup_{l=1}^{k} \{x_l + A\}\right\}.$$

We are going to prove the following inequality:

$$\mathbf{P}[G \in A] \ge \frac{2}{3} \quad \text{implies} \quad \sup_{\varepsilon>0} \varepsilon\sqrt{\log N(B_2, \varepsilon A)} \le c$$

for some universal constant c, where $B_2 = \{x \in \mathbb{R}^n : \|x\| \le 1\}$ is the Euclidean unit ball and A is any symmetric convex set. The proof of this result is one that we are quite familiar with: we will essentially use the same *volume argument* as was used in the proof of Lemma 5.13, but we will use the Gaussian measure $\mathbf{P}[G \in A]$ to measure the "volume" of the set A instead of the Lebesgue measure. The main difficulty is that the Gaussian measure, unlike the Lebesgue measure, is not translation-invariant, so we must first understand how to estimate the Gaussian measure of a translate of a set.

d. Let A be a symmetric set. Show that

$$\mathbf{P}[G \in x + A] \geq e^{-\|x\|^2/2}\, \mathbf{P}[G \in A] \quad \text{for all } x \in \mathbb{R}^n.$$

Hint: write out the probability as a Gaussian integral and use Jensen.

e. Let A be a symmetric set. Let $x_1, \ldots, x_k \in B_2$ be such that the translates $\{x_i + \varepsilon A\}$ are disjoint. Show that we can estimate

$$k\, e^{-1/2\varepsilon^2}\, \mathbf{P}[G \in A] \leq \sum_{i=1}^{k} \mathbf{P}[G \in \tfrac{x_i}{\varepsilon} + A] \leq 1.$$

f. Let A be a symmetric convex set. Show that

$$N(B, 2A) \leq \max\{k : \exists\, x_1, \ldots, x_k \in B \text{ s.t. } \{x_i + A\}_{i=1,\ldots,k} \text{ are disjoint}\}.$$

Hint: if $\{x + A\} \cap \{z + A\} \neq \varnothing$, then $z \in x + A - A$, and thus $z \in x + 2A$ as A is symmetric and convex (note that $A + A \neq 2A$ without convexity!)

g. Conclude that if A is a symmetric convex set and $\mathbf{P}[G \in A] \geq 2/3$, then

$$\sup_{\varepsilon > 0} \varepsilon \sqrt{\log N(B_2, \varepsilon A)} \leq c$$

for a universal constant c.

So far, the supremum of the Gaussian process does not appear. Let us correct this. Let T be a symmetric convex set, and define its *polar*

$$T^\circ := \{x \in \mathbb{R}^n : \langle t, x \rangle \leq 1 \text{ for all } t \in T\}.$$

Then evidently

$$\mathbf{P}[G \in aT^\circ] = \mathbf{P}\left[\sup_{t \in T} X_t \leq a\right] \geq 1 - \frac{1}{a}\, \mathbf{E}\left[\sup_{t \in T} X_t\right]$$

by Markov's inequality. So if we choose $A = 3\, \mathbf{E}[\sup_{t \in T} X_t]\, T^\circ$, we obtain

$$\sup_{\varepsilon > 0} \varepsilon \sqrt{\log N(B_2, \varepsilon T^\circ)} \leq 3c\, \mathbf{E}\left[\sup_{t \in T} X_t\right].$$

This result is known as the *dual Sudakov inequality*. The covering number on the right-hand side is not the same one that shows up in the Sudakov inequality: in Theorem 6.5, $N(B_2, \varepsilon T^\circ)$ is replaced by $N(T, d, \varepsilon) = N(T, \varepsilon B_2)$. To deduce the Sudakov inequality from the dual Sudakov inequality, we will use a convex duality argument to relate these two covering numbers.

h. Show that for every $x \in \mathbb{R}^n$

$$\|x\|^2 = \langle x, x \rangle \leq \sup_{t \in T} \langle t, x \rangle \sup_{t \in T^\circ} \langle t, x \rangle.$$

Hint: note that $x / \sup_{t \in T} \langle t, x \rangle \in T^\circ$.

i. Conclude from the previous part that $2T \cap \frac{\varepsilon^2}{2} T^\circ \subseteq \varepsilon B_2$, and therefore

$$N(T, \varepsilon B_2) \le N(T, 2T \cap \tfrac{\varepsilon^2}{2} T^\circ) = N(T, \tfrac{\varepsilon^2}{2} T^\circ).$$

j. Show that

$$N(T, \varepsilon B_2) \le N(T, 2\varepsilon B_2)\, N(2\varepsilon B_2, \tfrac{\varepsilon^2}{2} T^\circ).$$

Hint: construct a cover of T by translates of $\frac{\varepsilon^2}{2} T^\circ$ by first covering T by translates of $2\varepsilon B_2$, then covering each of the latter by translates of $\frac{\varepsilon^2}{2} T^\circ$.

k. Conclude that

$$\sup_{\varepsilon > 0} \varepsilon \sqrt{\log N(T, \varepsilon B_2)} \le 8 \sup_{\varepsilon > 0} \varepsilon \sqrt{\log N(B_2, \varepsilon T^\circ)},$$

so that the Theorem 6.5 follows from the dual Sudakov inequality.

6.2 Chaining in reverse and stationary processes

In the previous section we made a first step towards proving lower bounds for the suprema of Gaussian processes: we showed how one can make precise the intuition that well-separated points behave like independent variables. This allows us to obtain a lower bound in terms of the covering number at a single scale. However, in the upper bound we obtained by chaining, we necessarily must deal with infinitely many scales in order to eliminate the remainder term in the chaining method. In order to close the gap between our upper and lower bounds, our second challenge is therefore show how to obtain a *multiscale* lower bound. We will presently show how this can be done.

Let us recall the basic step in the chaining method: if $\mathrm{diam}(T) \le \varepsilon$ and if $N \subseteq T$ is an $\varepsilon/2$-net, then we have for some universal constant c_1

$$\mathbf{E}\left[\sup_{t \in T} X_t\right] \le c_1 \varepsilon \sqrt{\log |N|} + \mathbf{E}\left[\sup_{t \in T} \{X_t - X_{\pi(t)}\}\right].$$

This yields the contribution at a single scale ε, plus a remainder term. By iterating this bound, we can eliminate the remainder term and obtain a sum at infinitely many scales. To obtain a matching lower bound, we would like to mimick this procedure in the reverse direction. In order to do this, we would like to have an inequality of the following form: if $N \subseteq T$ is an ε-packing, then

$$\mathbf{E}\left[\sup_{t \in T} X_t\right] \ge c_2 \varepsilon \sqrt{\log |N|} \;+\; \text{a remainder term}$$

for some universal constant c_2. In the absence of the remainder term, this is precisely Sudakov's inequality proved in the previous section. However, without the remainder term, our lower bound necessarily terminates at a

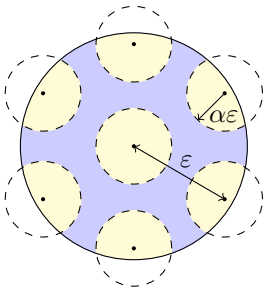
single scale. On the other hand, if we could prove an improvement of Sudakov's inequality that includes a remainder term (hopefully of a similar form to the one that appears in the chaining upper bound), then it becomes possible to iterate this inequality to obtain a multiscale lower bound. In essence, our aim is to develop an improved version of Sudakov's inequality that will allow us to run the chaining argument in reverse! This is the idea of the following result.

Theorem 6.14 (Super-Sudakov). *Let $\{X_t\}_{t\in T}$ be a separable Gaussian process and let N be an ε-packing of (T,d). Then we can estimate*

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \ge c\varepsilon\sqrt{\log|N|} + \min_{s\in N}\mathbf{E}\left[\sup_{t\in B(s,\alpha\varepsilon)} X_t\right],$$

where c and $\alpha < \frac{1}{2}$ are universal constants and $B(s,\varepsilon) := \{t \in T : d(t,s) \le \varepsilon\}$.

The geometry of Theorem 6.14 is illustrated in the following figure:



The set T (large circle) is packed with points at distance ε; around each point in the packing, we consider the set of parameters in a ball with radius $\alpha\varepsilon$ (small circles). The supremum of the process over the entire set is estimated from below by the lower bound obtained by applying Sudakov's inequality to the ε-packing, plus a remainder term which corresponds to the smallest expected supremum of the process over one of the disjoint balls.

The proof of Theorem 6.14 is not difficult. It will be deduced directly from Sudakov's inequality, together with the following basic consequence of the Gaussian concentration principle (Theorem 3.25).

Lemma 6.15 (Concentration of suprema). *Let $\{X_t\}_{t\in T}$ be a separable Gaussian process. Then $\sup_{t\in T} X_t$ is $\sup_{t\in T}\mathrm{Var}[X_t]$-subgaussian.*

Proof. By separability, we can approximate the supremum over T by the supremum over a finite set (cf. the proof of Theorem 5.24). It therefore suffices to prove the result for the maximum $\max_{i\le n} X_i$ of an n-dimensional Gaussian vector $X \sim N(0,\Sigma)$. It is convenient to write $X = \Sigma^{1/2}Z$ for $Z \in N(0,I)$. It then follows from Theorem 3.25 that $\max_{i\le n} X_i$ is $\|\nabla f\|_\infty^2$-subgaussian, where we have defined the function $f(z) := \max_{i\le n}(\Sigma^{1/2}z)_i$. Note that

$$\frac{\partial f}{\partial z_i}(z) = \sum_{j=1}^{n} \mathbf{1}_{j=i^*(z)} \Sigma_{ji}^{1/2} = \Sigma_{i^*(z)i}^{1/2},$$

where we defined $i^*(z) := \arg\max_{i \le n} (\Sigma^{1/2} z)_i$. Thus

$$\|\nabla f(z)\|^2 = \sum_{i=1}^{n} \Sigma_{i^*(z)i}^{1/2} \Sigma_{ii^*(z)}^{1/2} = \Sigma_{i^*(z)i^*(z)} \le \max_{i \le n} \Sigma_{ii}.$$

As $\Sigma_{ii} = \mathrm{Var}[X_i]$, the result follows immediately. □

We now complete the proof of Theorem 6.14.

Proof (Theorem 6.14). We can evidently estimate

$$\begin{aligned}
\mathbf{E}\left[\sup_{t\in T} X_t\right] &\ge \mathbf{E}\left[\max_{s\in N} \sup_{t\in B(s,\alpha\varepsilon)} X_t\right] \\
&= \mathbf{E}\left[\max_{s\in N} \left\{X_s + \mathbf{E}\left[\sup_{t\in B(s,\alpha\varepsilon)} X_t\right] + Y_s\right\}\right] \\
&\ge \mathbf{E}\left[\max_{s\in N} X_s\right] + \min_{s\in N} \mathbf{E}\left[\sup_{t\in B(s,\alpha\varepsilon)} X_t\right] - \mathbf{E}\left[\max_{s\in N}\{-Y_s\}\right],
\end{aligned}$$

where we defined

$$Y_s = \sup_{t\in B(s,\alpha\varepsilon)} \{X_t - X_s\} - \mathbf{E}\left[\sup_{t\in B(s,\alpha\varepsilon)} \{X_t - X_s\}\right].$$

By Lemma 6.15, Y_s is $\alpha^2\varepsilon^2$-subgaussian for all $s \in N$. Thus we obtain, bounding the first term using Theorem 6.5 and the last term using Lemma 5.1,

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \ge \{c - \alpha\sqrt{2}\}\varepsilon\sqrt{\log|N|} + \min_{s\in N} \mathbf{E}\left[\sup_{t\in B(s,\alpha\varepsilon)} X_t\right]$$

for some universal constant c. Choosing $\alpha = c/2\sqrt{2}$ completes the proof. □

Let us compare the lower bound of Theorem 6.14 to the chaining upper bound. An immediate difference between the two bounds is that the former is stated in terms of an ε-packing, while the latter is in terms of an ε-net. This will be taken care of using the duality between covering and packing, however, so that this difference is not a major concern at this stage. A more pressing concern is the minimum in the bound of Theorem 6.14. To emphasize this issue, let us reformulate the chaining upper bound to bring out the similarity between the two bounds: if $\mathrm{diam}(T) \le \varepsilon$ and $N \subseteq T$ is an $\alpha\varepsilon$-net, then

$$\begin{aligned}
\mathbf{E}\left[\sup_{t\in T} X_t\right] &\le c_1\varepsilon\sqrt{\log|N|} + \mathbf{E}\left[\max_{s\in N} \sup_{t\in B(s,\alpha\varepsilon)} \{X_t - X_s\}\right] \\
&\le c_1'\varepsilon\sqrt{\log|N|} + \max_{s\in N} \mathbf{E}\left[\sup_{t\in B(s,\alpha\varepsilon)} X_t\right].
\end{aligned}$$

The first inequality follows trivially from the chaining upper bound as stated at the beginning of this section, while the second bound is readily obtained by using Gaussian concentration as in the proof of Theorem 6.14. In contrast, the bound of Theorem 6.14 states that if N is an ε-packing, then

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \ge c\varepsilon\sqrt{\log|N|} + \min_{s\in N}\mathbf{E}\left[\sup_{t\in B(s,\alpha\varepsilon)} X_t\right].$$

When phrased in this manner, the two bounds appear to be very similar, with one crucial difference: in the chaining upper bound, the remainder term is the *largest* expected supremum of the Gaussian process over a ball centered at one of the points in N, while the remainder term in Theorem 6.14 is the *smallest* expected supremum over such a ball. There is no reason why the supremum of the Gaussian process over two balls of the same radius should be of the same order: in general, the remainder terms in our upper and lower bounds can be of a very different order of magnitude. The major remaining question, to be addressed in the next section, is how to overcome this problem.

For the time being, however, we would like to illustrate the idea of chaining in reverse without having to cope with the complications arising from the above problem. To this end, we will investigate in the remainder of this section a special class of Gaussian processes for which this problem does not arise.

Definition 6.16 (Stationary Gaussian process). *The Gaussian process $\{X_t\}_{t\in T}$ is called* stationary *if there exists a group G acting on T such that*

1. $d(g(t),g(s)) = d(t,s)$ for all $t,s\in T$, $g\in G$ (translation invariance).
2. For every $t,s\in T$, there exists $g\in G$ such that $t=g(s)$ (transitivity).

Of course, the key point of this definition is that for a stationary Gaussian process all balls are created equal: indeed, we have equality in distribution

$$\{X_t - X_s : t\in B(s,\varepsilon)\} \stackrel{\mathrm{d}}{=} \{X_t - X_{s'} : t\in B(s',\varepsilon)\} \quad \text{for all } s,s'\in T.$$

To see this, recall that the law of the increments of a Gaussian process is entirely determined by the natural metric d, and note that if $g\in G$ is such that $s'=g(s)$, then g maps $B(s,\varepsilon)$ isometrically onto $B(s',\varepsilon)$. Thus

$$\max_{s\in T}\mathbf{E}\left[\sup_{t\in B(s,\varepsilon)} X_t\right] = \min_{s\in T}\mathbf{E}\left[\sup_{t\in B(s,\varepsilon)} X_t\right],$$

so our upper and lower bounds are of the same order in this case.

Example 6.17 (Brownian motion). Let $\{B_t\}_{t\in\mathbb{R}}$ be two-sided Brownian motion (that is, $B_t = B'_t$ for $t\ge 0$ and $B_t = B''_{-t}$ for $t<0$, where $\{B'_t\}_{t\ge0}$ and $\{B''_t\}_{t\ge0}$ are independent standard Brownian motions). We can view the index set $\mathbb{R}$ itself as a group $G=(\mathbb{R},+)$ under addition. It is now easily seen that Brownian motion is a stationary Gaussian process: transitivity is obvious, while translation invariance can be read off from $d(t,s)=\sqrt{|t-s|}$.

Example 6.18 (Random Fourier series). A classical application of stationary Gaussian processes is to develop an understanding of Fourier series with random coefficients. Let g_k and g'_k be i.i.d. $N(0,1)$ random variables, and let c_k be coefficients such that $\sum_k c_k^2 < \infty$. Define for $t \in S^1 = [0, 2\pi[$ the process

$$X_t = \sum_{k=0}^{\infty} c_k \{g_k \sin kt + g'_k \cos kt\}.$$

Then $\{X_t\}_{t\in S^1}$ is a stationary Gaussian process for the group of rotations of the circle S^1. Indeed, transitivity is obvious, and is it not difficult to compute $d(t,s)^2 = 2\sum_k c_k^2\{1 - \cos(k(t-s))\}$ which is evidently translation-invariant.

Under the stationarity assumption, we have seen that the upper bound we have used in a single iteration of the chaining argument is matched by an essentially equivalent lower bound. Therefore, in this setting, we expect that the chaining bound obtained in the previous chapter is tight. To prove this, little remains but to run the chaining argument in reverse.

Theorem 6.19 (Fernique). *Let $\{X_t\}_{t\in T}$ be a stationary separable Gaussian process. Then we can estimate for some universal constants c_1, c_2*

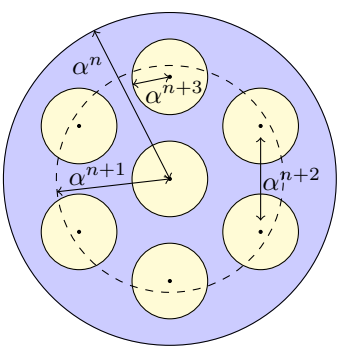
$$c_1 \int_0^\infty \sqrt{\log N(T,d,\varepsilon)}\, d\varepsilon \le \mathbf{E}\left[\sup_{t\in T} X_t\right] \le c_2 \int_0^\infty \sqrt{\log N(T,d,\varepsilon)}\, d\varepsilon.$$

Proof. As the Gaussian process is stationary, all balls behave in the same way. Thus we will lighten our notation by defining $B(\varepsilon) = B(t_0,\varepsilon)$ for some fixed by arbitrary point $t_0 \in T$. This will play the role of our "representative ball".

Let us begin by applying Theorem 6.14 at the scale α^n. Choose N_n to be a maximal α^{n+2}-packing of the ball $B(\alpha^{n+1})$. Then we have

$$\bigcup_{s\in N_n} B(s,\alpha^{n+3}) \subseteq B(\alpha^n),$$

as $d(t_0,t) \le d(t_0,s) + d(s,t) \le \alpha^{n+1} + \alpha^{n+3} \le \alpha^n$ for every $s \in N_n$ and $t \in B(s,\alpha^{n+3})$. This situation is illustrated in the following figure:



By the maximality of the packing N_n, the duality between packing and covering numbers yields $|N_n| \geq N(B(\alpha^{n+1}), d, \alpha^{n+2})$. Thus Theorem 6.14 yields

$$\mathbf{E}\left[\sup_{t\in B(\alpha^n)} X_t\right] \geq c\alpha^{n+2}\sqrt{\log N(B(\alpha^{n+1}), d, \alpha^{n+2})} + \mathbf{E}\left[\sup_{t\in B(\alpha^{n+3})} X_t\right],$$

where we have used stationarity and $B(s, \alpha^{n+3}) \subset B(\alpha^n)$ to conclude that

$$\min_{s\in N_n} \mathbf{E}\left[\sup_{t\in B(\alpha^n)\cap B(s,\alpha^{n+3})} X_t\right] = \mathbf{E}\left[\sup_{t\in B(\alpha^{n+3})} X_t\right].$$

(the term on the left being the one that arises in Theorem 6.14).

We now iterate this bound. Let k_0 be the largest integer such that $\alpha^{k_0} \geq \mathrm{diam}(T)$. If we start the iteration at any $n \leq k_0$, then we obtain

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \geq c\sum_{k\geq 0} \alpha^{n+3k+2}\sqrt{\log N(B(\alpha^{n+3k+1}), d, \alpha^{n+3k+2})}.$$

This completes the core part of the proof of Theorem 6.19: we have obtained a multiscale lower bound on the supremum of the Gaussian process by "chaining in reverse". However, at first sight the lower bound looks a little different than the upper bound of Theorem 5.24. The difference proves to be cosmetic, and we will presently "fix" the discrepancy between the two bounds.

First, note that the terms in the above sum "skip" from scale α^k to α^{k+3}, rather than summing over all $k \in \mathbb{Z}$. As the starting point n is arbitrary, however, we can fix this by averaging over $n = k_0, k_0 - 1, k_0 - 2$. This yields

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \geq \frac{c}{3}\sum_{k\in\mathbb{Z}} \alpha^{k+1}\sqrt{\log N(B(\alpha^k), d, \alpha^{k+1})}.$$

The remaining problem with this lower bound is that it contains covering numbers of the form $N(B(\alpha^k), d, \alpha^{k+1})$, while our upper bound is phrased in terms of covering numbers of the entire set $N(T, d, \alpha^{k+1})$. To fix this, let us do some covering number gymnastics. Suppose we can cover T by m balls of radius α^k, and that each ball of radius α^k can be covered by m' balls of radius α^{k+1}. Then clearly T can be covered by mm' balls of radius α^{k+1}. We can choose $m = N(T, d, \alpha^k)$ and $m' = N(B(\alpha^k), d, \alpha^{k+1})$ (using stationarity to argue that the covering number of any ball $B(s, \alpha^k)$ is equal to that of our representative ball $B(\alpha^k)$). A moment's reflection will show that we proved

$$N(T, d, \alpha^{k+1}) \leq N(T, d, \alpha^k)\, N(B(\alpha^k), d, \alpha^{k+1}).$$

This sort of reasoning is useful in many problems involving covering numbers. In the present setting, plugging this identity into the above bound yields

$$\begin{aligned}\mathbf{E}\left[\sup_{t\in T} X_t\right] &\geq \frac{c}{3}\sum_{k\in\mathbb{Z}} \alpha^{k+1}\sqrt{\log N(T,d,\alpha^{k+1})} - \frac{c}{3}\sum_{k\in\mathbb{Z}} \alpha^{k+1}\sqrt{\log N(T,d,\alpha^{k})} \\ &= \frac{c(1-\alpha)}{3}\sum_{k\in\mathbb{Z}} \alpha^{k+1}\sqrt{\log N(T,d,\alpha^{k+1})} \\ &\geq c' \int_0^\infty \sqrt{\log N(T,d,\varepsilon)}\,d\varepsilon\end{aligned}$$

for some universal constant c', where we estimated the sum by an integral in the usual manner (cf. Problem 5.9). Note that in order to prove that the two terms in the first inequality are of the same order, we used the fact that the sum runs over all $k \in \mathbb{Z}$ and not just over multiples of three. This minor annoyance in the proof therefore does serve a purpose.

We have now proved the lower bound. The corresponding upper bound follows immediately from the previous chapter (Corollary 5.25). □

Problems

6.1 (An alternative proof of super-Sudakov). We deduced the super-Sudakov inequality from the ordinary Sudakov inequality together with Gaussian concentration. It is also possible, however, to obtain Theorem 6.14 directly from the Slepian-Fernique inequality by modifying the proof of the Sudakov inequality. The advantage of this is that it yields somewhat sharper constants. The aim of this problem is to develop this alternative proof.

For simplicity, let $\{X_t\}_{t\in T}$ be a Gaussian process on a *finite* index set T (the extension to the case of a separable Gaussian process follows readily as in the proof of Theorem 5.24). Let N be an ε-packing of (T,d).

a. For every $s \in N$, let $T_s := \{t \in T : d(t,s) \leq \frac{1}{4}\varepsilon\}$ and

$$Z_t = X_t^{(s)} - X_s^{(s)} + \tfrac{1}{4}\varepsilon g_s \quad \text{for } t \in T_s,\ s \in N,$$

where $\{X_t^{(s)}\}_{t\in T}$ are independent copies of $\{X_t\}_{t\in T}$ and g_s are independent $N(0,1)$ random variables for $s \in N$. Show that we have

$$\mathbf{E}|X_t - X_{t'}|^2 \geq \mathbf{E}|Z_t - Z_{t'}|^2 \quad \text{for all } t,t' \in \bigcup_{s\in N} T_s.$$

b. Conclude from Theorem 6.8 that

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \geq \mathbf{E}\left[\max_{s\in N}\left\{\frac{\varepsilon}{4}\,g_s + \sup_{t\in T_s}\{X_t^{(s)} - X_s^{(s)}\}\right\}\right].$$

c. Use Jensen's inequality conditionally on $\{g_s\}_{s\in N}$ to conclude that

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \geq \frac{\varepsilon}{4}\,\mathbf{E}\left[\max_{s\in N} g_s\right] + \min_{s\in N}\mathbf{E}\left[\sup_{t\in T_s} X_t\right],$$

and conclude that Theorem 6.14 holds for $\alpha = \frac{1}{4}$.

6.2 (Rectangles). Consider the Gaussian process $\{X_t\}_{t\in\{-1,1\}^n}$ of the form

$$X_t = \sum_{k=1}^{n} g_k t_k a_k,$$

where $a_1 > \cdots > a_n > 0$ are given constants and $g_1, \ldots, g_n$ are i.i.d. $N(0,1)$. Such a process is called a rectangle (as the index set $(\{-1,1\}^n, d)$ has the same geometry as the corners of a rectangle in $(\mathbb{R}^n, \|\cdot\|)$).

a. Show that

$$\mathbf{E}\left[\sup_{t\in\{-1,1\}^n} X_t\right] = \sqrt{\frac{2}{\pi}} \sum_{k=1}^{n} a_k.$$

b. Argue that $\{X_t\}_{t\in\{-1,1\}^n}$ is a stationary Gaussian process, so that

$$\int_0^\infty \sqrt{\log N(\{-1,1\}^n, d, \varepsilon)}\, d\varepsilon \asymp \sum_{k=1}^{n} a_k.$$

c. Attempt to verify this conclusion by estimating covering numbers and computing the entropy integral directly. (This is surprisingly hard!)

d. Let $a_k = 1/k$. Show that for every $n \geq 1$

$$\sup_{\varepsilon>0} \varepsilon\sqrt{\log N(\{-1,1\}^n, d, \varepsilon)} \leq c \quad \text{and} \quad \sum_{k=1}^{n} a_k \gtrsim \log n$$

for some universal constant c. Therefore, while the chaining bound of Theorem 5.24 is sharp, Sudakov's inequality is far from sharp in this example.

6.3 (A nonstationary process). Consider the Gaussian process $\{X_n\}_{n\in\mathbb{N}}$

$$X_n = \frac{g_n}{\sqrt{1+\log n}},$$

where $\{g_n\}_{n\in\mathbb{N}}$ are i.i.d. $N(0,1)$. This process is most definitely not stationary.

a. Show that

$$\mathbf{E}\left[\sup_{n\in\mathbb{N}} X_n\right] < \infty.$$

b. Show that

$$\int_0^\infty \sqrt{\log N(\mathbb{N}, d, \varepsilon)}\, d\varepsilon = \infty,$$

so the conclusion of Theorem 6.19 can indeed fail in the nonstationary case.

c. To gain some insight into the problem, compute the quantity

$$\mathbf{E}\left[\sup_{d(n,m)\le\varepsilon} X_n\right]$$

for different $m \in \mathbb{N}$. Conclude that while one needs $N(\mathbb{N}, d, \varepsilon)$ balls of radius ε to cover $\mathbb{N}$ (and $N(\mathbb{N}, d, \varepsilon) \uparrow \infty$ as $\varepsilon \downarrow 0$), the expected supremum of the Gaussian process over all but one of these balls vanishes. Thus the remainder terms in our chaining upper and lower bounds are not comparable (in fact, in this case it is clearly the *upper* bound that it inefficient).

6.4 (An improved chaining argument). Let $\{X_t\}_{t\in T}$ be a (nonstationary) Gaussian process. In order to compare the super-Sudakov inequality to the chaining upper bound, we used Gaussian concentration to reformulate the upper bound as follows: if $\mathrm{diam}(T) \le \varepsilon$ and $N \subseteq T$ is an $\alpha\varepsilon$-net, then

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \le c\varepsilon\sqrt{\log|N|} + \max_{s\in N}\mathbf{E}\left[\sup_{t\in B(s,\alpha\varepsilon)} X_t\right].$$

To goal of this problem is to note that chaining using this improved inequality will in fact yield a slightly improved version of Corollary 5.25:

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \le c_1 \sup_{t\in T}\int_0^\infty \sqrt{\log N(B(t, c_2\varepsilon), d, \varepsilon)}\, d\varepsilon$$

for universal constants $c_1, c_2 > 1$.

a. Prove the above inequality.

b. Find an example where this inequality is sharp, but Corollary 5.25 is not.
Hint: let $\mathbb{T}$ be a (not necessarily regular) finite rooted tree with root $t_0 \in \mathbb{T}$ and leaves $T \subseteq \mathbb{T}$. Assume that all leaves have the same depth n. For every leaf $t \in T$, denote by $\pi_0(t), \pi_1(t), \ldots, \pi_n(t)$ the unique path in the tree from the root $\pi_0(t) = t_0$ to the leaf $\pi_n(t) = t$. Attach to each vertex $s \in \mathbb{T}$ an i.i.d. $N(0,1)$ random variable ξ_s, and define $\{X_t\}_{t\in T}$ as $X_t = \sum_{k=0}^n \beta^k \xi_{\pi_k(t)}$. Choose $\beta < 1$ and an irregular tree $\mathbb{T}$ carefully to construct the example.

c. Find an example where also the present inequality is not sharp.
Hint: consider Problem 6.3.

6.3 The majorizing measure theorem

In the previous section we developed the machinery needed to run the chaining argument in reverse. However, our upper bound involved a maximum over the expected supremum of different balls, while our lower bound involved a minimum over the expected supremum of different balls. In the stationary case,

these quantities are of the same order and we were able to run the chaining argument to its completion. In the general case, however, the supremum over different balls of the same radius can be of a very different order of magnitude, and thus our upper and lower bounds do not match. To close this gap, it will be essential to take the inhomogeneity of the process into account.

In this section, we will develop our most efficient incarnation of the chaining method that achieves precisely this goal. There are two problems to be overcome. First, we must understand how to obtain matching upper and lower bounds at the level of a single iteration of the chaining argument. This will prove to be surprisingly straightforward: we have already encountered most of the ideas in the previous section, and it remains to note that they can be implemented more efficiently. Next, we must understand how to iterate these inequalities so that we ultimately obtain matching upper and lower bounds. This will prove to be the most clever part of the argument, and we will see that we must organize the chaining argument carefully in order to retain the duality between packing and covering at different scales. The payoff, however, will be a remarkable achievement: a complete understanding of the expected supremum of a Gaussian process in terms of chaining! With that accomplishment to look forward to, let us proceed to making it happen.

Our first step is a seemingly innocuous observation. In the super-Sudakov inequality of Theorem 6.14, we could choose N to be any ε-packing. If we did not have the remainder term, then the best possible bound would be obtained by choosing a *maximal* packing, as we did in the Sudakov inequality of Theorem 6.5. However, in the super-Sudakov inequality, this is not necessarily the best idea: if we increase the size of the packing, then evidently the size of the remainder term will decrease, and thus we could "miss" important parts of the index set that will arise in a later iteration of the chaining argument. By resisting the temptation to be greedy, we obtain an immediate improvement of the super-Sudakov inequality without any additional effort.

Corollary 6.20 (Super-Sudakov improved). *Let $\{X_t\}_{t\in T}$ be a separable Gaussian process and let $N = \{t_1, \ldots, t_r\}$ be an ε-packing of (T, d). Then*

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \geq \min_{\sigma} \max_{k\leq r} \left\{ c\varepsilon\sqrt{\log \sigma(k)} + \mathbf{E}\left[\sup_{t\in B(t_k,\alpha\varepsilon)} X_t\right]\right\}.$$

where the minimum is over all permutations σ of $\{1, \ldots, r\}$.

While we have phrased this result as a minimum over permutations for aesthetic reasons, note that it is clear what is the optimal permutation: it is given by $\sigma(k_i) = i$ if we rank the remainder terms in decreasing order

$$\mathbf{E}\left[\sup_{t\in B(t_{k_1},\alpha\varepsilon)} X_t\right] \geq \mathbf{E}\left[\sup_{t\in B(t_{k_2},\alpha\varepsilon)} X_t\right] \geq \cdots \geq \mathbf{E}\left[\sup_{t\in B(t_{k_r},\alpha\varepsilon)} X_t\right].$$

Thus the permutation σ captures precisely the inhomogeneity of the process: "fatter" balls $B(t_k, \alpha\varepsilon)$ end up with smaller labels $\sigma(k)$.

Proof. Sort the packing $N = \{t_{k_1}, \ldots, t_{k_r}\}$ as indicated above. If we apply Theorem 6.14 to the smaller packing $\{t_{k_1}, \ldots, t_{k_\ell}\}$ only, we evidently obtain

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \ge c\varepsilon\sqrt{\log \ell} + \mathbf{E}\left[\sup_{t\in B(t_{k_\ell},\alpha\varepsilon)} X_t\right] \quad \text{for any } \ell \le r.$$

The result follows immediately by optimizing this bound over ℓ. □

It might be unclear at this point that we have made significant progress. Indeed, while we now capture the inhomogeneity of the Gaussian process in the lower bound, we have essentially just rearranged our previous lower bound without making any fundamental improvement. In particular, we are still far removed from our chaining upper bound. However, now that we have reformulated our lower bound in this illuminating manner, it will quickly become clear that it is in fact the *upper* bound that is inefficient and fails to capture the inhomogeneity of the process. We will presently correct this.

Proposition 6.21 (Super-chaining). *Let $\{X_t\}_{t\in T}$ be a separable Gaussian process. If $\operatorname{diam}(T) \le \varepsilon$ and $\{A_1, \ldots, A_r\}$ is a partition of T, then*

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \le \min_{\sigma} \max_{k\le r} \left\{3\varepsilon\big\{1 + \sqrt{\log \sigma(k)}\big\} + \mathbf{E}\left[\sup_{t\in A_k} X_t\right]\right\}.$$

The improved upper bound of Proposition 6.21 captures the inhomogeneity of the Gaussian process in a completely analogous manner to the lower bound of Corollary 6.20. To prove this result, we must eliminate the inefficiency in the proof of our previous upper bound. Somewhat surprisingly, it turns out that this inefficiency arises in the very first result we proved about maxima of random variables: Lemma 5.1. The following apparently minor improvement, which is proved using a simple union bound, yields precisely what we need.

Lemma 6.22. *Let $Z_1, \ldots, Z_n$ be σ^2-subgaussian random variables. Then*

$$\mathbf{E}\left[\max_{k\le n}\{Z_k - \mathbf{E}[Z_k] - 2\sigma\sqrt{\log k}\}\right] \le 3\sigma.$$

Proof. We can assume without loss of generality that $\mathbf{E}[Z_k] = 0$ for all k. Using a union bound and the subgaussian property, we evidently have

$$\begin{aligned}\mathbf{P}\left[\max_{k\le n}\{Z_k - 2\sigma\sqrt{\log k}\} \ge t\right] &\le \sum_{k=1}^n \mathbf{P}[Z_k \ge 2\sigma\sqrt{\log k} + t] \\ &\le \sum_{k=1}^n e^{-(2\sigma\sqrt{\log k}+t)^2/2\sigma^2} \le e^{-t^2/2\sigma^2} \sum_{k=1}^n \frac{1}{k^2}.\end{aligned}$$

We therefore estimate

$$\mathbf{E}\left[\max_{k\le n}\{Z_k - 2\sigma\sqrt{\log k}\}\right] \le \int_0^\infty e^{-t^2/2\sigma^2}\, dt \sum_{k=1}^\infty \frac{1}{k^2} = \frac{\pi^{5/2}}{6\sqrt{2}}\,\sigma.$$

For simplicity we estimate the ugly constant $\pi^{5/2}/6\sqrt{2} \approx 2.06$ by 3. □

We can now complete the proof of Proposition 6.21.

Proof (Proposition 6.21). Fix any $t_0 \in T$. As $\mathbf{E}[X_{t_0}] = 0$, we can estimate

$$\begin{aligned}\mathbf{E}\left[\sup_{t\in T} X_t\right] &= \mathbf{E}\left[\max_{k\le r}\sup_{t\in A_k}\{X_t - X_{t_0}\}\right] \\ &= \mathbf{E}\left[\max_{k\le r}\left\{2\varepsilon\sqrt{\log k} + \mathbf{E}\left[\sup_{t\in A_k} X_t\right] + \{Y_k - 2\varepsilon\sqrt{\log k}\}\right\}\right],\end{aligned}$$

where we have defined

$$Y_k = \sup_{t\in A_k}\{X_t - X_{t_0}\} - \mathbf{E}\left[\sup_{t\in A_k}\{X_t - X_{t_0}\}\right].$$

As $d(t,t_0) \le \mathrm{diam}(T) \le \varepsilon$, the random variables Y_k are ε^2-subgaussian by Lemma 6.15. Thus Lemma 6.22 immediately yields

$$\mathbf{E}\left[\max_{k\le r}\{Y_k - 2\varepsilon\sqrt{\log k}\}\right] \le 3\varepsilon,$$

and thus we obtain

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \le \max_{k\le r}\left\{3\varepsilon\{1+\sqrt{\log k}\} + \mathbf{E}\left[\sup_{t\in A_k} X_t\right]\right\}.$$

But note that this result holds for any ordering of $\{A_1,\ldots,A_r\}$. Replacing A_i by $A_{\sigma^{-1}(i)}$ and optimizing over permutations σ concludes the proof. □

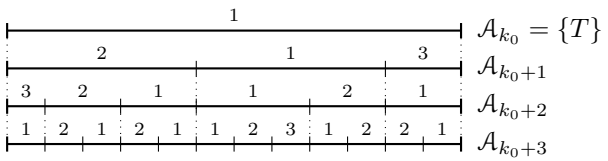
Up to the duality between packing and covering, we have now essentially obtained matching upper and lower bounds in Corollary 6.20 and Proposition 6.21 for a single iteration of the chaining argument. We have therefore finally reached a point at which it should no longer appear to be a major miracle that we can obtain matching upper and lower bounds on supremum of a Gaussian process. However, these bounds will be necessarily more sophisticated than in Theorem 5.24, as we must now explicitly keep track of the inhomogeneity of the process in each iteration of the chaining argument. In particular, it is no longer sufficient just to choose any sequence of coverings of the index set T at different scales: we must sort each of the covers in accordance with the permutations σ in Corollary 6.20, which should be thought of as ranking the elements of the cover in order of decreasing "fatness". This requires some amount of bookkeeping, which can be done in different ways. The device that we will choose for this purpose, given in the following definition, is designed to be as close as possible to the statement of Proposition 6.21.

Recall that an increasing sequence of partitions $\{\mathcal{A}_n\}_{n\in\mathbb{Z}}$ is a family of partitions $\mathcal{A}_n$ such that every $B \in \mathcal{A}_{n+1}$ is contained in some set $A \in \mathcal{A}_n$. The set of *children* of a set $A \in \mathcal{A}_n$ is denoted $c(A) := \{B \in \mathcal{A}_{n+1} : B \subseteq A\}$. For any $t \in T$, we denote by $A_n(t)$ the unique set $A \in \mathcal{A}_n$ that contains t.

Definition 6.23 (Labelled net). *A pair* $(\mathcal{A}, \ell)$ *if called a* labelled net *if*

1. $\mathcal{A} = \{\mathcal{A}_n\}_{n\in\mathbb{Z}}$ *is an increasing sequence of partitions of* T.
2. $\operatorname{diam}(A) \le 2\alpha^n$ *for every* $A \in \mathcal{A}_n$, $n \in \mathbb{Z}$.
3. $\ell : \mathcal{A} \to \mathbb{N}$ *satisfies* $\{\ell(B) : B \in c(A)\} = \{1, \ldots, |c(A)|\}$ *for all* $A \in \mathcal{A}$.

That is, a labelled net is an increasing family of partitions $\mathcal{A}$, together with a labeling ℓ that defines an ordering among all elements of each partition that share the same parent. Such a construction is illustrated in the following figure.



Each horizontal interval represents a partition of T, and the numbers indicate an assignment of labels to each partition element. The dotted lines indicate the children of each partition element. Note that each $t \in T$ defines a vertical slice through this picture. Listing the labels one encounters along this slice from top to bottom gives the sequence $\ell(A_{k_0}(t))$, $\ell(A_{k_0+1}(t))$, ...

We are now ready to state a form of the ultimate chaining bound for Gaussian processes due to Talagrand.

Theorem 6.24 (The majorizing measure theorem). *Let* $\{X_t\}_{t\in T}$ *be a separable Gaussian process. Then we have for universal constants* c_1, c_2, α

$$c_1\gamma(T) \le \mathbf{E}\left[\sup_{t\in T} X_t\right] \le c_2\gamma(T).$$

Here we defined

$$\gamma(T) := \inf_{(\mathcal{A},\ell)} \sup_{t\in T} \sum_{k\in\mathbb{Z}} \alpha^k \sqrt{\log \ell(A_k(t))},$$

where the infimum is taken over all labelled nets $(\mathcal{A}, \ell)$.

Let us take a moment to consider what we have achieved. Theorem 6.24 gives matching upper and lower bounds for the expected supremum of a Gaussian process. We can therefore conclude that we have completely understood the magnitude of the supremum of Gaussian processes in terms of chaining! On the other hand, the chaining object that arises in Theorem 6.24 is of a very sophisticated form (necessarily so, as we must account explicitly for the inhomogeneity of the Gaussian process): to find a good bound in this manner we must be able to construct a "good" labelled net. Unlike the covering numbers that arose in Theorem 5.24, which are often easy to estimate, constructing good labelled nets "by hand" in inhomogeneous situations is generally an exceedingly difficult task. It may therefore be unclear at this point that Theorem

6.24 has any practical utility. It turns out that Theorem 6.24 is a powerful tool that makes it possible to prove useful and deep results about the suprema of random processes that do not appear to be readily established by other means. We will encounter some examples of such results in the next section.

Remark 6.25. The bookkeeping in the chaining argument can be done in several different ways. We have chosen the labelled net as the basic object in our development of Theorem 6.24 as its definition is tailored to the application of Proposition 6.21. The name "majorizing measure theorem" refers to a different method of bookkeeping that was used in the original formulation of Theorem 6.24, where role of the labels ℓ is replaced by the definition of a measure on the index set T that assigns larger mass to "fatter" partition elements. This idea will be developed in Problem 6.7 below. Yet another formulation, in terms of admissible nets, dispenses entirely of the need for explicitly labelling partition elements. This idea will be developed in the next section.

Let us turn to the proof of Theorem 6.24. We begin by proving the upper bound, which is an almost immediate consequence of Proposition 6.21.

Proof (Upper bound). As in the proof of Theorem 5.24, it suffices to consider the case that T is a finite set. In the following, we fix a labelled net $(\mathcal{A}, \ell)$, and let k_0 be the largest integer such that $\mathcal{A}_{k_0} = \{T\}$. We aim to show that

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \le c' \sup_{t\in T} \sum_{k>k_0} \alpha^k \sqrt{\log \ell(A_k(t))}.$$

Note that if $k_0 = -\infty$, then the right-hand side of this inequality is infinite and the statement is trivial. We may therefore assume that $k_0 > -\infty$.

The proof is now easily completed. By Proposition 6.21, we have

$$\mathbf{E}\left[\sup_{t\in A} X_t\right] \le \max_{B\in c(A)} \left\{6\alpha^k\big\{1+\sqrt{\log \ell(B)}\big\} + \mathbf{E}\left[\sup_{t\in B} X_t\right]\right\}$$

for any $A \in \mathcal{A}_k$. Iterating this inequality n times starting at $k = k_0$ yields

$$\begin{aligned}\mathbf{E}\left[\sup_{t\in T} X_t\right] &\le \sup_{t\in T}\left\{\sum_{k=k_0}^{k_0+n-1} 6\alpha^k\big\{1+\sqrt{\log \ell(A_{k+1}(t))}\big\} + \mathbf{E}\left[\sup_{s\in A_{k_0+n}(t)} X_s\right]\right\} \\ &\le \frac{6\alpha^{k_0}}{1-\alpha} + \frac{6}{\alpha}\sup_{t\in T}\sum_{k>k_0}\alpha^k\sqrt{\log \ell(A_k(t))}\end{aligned}$$

provided that n is chosen sufficiently large. Here we have used that as T is assumed to be finite, the remainder term vanishes uniformly in t for large n.

It remains to eliminate the additive constant. To this end, note that by the definition of k_0, there exists $t \in T$ such that $\ell(A_{k_0+1}(t)) = 2$, so that

$$\alpha^{k_0+1}\sqrt{\log 2} \le \sup_{t\in T}\sum_{k>k_0}\alpha^k\sqrt{\log \ell(A_k(t))}.$$

The proof is now easily completed with $c_2 = 6\alpha^{-1}\{1 + 1/(1-\alpha)\sqrt{\log 2}\}$. □

We now turn to the lower bound. The difficulty here is that the lower bound of Corollary 6.20 requires a packing, while the labelled net is defined in terms partitions. Of course, the duality between packing and covering will be essential here, but the situation proves to be somewhat more delicate than we have previously encountered. To understand the problem, let us try to apply a naive duality argument to the first chaining iteration. Assume for simplicity that $\operatorname{diam}(T) = \alpha^{k_0}$. To apply the lower bound, we first choose a maximal α^{k_0+1}-packing $N_{k_0+1} = \{t_1, \ldots, t_r\}$ of T. Then Corollary 6.20 gives

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \geq \max_{k\leq r}\left\{c'\alpha^{k_0+1}\sqrt{\log\sigma(k)} + \mathbf{E}\left[\sup_{t\in B(t_k,\alpha^{k_0+2})} X_t\right]\right\}$$

for a suitable choice of σ. We now define the first nontrivial partition $\mathcal{A}_{k_0+1} = \{A_1, \ldots, A_r\}$ of our labelled net by setting $A_k = \{t \in T : \pi_{k_0+1}(t) = t_k\}$, and define the label $\ell(A_k) = \sigma(k)$. By maximality of the packing, each set A_k has diameter at most $2\alpha^{k_0+1}$ as required. Then Proposition 6.21 gives

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \leq \max_{k\leq r}\left\{c\alpha^{k_0+1}\sqrt{\log\sigma(k)} + \mathbf{E}\left[\sup_{t\in A_k} X_t\right]\right\}.$$

Unfortunately, we are now stuck: while the primary terms in the upper and lower bounds match, the remainder terms are not necessarily comparable. Indeed, in the lower bound, we only see the supremum of the process over small balls $B(t_k, \alpha^{k_0+2})$ centered at each point in the packing, while in the upper bound we have the supremum over every element of a partition of the set. If we attempt to iterate this procedure, we will therefore miss in the lower bound all elements of the partitions $\mathcal{A}_n$ in subsequent stages $n \geq k_0 + 1$ that are not included in one of the balls $B(t_k, \alpha^{k_0+2})$.

The solution to this problem lies in a clever organization of the duality argument. Rather than choosing *any* maximal packing N_{k_0+1}, we will choose the points $t_1, \ldots, t_r$ in such a way that the expected supremum of the process over each of the balls $B(t_k, \alpha^{k_0+2})$ is maximized. Because of this choice, the expected supremum of any element of a partition at a smaller scale is bounded above by the expected supremum over $B(t_k, \alpha^{k_0+2})$, and we can therefore recover all elements of the labelled net in the lower bound. In the end, the argument is not any more difficult than the naive duality argument: the key to the proof is the insight that one must organize the duality argument at a given scale with subsequent iterations of the chaining argument in mind.

Proof (Lower bound). Define for any subset $A \subseteq T$

$$G(A) := \mathbf{E}\left[\sup_{t\in A} X_t\right].$$

We can assume that $G(T) < \infty$, as the lower bound is trivial otherwise. This implies that $N(T, d, \varepsilon) < \infty$ for all $\varepsilon > 0$ by Sudakov's inequality, and thus $\operatorname{diam}(T) < \infty$. Let k_0 be the largest integer such that $2\alpha^{k_0} \geq \operatorname{diam}(T)$.

To prove the lower bound, we must construct a labelled net $(\mathcal{A}, \ell)$ so that

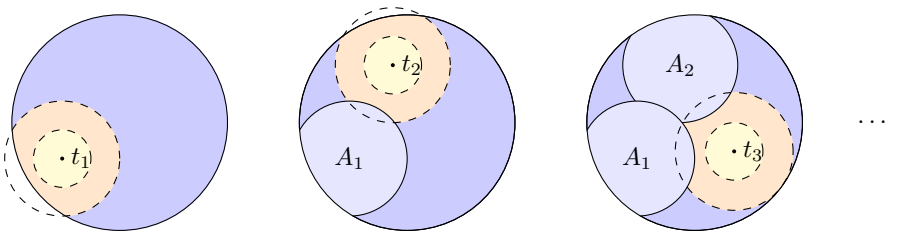
$$G(T) \geq c_1 \sum_{k \in \mathbb{Z}} \alpha^k \sqrt{\log \ell(A_k(t))}$$

for every $t \in T$. To this end, we first let $\mathcal{A}_k = \{T\}$ for all $k \leq k_0$ (with $\ell(T) = 1$). We now construct $\mathcal{A}_k$ for $k > k_0$ iteratively in the following manner.

Suppose $\mathcal{A}_k$ has been constructed. We will construct $\mathcal{A}_{k+1}$ by partitioning every element $A \in \mathcal{A}_k$ into smaller subsets as follows.

1. Choose $t_1 \in A$ so that $G(A \cap B(t_1, \alpha^{k+2}))$ is maximized.
2. Let $A_1 = A \cap B(t_1, \alpha^{k+1})$ and $\ell(A_1) = 1$.
3. Choose $t_2 \in A \backslash A_1$ so that $G(A \backslash A_1 \cap B(t_2, \alpha^{k+2}))$ is maximized.
4. Let $A_2 = A \backslash A_1 \cap B(t_2, \alpha^{k+1})$ and $\ell(A_2) = 2$.
5. Choose $t_3 \in A \backslash (A_1 \cup A_2)$ so that $G(A \backslash (A_1 \cup A_2) \cap B(t_3, \alpha^{k+2}))$ is maximized.
6. ... etc.

This construction is illustrated in the following figure:



The optimization over the choice of t_i ensures that $G(H) \leq G(A_i \cap B(t_i, \alpha^{k+2}))$ for any set $H \subseteq A_i$ that is contained in a ball of radius α^{k+2}. This will allow us to control the remainder term in Corollary 6.20. On the other hand, in each stage we remove from the set A a ball $B(t_i, \alpha^{k+1})$ with a larger radius α^{k+1}. This ensures that $d(t_i, t_j) \geq \alpha^{k+1}$, so that $\{t_1, t_2, \ldots\}$ form an α^{k+1}-packing of A as is required to apply Corollary 6.20. This also implies that the above construction must terminate after a finite number of steps, as the set T has finite packing numbers (as $N(T, d, \varepsilon) < \infty$ for all $\varepsilon > 0$).

Suppose that the above construction terminates after r steps. Then $\{A_1, \ldots, A_r\}$ must be a partition of A, each A_i has a distinct label $\ell(A_i) = i$, and $\operatorname{diam}(A_i) \leq 2\alpha^{k+1}$ by construction. By partitioning every $A \in \mathcal{A}_k$ in this manner, we have constructed a labelled partition $\mathcal{A}_{k+1}$ of T that satisfies all the properties required of a labelled net. We now iterate this process to construct $\mathcal{A}_{k+2}, \mathcal{A}_{k+3}$, and so forth, to obtain a labelled net $(\mathcal{A}, \ell)$.

Now consider again $A \in \mathcal{A}_k$ and the partition $\{A_1, \ldots, A_r\}$ and packing $\{t_1, \ldots, t_r\}$ constructed above. As $G(B(t_i, \alpha^{k+2}))$ is decreasing in i, we have

$$G(A) \geq \max_{i \leq r} \{c\alpha^{k+1} \sqrt{\log \ell(A_i)} + G(B(t_i, \alpha^{k+2}))\}$$

by Corollary 6.20. Now note that for any $t \in A_i$, we have $A_k(t) = A$, $A_{k+1}(t) = A_i$, $A_{k+3}(t) \subseteq A_i$, and $\operatorname{diam}(A_{k+3}(t)) \le 2\alpha^{k+3} \le \alpha^{k+2}$. Thus $G(A_{k+3}(t)) \le G(B(t_i, \alpha^{k+2}))$ by the maximality property of t_i, and we obtain

$$G(A_k(t)) \ge c\alpha^{k+1}\sqrt{\log \ell(A_{k+1}(t))} + G(A_{k+3}(t)).$$

This identity holds for every $t \in T$ and $k \ge k_0$. As in the proof of Theorem 6.19, this inequality "skips" from scale α^k to α^{k+3}, so we can iterate starting at $k = k_0, k_0 - 1, k_0 - 2$ and average these lower bounds to obtain

$$G(T) \ge \frac{c}{3}\sum_{k\in\mathbb{Z}} \alpha^k \sqrt{\log \ell(A_k(t))}.$$

As this holds for every $t \in T$, the proof is complete. □

Remark 6.26. Throughout this section, we have fixed α as defined in Theorem 6.14. All our constructions, including the definition of a labelled net, were stated in terms of this universal constant. However, it should be noted that while α must be sufficiently small to ensure the validity of Theorem 6.14, the precise value of α has no particular significance: in particular, we can replace α by any $\beta < \alpha$ throughout at the expense only of changing the universal constants that appear in Theorem 6.24. In view of Problem 6.1, we may therefore fix an arbitrary value $\alpha \le \frac{1}{4}$ throughout this section.

Problems

6.5 (Classical chaining and labelled nets). As the chaining functional $\gamma(T)$ of Theorem 6.24 is equivalent to the supremum of the Gaussian process up to universal constants, any upper bound on the latter must also be an upper bound for $\gamma(T)$ up to a universal constant. This is the case, in particular, for all the chaining bounds that we constructed previously. It is straightforward but instructive, however, to give a direct proof that

$$\gamma(T) \lesssim \int_0^\infty \sqrt{\log N(T, d, \varepsilon)}\, d\varepsilon$$

by constructing a simple labelled net that witnesses the upper bound. Similarly, give a direct proof of the improved chaining bound

$$\gamma(T) \lesssim \sup_{t\in T} \int_0^\infty \sqrt{\log N(B(t, c\varepsilon), d, \varepsilon)}\, d\varepsilon$$

that was investigated in Problem 6.4 above.

6.6 (A nonstationary process revisited). In Problem 6.3 we considered the decidedly nonstationary Gaussian process $\{X_n\}_{n\in\mathbb{N}}$ defined by

$$X_n = \frac{g_n}{\sqrt{1 + \log n}},$$

where $\{g_n\}_{n\in\mathbb{N}}$ are i.i.d. $N(0,1)$. The expected supremum of this process is finite, but none of the chaining bounds that we obtained previously was able to capture this fact (see Problems 6.3 and 6.4). As Theorem 6.24 is sharp, however, there must exist a labelled net that witnesses the finiteness of $\mathbf{E}[\sup_n X_n]$. Construct such a labelled net explicitly.

Hint: choose partitions of the form $\mathcal{A}_k = \{\{1\},\{2\},\ldots,\{n_k\},\mathbb{N}\cap]n_k,\infty[\}$.

6.7 (Majorizing measures). In the original formulation of Theorem 6.24, the bookkeeping in the chaining argument was not done in terms of labelled nets but rather in terms of "majorizing measures". The goal of this problem is to develop this alternative formulation of Theorem 6.24.

We begin by proving a discrete version of the majorizing measure bound

$$\gamma(T) \asymp \inf_{(\mathcal{A},\mu)} \sup_{t\in T} \sum_{k\in\mathbb{Z}} \alpha^k \sqrt{\log \frac{1}{\mu(A_k(t))}} =: \tilde\gamma(T),$$

where $\mathcal{A} = \{A_k\}_{k\in\mathbb{Z}}$ is an increasing sequence of partitions of T such that $\mathrm{diam}(A) \le 2\alpha^n$ for all $A \in \mathcal{A}_n$, and μ is a probability measure on T. The *majorizing measure* μ here plays the role of the labels in the definition of $\gamma(T)$: evidently μ should assign larger mass to "fatter" partition elements.

a. Show that $\gamma(T) \le \tilde\gamma(T)$.

Hint: if $p_1 \ge p_2 \ge \cdots \ge p_r \ge 0$ and $\sum_{i=1}^r p_i \le 1$, then $p_i \le 1/i$ for every i.

To establish the converse inequality, we must be able to construct a majorizing measure μ from labels ℓ. The problem here is that $1/\mu(A_k(t))$ must be increasing in k, while there is no ordering relation between the labels $\ell(A_k(t))$. The appropriate property is easily engineered, however, by "integrating by parts".

b. Let $\{b_k\}_{k\in\mathbb{Z}}$ be any sequence such that $b_k = 0$ for all k sufficiently small. Prove the elementary "integration by parts" identity

$$\sum_{k\in\mathbb{Z}} \alpha^k b_k = (1-\alpha)\sum_{k\in\mathbb{Z}} \alpha^k B_k, \qquad B_k := \sum_{m\le k} b_m.$$

c. Conclude that

$$\gamma(T) \gtrsim \inf_{(\mathcal{A},\ell)} \sup_{t\in T} \sum_{k\in\mathbb{Z}} \alpha^k \sqrt{\log \prod_{m\le k} \ell(A_m(t))}.$$

d. Let $(\mathcal{A},\ell)$ be a labelled net, and let k_0 be the largest integer such that $\mathcal{A}_{k_0} = \{T\}$. Fix an arbitrary $t_A \in A$ for every $A \in \mathcal{A}_n$, $n \in \mathbb{Z}$. Show that

$$\sum_{A\in\mathcal{A}_k} \prod_{m\le k} \frac{1}{\ell(A_m(t_A))^2} \le \left(\frac{\pi^2}{6}\right)^{k-k_0} \le 2^{k-k_0}.$$

e. In the setting of the previous part, define the probability measure

$$\mu \propto \sum_{k>k_0} 2^{-2(k-k_0)} \sum_{A\in\mathcal{A}_k} \delta_{t_A} \prod_{m\le k} \frac{1}{\ell(A_m(t_A))^2}.$$

Show that for every $t \in T$ and $k \in \mathbb{Z}$

$$\log \frac{1}{\mu(A_k(t))} \le 2(k-k_0)\log 2 + 2\log \prod_{m\le k} \ell(A_m(t)).$$

f. Conclude that $\gamma(T) \gtrsim \tilde{\gamma}(T)$.

The original formulation of the majorizing measure theorem was in terms of an integral rather than a sum, in analogy to Corollary 5.25:

$$\gamma(T) \asymp \inf_\mu \sup_{t\in T} \int_0^\infty \sqrt{\log \frac{1}{\mu(B(t,\varepsilon))}}\, d\varepsilon =: \bar{\gamma}(T).$$

It might seem at first sight that the continuous formulation is simpler, as it does not explicitly involve a choice of partitions. However, in applications of the majorizing measure theorem, the discrete formulation is often easier to use and more natural as it is closer to the underlying chaining mechanism.

We will presently prove the continuous formulation as well.

g. Deduce from the discrete majorizing measure bound that $\gamma(T) \gtrsim \bar{\gamma}(T)$.

The converse inequality is much more difficult, as we must now construct a sequence of partitions which was somehow lost in the continuous formulation of the majorizing measure bound. In fact, we might as well construct an entire labelled net. To this end, let us define for every $A \subseteq T$ the functional

$$F(A) := \inf_\mu \sup_{t\in A} \int_0^{\mathrm{diam}(A)} \sqrt{\log \frac{1}{\mu(B(t,\varepsilon))}}\, d\varepsilon.$$

It turns out that $F(A)$ behaves very much like $G(A) := \mathbf{E}[\sup_{t\in A} X_t]$.

h. Suppose that $\alpha \le \frac{1}{8}$. Prove the following "super-Sudakov inequality" for the functional F: if N is an ε-packing of $A \subseteq T$, then

$$F(A) \ge c\varepsilon\sqrt{\log|N|} + \min_{s\in N} F(A\cap B(s,\alpha\varepsilon)).$$

Hint: use that if $B_1,\ldots,B_r$ are disjoint, then $\mu(B_i) \le 1/r$ for some i.

i. Repeat the proof of Theorem 6.24 to show that $\gamma(T) \lesssim F(T) = \bar{\gamma}(T)$.

6.4 The generic chaining, admissible nets, and trees

The majorizing measure theorem developed in the previous section completely characterizes the supremum of Gaussian processes in terms of chaining. From the fundamental viewpoint, this provides us with substantial insight into the nature of Gaussian processes. On the other hand, it is far from clear at this point that this is a *useful* result: labelled nets are intricate chaining objects that are usually difficult to construct for any given problem. In this section, we will develop some alternative formulations of the majorizing measure theorem and show how they can be used to prove some highly nontrivial results about Gaussian and subgaussian processes. While we only scratch the surface of what can be done with this machinery, the results developed in this section give a flavor of the manner in which such machinery is applied.

We begin with a simple but very important extension of Theorem 6.24. In both the upper bound and lower bound of Theoren 6.24, we have used the Gaussian nature of the process $\{X_t\}_{t\in T}$. In the lower bound, of course, we already heavily used the Gaussian property even to prove Sudakov's inequality at a single scale. In the upper bound, however, we only used Gaussian concentration in Proposition 6.21 to handle the remainder term; the rest of the proof used a simple union bound and did not use any special properties of Gaussians. On the other hand, note that all we will do with the remainder term in Proposition 6.21 is to apply the same result to it again in the next iteration of the chaining argument. If, rather than running our chaining argument one iteration at a time, we were to bound all the links in the chain at once as we did in the proof of Theorem 5.29, then Gaussian concentration is no longer needed in the upper bound. In particular, this implies that the *upper* bound in Theorem 6.24 only requires that $\{X_t\}_{t\in T}$ is subgaussian!

Theorem 6.27 (Generic chaining). *Let $\{X_t\}_{t\in T}$ be a separable subgaussian process on (T,d). Then we have for a universal constant c*

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \le c\,\gamma(T).$$

Proof. We begin by arguing as in the proof of Theorem 5.29. As usual, it suffices to assume that T is a finite set. Let $(\mathcal{A},\ell)$ be any labelled net, and let k_0 be the largest integer such that $\mathcal{A}_{k_0} = \{T\}$. Choose for every $A\in\mathcal{A}$ an arbitrary point $t_A \in A$, and define $\pi_k(t) := t_{A_k(t)}$ for every $t\in T$. As T is finite and the diameter of $A_k(t)$ decreases to zero, we evidently have

$$X_t - X_{t_0} = \sum_{k>k_0} \{X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\},$$

where $t_0 = \pi_{k_0}(t)$. This is the usual chaining identity.

Let us define a suitable function $\mathbf{u} : \mathcal{A} \to [1,\infty[$ to be chosen later. Then it follows immediately from the subgaussian assumption that

$$\mathbf{P}\big[X_{\pi_k(t)} - X_{\pi_{k-1}(t)} \ge x\alpha^{k-1}\sqrt{\log \mathbf{u}(A_k(t))}\big] \le \mathbf{u}(A_k(t))^{-x^2/8},$$

where we have used that $d(\pi_k(t), \pi_{k-1}(t)) \le \mathrm{diam}(A_{k-1}(t)) \le 2\alpha^{k-1}$ by the definition of a labelled net. We therefore obtain by the union bound that

$$\begin{aligned}\mathbf{P}[\Omega_x] &:= \mathbf{P}\big[\exists\, k > k_0,\ t \in T \text{ s.t. } X_{\pi_k(t)} - X_{\pi_{k-1}(t)} \ge x\alpha^{k-1}\sqrt{\log \mathbf{u}(A_k(t))}\big] \\ &\le \sum_{k>k_0} \sum_{A\in\mathcal{A}_k} \mathbf{u}(A)^{-x^2/8},\end{aligned}$$

while we evidently have on the event Ω_x^c

$$\sup_{t\in T}\{X_t - X_{t_0}\} \le \frac{x}{\alpha} \sup_{t\in T} \sum_{k>k_0} \alpha^k \sqrt{\log \mathbf{u}(A_k(t))}.$$

This simple computation contains the entire idea behind the generic chaining bound. The challenge is to choose the function $\mathbf{u}$ such that the bound on the supremum of the Gaussian process is as small as possible, while we can still control the probability of the bad events Ω_x (once we have a good bound on the probabilities, we obtain a bound on the expectation as usual by integration). In view of Theorem 6.24 we would really like to choose $\mathbf{u}(A) = \ell(A)$, but this is clearly not a good idea: there are many sets $A \in \mathcal{A}$ with label $\ell(A) = 1$, and thus one cannot control our bound on $\mathbf{P}[\Omega_x]$ in this manner.

To get around this problem, note that we have a lot of freedom in how to arrange a geometric sum. This idea is extremely useful in chaining arguments.

Lemma 6.28. *Let $\alpha < 1$ and $u_k \ge 1$ for all $k > k_0$. Then*

$$(1-\alpha)\sum_{k>k_0} \alpha^k \sqrt{\log U_k} \le \sum_{k>k_0} \alpha^k \sqrt{\log u_k} \quad \textit{with} \quad U_k := \prod_{k_0<m\le k} u_m.$$

Proof. As $U_k = U_{k-1}u_k$ for $k > k_0 + 1$, we can estimate

$$\begin{aligned}\sum_{k>k_0} \alpha^k \sqrt{\log U_k} &\le \sum_{k>k_0+1} \alpha^k \sqrt{\log U_{k-1}} + \sum_{k>k_0} \alpha^k \sqrt{\log u_k} \\ &= \alpha \sum_{k>k_0} \alpha^k \sqrt{\log U_k} + \sum_{k>k_0} \alpha^k \sqrt{\log u_k}.\end{aligned}$$

The inequality now follows readily. □

The advantage of this simple reformulation is that U_k is much larger than u_k, while the geometric sum differs by at most a constant factor. To put this idea to good use, let us define for every $k > k_0$ and $t \in T$

$$\mathbf{u}(A_k(t)) = 2^{k-k_0} \prod_{k_0<m\le k} \ell(A_m(t))^2.$$

Then we have on the event Ω_x^c

$$\sup_{t\in T}\{X_t - X_{t_0}\} \le \alpha^{k_0-1} x \sum_{k>0} \alpha^k \sqrt{k \log 2} + \frac{x\sqrt{2}}{\alpha(1-\alpha)} \sup_{t\in T} \sum_{k>k_0} \alpha^k \sqrt{\log \ell(A_k(t))}$$
$$\le c_1 x \sup_{t\in T} \sum_{k>k_0} \alpha^k \sqrt{\log \ell(A_k(t))}$$

using Lemma 6.28, where c_1 is a constant that depends on α only and where the second inequality follows as in the upper bound proof of Theorem 6.24. On the other hand, note that by the definition of a labelled net

$$\sum_{B\in c(A)} \frac{1}{\ell(B)^2} = \sum_{m=1}^{|c(A)|} \frac{1}{m^2} < 2$$

for every $A \in \mathcal{A}$, so that we can estimate

$$\sum_{A\in\mathcal{A}_k} \prod_{k_0<m\le k} \frac{1}{\ell(A_m(t_A))^2} = \sum_{A\in\mathcal{A}_{k-1}} \sum_{B\in c(A)} \frac{1}{\ell(B)^2} \prod_{k_0<m\le k-1} \frac{1}{\ell(A_m(t_A))^2}$$
$$< 2 \sum_{A\in\mathcal{A}_{k-1}} \prod_{k_0<m\le k-1} \frac{1}{\ell(A_m(t_A))^2} < \cdots < 2^{k-k_0}.$$

We can therefore estimate for every $x \ge 4$

$$\mathbf{P}[\Omega_x] \le \sum_{k>k_0} 2^{-(k-k_0)x^2/8} \sum_{A\in\mathcal{A}_k} \prod_{k_0<m\le k} \frac{1}{\ell(A_m(t_A))^2} \le c_2 2^{-x^2/8},$$

where c_2 is a universal constant. We have now finally proved that

$$\mathbf{P}\left[\sup_{t\in T}\{X_t - X_{t_0}\} \ge c_1 x \sup_{t\in T} \sum_{k>k_0} \alpha^k \sqrt{\log \ell(A_k(t))}\right] \le c_2 2^{-x^2/8}.$$

for $x \ge 4$. Using $\mathbf{E}[Z] \le \int_0^\infty \mathbf{P}[Z \ge x]\,dx \le 4 + \int_4^\infty \mathbf{P}[Z \ge x]\,dx$ and optimizing over all labelled nets $(\mathcal{A}, \ell)$ completes the proof of the Theorem. □

We now immediately obtain our first nontrivial application of the majorizing measure theorem. The statement of this result is so simple that one would expect that there must be an elementary proof; but no other proof is known.

Corollary 6.29 (Subgaussian comparison theorem). *Let $\{Y_t\}_{t\in T}$ be a separable Gaussian process with natural metric d, and let $\{X_t\}_{t\in T}$ be a separable subgaussian process on (T, d). Then for a universal constant C*

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \le C\,\mathbf{E}\left[\sup_{t\in T} Y_t\right].$$

Proof. Combine Theorems 6.27 and 6.24. □

Remark 6.30. A comparison theorem of this kind can be very useful in practice. In many problems, it is possible to explicitly compute the supremum of a Gaussian process by exploiting special properties of Gaussians (e.g., rotation invariance). One can then invoke Corollary 6.29 to show that the same bound applies when the Gaussian variables are replaced by subgaussian ones, even though one cannot perform explicit computations in the general setting.

While Corollary 6.29 is a trivial consequence of the generic chaining method, most applications require one to work in a nontrivial manner with the chaining bounds. So far we have taken care of the bookkeeping in the chaining argument in terms of labelled nets, as this formulation arose in the most natural manner from the investigation of Gaussian processes. A labelled net is a somewhat unwieldy object, however: not only must one construct increasing partitions, but one must also keep track of labels along the way. We will presently develop an alternative way to organize the generic chaining bounds that dispenses with the need to keep track of the labels.

The basic idea that will be used in the sequel is as follows. In all the chaining arguments that we have used above, we fixed at each scale the diameter of the sets $A \in \mathcal{A}_k$ but allowed an arbitrary number of such sets. An alternative way of organizing the chaining argument is to fix the number of sets in the partition $\mathcal{A}_k$, but to allow their diameters to vary. As a warm-up execise, let us reformulate the simple entropy integral bound from the previous chapter (Corollary 5.25) in this manner. Recall that the covering number $N(T,d,\varepsilon)$ denotes the smallest number of ε-balls needed to cover T. If we define

$$e_n(T) := \inf\{\varepsilon : N(T,d,\varepsilon) < 2^{2^n}\},$$

then the *entropy number* $e_n(T)$ is the smallest radius ε for which one can cover T by less than 2^{2^n} ε-balls (the mysterious 2^{2^n} will be explained shortly). To formulate the chaining bound in terms of entropy numbers, note that

$$\int_0^\infty \sqrt{\log N(T,d,\varepsilon)}\,d\varepsilon = \sum_{n\ge 0}\int_{e_{n+1}(T)}^{e_n(T)} \sqrt{\log N(T,d,\varepsilon)}\,d\varepsilon.$$

Using that $2^{2^n} \le N(T,d,\varepsilon) < 2^{2^{n+1}}$ when $e_{n+1}(T) < \varepsilon < e_n(T)$, we obtain

$$\int_0^\infty \sqrt{\log N(T,d,\varepsilon)}\,d\varepsilon \asymp \sum_{n\ge 0} 2^{n/2}\{e_n(T) - e_{n+1}(T)\} \asymp \sum_{n\ge 0} 2^{n/2} e_n(T).$$

Thus we obtain a bound in terms of entropy numbers that is entirely equivalent, up to the constants, to the entropy integral of Corollary 5.25.

Remark 6.31. Let $\{\beta_n\}$ be an increasing sequence with $\beta_0 = 2$, and define the β-entropy numbers $e_n^\beta = \inf\{\varepsilon : N(T,d,\varepsilon) < \beta_n\}$. Then we can estimate

$$\sum_{n\ge 0}\sqrt{\log\beta_n}\{e_n^\beta - e_{n+1}^\beta\} \le \int_0^\infty \sqrt{\log N(T,d,\varepsilon)}\,d\varepsilon \le \sum_{n\ge 0}\sqrt{\log\beta_{n+1}}\{e_n^\beta - e_{n+1}^\beta\}$$

by arguing as above. In order for the left- and right-hand sides to be comparable, we must have $\log \beta_{n+1} \lesssim \log \beta_n$, which means that $\log \beta_n$ should increase at most exponentially. This explains why we chose $\beta_n = 2^{2^n}$ above (of course, any a^{b^n} for $a, b > 1$ would give equivalent results up to universal constants.)

We now develop a formulation of the generic chaining bound along these lines. The remarkable feature of this formulation is that, somewhat surprisingly, there is no longer a need to keep track of a label for each partition element: the labels are "hidden" in the diameters of the partition elements.

Definition 6.32 (Admissible net). *An increasing sequence of partitions $\mathcal{A} = \{\mathcal{A}_n\}_{n\ge 0}$ of T is called an* admissible net *if $|\mathcal{A}_n| < 2^{2^n}$ for every $n \ge 0$.*

Theorem 6.33 (Labelled and admissible nets). *There exist universal constants c_1, c_2 such that $c_1\gamma'(T) \le \gamma(T) \le c_2\gamma'(T)$. Here we defined*

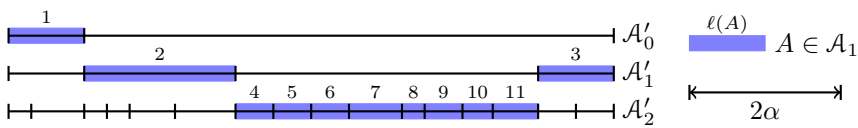
$$\gamma'(T) := \inf_{\mathcal{A}} \sup_{t\in T} \sum_{n\ge 0} 2^{n/2} \operatorname{diam}(A_n(t)),$$

where the infimum is taken over all admissible nets $\mathcal{A}$.

To illustrate the idea of the proof, consider the upper bound $\gamma(T) \lesssim \gamma'(T)$. For any admissible net $\mathcal{A}'$, we must construct an labelled net $(\mathcal{A}, \ell)$ such that

$$\sup_{t\in T} \sum_{k\in\mathbb{Z}} \alpha^k \sqrt{\log \ell(A_k(t))} \lesssim \sup_{t\in T} \sum_{n\ge 0} 2^{n/2} \operatorname{diam}(A'_n(t)).$$

We can view any increasing sequence of partitions as a partition tree with a directed edge from A to B if $B \in c(A)$. A *cut* in the tree is a set of vertices $\mathcal{B}$ such that every branch of the tree contains exactly one element of $\mathcal{B}$. Clearly any cut of a partition tree is itself a partition. The idea of the proof is to define each partition $\mathcal{A}_n$ by taking the smallest possible cut in $\mathcal{A}'$ such that each element of $\mathcal{A}_n$ has diameter at most $2\alpha^n$. Then the above inequality will follow if we assign labels in order of increasing depth of the elements in the original tree $\mathcal{A}'$. This construction is illustrated in the following figure.



Proof (Upper bound). Let $\mathcal{A}'$ be an admissible net, and define

$$n_k(t) = \inf\{n : \operatorname{diam}(A'_n(t)) \le 2\alpha^k\}$$

for every $k \in \mathbb{Z}$ and $t \in T$ (we may assume that $n_k(t) < \infty$ for every k, t, as otherwise the quantity in the definition of $\gamma'(T)$ will be infinite). Let k_0 be the largest integer such that $\operatorname{diam}(T) \le 2\alpha^{k_0}$, and define $\mathcal{A} = \{\mathcal{A}_k\}_{k\in\mathbb{Z}}$ as

$$\mathcal{A}_k = \{T\} \quad \text{for } k \le k_0, \qquad \mathcal{A}_k = \{A'_{n_k(t)}(t) : t \in T\} \quad \text{for } k > k_0.$$

Clearly $\mathcal{A}_k$ defines a cut in $\mathcal{A}'$, and thus $\mathcal{A}$ is an increasing sequence of partitions as in the definition of a labelled net. We now assign labels such that if $A_{k-1}(t) = A_{k-1}(t')$, then $\ell(A_k(t)) > \ell(A_k(t'))$ whenever $n_k(t) > n_k(t')$.

Now note that we can reorganize the sum in the definition of $\gamma'(T)$ as

$$\begin{aligned}\sum_{n\ge 0} 2^{n/2}\operatorname{diam}(A'_n(t)) &= \sum_{k>k_0}\sum_{n_{k-1}(t)\le n<n_k(t)} 2^{n/2}\operatorname{diam}(A'_n(t)) \\ &\ge \sum_{k>k_0} 2\alpha^k \sum_{n_{k-1}(t)\le n<n_k(t)} 2^{n/2} \\ &\ge \sqrt{2}\sum_{k>k_0} \alpha^k 2^{n_k(t)/2}\mathbf{1}_{n_k(t)\neq n_{k-1}(t)}.\end{aligned}$$

We now claim that $2^{n_k(t)/2}\mathbf{1}_{n_k(t)\neq n_{k-1}(t)}\sqrt{\log 2} \ge \sqrt{\log \ell(A_k(t))}$. To see this, note that if $n_k(t) = n_{k-1}(t)$, then $A_k(t)$ is the only child of $A_{k-1}(t)$ and thus $\ell(A_k(t)) = 1$, while we must have $\ell(A_k(t)) \le |\mathcal{A}'_{n_k(t)}| < 2^{2^{n_k(t)}}$ as the labels are sorted by increasing depth in $\mathcal{A}'$. Thus we have shown that for every admissible net $\mathcal{A}'$, there exists a labelled net $(\mathcal{A}, \ell)$ such that

$$\sum_{n\ge 0} 2^{n/2}\operatorname{diam}(A'_n(t)) \ge \sqrt{\frac{2}{\log 2}}\sum_{k\in\mathbb{Z}} \alpha^k \sqrt{\log \ell(A_k(t))}$$

for all $t \in T$. Taking the supremum over t, the infimum over $(\mathcal{A}, \ell)$, and then the infimum over $\mathcal{A}$ yields $\gamma(T) \le c_2\gamma'(T)$ with $c_2 = \sqrt{2^{-1}\log 2}$. □

The proof of the lower bound follows along very similar lines: starting from a labelled net $(\mathcal{A}, \ell)$, we will choose cuts $\mathcal{A}'_n$ such that $|\mathcal{A}'_n| < 2^{2^n}$.

Proof (Lower bound). This time we start with a labelled net $(\mathcal{A}, \ell)$. Let k_0 be the largest integer such that $\mathcal{A}_{k_0} = \{T\}$, and define the quantity

$$\mathbf{u}(A_k(t)) = 4^{k-k_0}\prod_{k_0<m\le k} \ell(A_m(t))^2.$$

Then we have as in the proof of Theorem 6.27 for a universal constant c

$$\sup_{t\in T}\sum_{k\in\mathbb{Z}} \alpha^k\sqrt{\log \ell(A_k(t))} \ge c\sup_{t\in T}\sum_{k>k_0} \alpha^k\sqrt{\log \mathbf{u}(A_k(t))}.$$

We now define a cut in $\mathcal{A}$ by setting

$$k_n(t) = \sup\{k \ge k_0 : \mathbf{u}(A_k(t)) < 2^{2^n}\}.$$

Note that $k_n(t) < \infty$ as $\mathbf{u}(A_k(t))$ increases to infinity (this is the reason why we work with the cumulative labels $\mathbf{u}(A_k(t))$ rather than the labels $\ell(A_k(t))$). Thus we can define the increasing sequence of partitions $\mathcal{A}' = \{\mathcal{A}'_n\}_{n\ge 0}$ as

$$\mathcal{A}'_n = \{A_{k_n(t)}(t) : t \in T\}.$$

As $\mathbf{u}(A_k(t)) \ge 2^{2^n}$ when $k > k_n(t)$, we can estimate

$$\begin{aligned}
\sum_{k>k_0} \alpha^k \sqrt{\log \mathbf{u}(A_k(t))} &= \sum_{n\ge 0} \sum_{k_n(t)<k\le k_{n+1}(t)} \alpha^k \sqrt{\log \mathbf{u}(A_k(t))} \\
&\ge \sqrt{\log 2}\,\frac{\alpha}{1-\alpha} \sum_{n\ge 0} 2^{n/2}\{\alpha^{k_n(t)} - \alpha^{k_{n+1}(t)}\} \\
&\ge \sqrt{\log 2}\left(1-\frac{1}{\sqrt{2}}\right)\frac{\alpha}{1-\alpha} \sum_{n\ge 0} 2^{n/2}\alpha^{k_n(t)} \\
&\ge \frac{\sqrt{\log 2}}{2}\left(1-\frac{1}{\sqrt{2}}\right)\frac{\alpha}{1-\alpha} \sum_{n\ge 0} 2^{n/2}\,\mathrm{diam}(A'_n(t)).
\end{aligned}$$

Thus the only thing that remains to be proved is that $\mathcal{A}'$ is an admissible net. If this is the case, then taking the supremum over t, the infimum over $\mathcal{A}'$, and then the infimum over $\mathcal{A}$ yields the result $c_1\gamma'(T) \le \gamma(T)$.

It therefore remains to show that $|\mathcal{A}'_n| < 2^{2^n}$. To this end, note that by the definition of a labelled net, every partition element $A_k(t) \in \mathcal{A}_k$ gives rise to a distinct sequence of labels $\ell(A_{k_0+1}(t)), \ldots, \ell(A_k(t))$. Thus

$$\begin{aligned}
|\mathcal{A}'_n| &\le \sum_{k>k_0} \sum_{\ell_{k_0+1},\ldots,\ell_k\in\mathbb{N}} \mathbf{1}_{4^{k-k_0}\prod_{k_0<m\le k}\ell_m^2<2^{2^n}} \\
&\le 2^{2^n} \sum_{k\ge 1} 4^{-k} \sum_{\ell_1,\ldots,\ell_k\in\mathbb{N}} \prod_{1\le m\le k} \frac{1}{\ell_m^2} < 2^{2^n},
\end{aligned}$$

as $\sum_k 4^{-k} \sum_{\ell_1,\ldots,\ell_k} \prod_m \frac{1}{\ell_m^2} = \sum_k \left(\frac{\pi^2}{24}\right)^k \approx 0.7$. □

While the formulation in terms of admissible nets is entirely equivalent to the formulation in terms of labelled nets, the former can often be simpler to use in applications as there are no labels to keep track of. To illustrate a nontrivial result that can now readily be obtained, let us prove a remarkable fact about the geometry of Gaussian processes on $\mathbb{R}^n$.

For any subset $T \subseteq \mathbb{R}^n$, let us define the *Gaussian width* $g(T)$ as

$$g(T) := \mathbf{E}\left[\sup_{t\in T} \sum_{i=1}^n g_i t_i\right], \qquad g_1, \ldots, g_n \sim \text{i.i.d. } N(0,1).$$

That is, $g(T)$ is the expected supremum over T of the Gaussian process whose natural distance is the Euclidean distance. We begin with an easy example.

Lemma 6.34. *Let $T = \{t_k : k \ge 2\}$ with $\sup_k \|t_k - s\|\sqrt{\log k} \le a$ for some $s \in \mathbb{R}^n$. Then $g(T) \le Ca$ for a universal constant C.*

Proof. As $X_k = \sum_{i=1}^n g_i\{t_{ki} - s_i\} \sim N(0, \|t_k - s\|^2)$, the union bound gives

$$\mathbf{P}\left[\sup_{k\ge 2} X_k \ge x\right] \le \sum_{k\ge 2} e^{-x^2/2\|t_k-s\|^2} \le \sum_{k\ge 2} k^{-x^2/2a^2}.$$

For $x \ge 2a$, the right-hand side is $\le C' 2^{-x^2/2a^2}$ for a universal constant C'. Thus $g(T) \le 2a + C' \int_{2a}^\infty 2^{-x^2/2a^2}\, dx \le Ca$ for a universal constant C. □

We now make a trivial observation: as the supremum of a *linear* function $L(t)$ over $T \subseteq \mathbb{R}^n$ equals the supremum over the closed convex hull $\overline{\mathrm{conv}}\, T$, we immediately obtain $g(T) = g(\overline{\mathrm{conv}}\, T)$ for any set T. This implies:

Corollary 6.35. *Let $T \subseteq \overline{\mathrm{conv}}\{t_k : k \ge 2\}$ with $\sup_k \|t_k - s\|\sqrt{\log k} \le a$ for some $s \in \mathbb{R}^n$. Then $g(T) \le Ca$ for a universal constant C.*

This easy example gives us a simple geometric principle to control the Gaussian width: if T is contained in the convex hull of a sequence of points $t_k \to s$ that converge at rate $a/\sqrt{\log k}$, then its Gaussian width $g(T)$ is controlled by a. However, this sort of principle appears to be completely arbitrary: we could have started with *any* example in which we can compute explicitly the Gaussian width (for example, ellipsoids or squares) and deduce an analogous geometric principle. The completely unexpected feature of Corollary 6.35, however, is that it admits a sharp converse.

Theorem 6.36. *There is a universal constant K such that whenever $g(T) \le Ka$, there exist $s, \{t_k\}$ with $\sup_k \|t_k - s\|\sqrt{\log k} \le a$ and $T \subseteq \overline{\mathrm{conv}}\{t_k : k \ge 2\}$.*

Combining Corollary 6.35 and Theorem 6.36 immediately yields the following geometric characterization of the Gaussian width:

$$g(T) \asymp \inf\left\{\sup_{k\ge 2} \|t_k - s\|\sqrt{\log k} : T \subseteq \overline{\mathrm{conv}}\{t_k : k \ge 2\}\right\}.$$

This remarkable result appears as a complete mystery at this point. However, much of the mystery is about to disappear: as we will see presently, Theorem 6.36 is little more than a reformulation of the majorizing measure theorem. The key idea is that the points t_k are none other than rescaled versions of the "links" $\pi_n(t) - \pi_{n-1}(t)$ that appear in the chaining argument.

Proof. By the majorizing measure theorem, there is an admissible net $\mathcal{A}$ with

$$\sum_{n\ge 0} 2^{n/2}\,\mathrm{diam}(A_n(t)) \le c\, g(T)$$

for all $t \in T$. Choose for every $A \in \mathcal{A}$ an arbitrary point $t_A \in A$, and define $\pi_n(t) := t_{A_n(t)}$. Fix also an arbitrary point $s \in T$ and let $\pi_{-1}(t) := s$. Define

$$\beta_n(t) = \frac{2^{n/2}\|\pi_n(t) - \pi_{n-1}(t)\|}{Cg(T)}, \qquad x_n(t) = \frac{Cg(T)}{2^{n/2}} \frac{\pi_n(t) - \pi_{n-1}(t)}{\|\pi_n(t) - \pi_{n-1}(t)\|}$$

for $n \ge 0$. As $\|\pi_n(t) - t\| \le \mathrm{diam}(A_n(t)) \to 0$, we have

$$t = s + \sum_{n\ge 0}\{\pi_n(t) - \pi_{n-1}(t)\} = s + \sum_{n\ge 0}\beta_n(t)x_n(t).$$

As $g(T) \ge \mathbf{E}[\langle g, t\rangle \vee \langle g, s\rangle] = \mathbf{E}|\langle g, \frac{t-s}{2}\rangle| = \|t-s\|/\sqrt{2\pi}$ for all $t \in T$,

$$\sum_{n\ge 0}\beta_n(t) \le \frac{\sqrt{2\pi}}{C} + \frac{1}{Cg(T)}\sum_{n\ge 1} 2^{n/2}\,\mathrm{diam}(A_{n-1}(t)) \le 1$$

if we choose $C = \sqrt{2\pi} + c\sqrt{2}$. Thus

$$T \subseteq \overline{\mathrm{conv}}\{s + x_n(t) : n \ge 0, t \in T\} =: \overline{\mathrm{conv}}\{z_k : k \ge 1\},$$

where z_k have been sorted such that $\|z_k - s\|$ is nonincreasing.

Now note that $\|x_n(t)\| = Cg(T)2^{-n/2}$, while there are at most $|\mathcal{A}_n||\mathcal{A}_{n-1}|$ such terms. We can therefore readily estimate

$$\max\{k : \|z_k - s\| > Cg(T)2^{-n/2}\} \le \sum_{k=0}^{n-1} 2^{2^k}2^{2^{k-1}} \le n2^{2^n} \le 2^{2^{n+1}}.$$

Thus we have for all $n \ge 0$ and $2^{2^{n+1}} < k \le 2^{2^{n+2}}$

$$\|z_k - s\| \le \frac{Cg(T)}{2^{n/2}} = \frac{2\sqrt{\log 2}\,Cg(T)}{\sqrt{\log 2^{2^{n+2}}}} \le \frac{2\sqrt{\log 2}\,Cg(T)}{\sqrt{\log k}}.$$

Setting $t_{k+1} = z_k$ so that $T \subseteq \overline{\mathrm{conv}}\{t_k : k \ge 2\}$, we can readily choose K such that $g(T) \le Ka$ implies $\|t_k - s\| \le a/\sqrt{\log k}$ for all $k \ge 2$. □

We have seen above several different but closely related formulations of the generic chaining bound: in terms of labelled nets (Theorems 6.24 and 6.27), in terms of majorizing measures (Problem 6.7), and in terms of admissible nets (Theorem 6.33). We conclude this section by developing a *dual* formulation of the generic chaining bound. Beside that this very useful result is of significant interest in its own right, we will isolate along the way a fundamental idea that underlies many applications of the generic chaining machinery.

Let us begin by motivating why we develop yet another formulation of the generic chaining. The definition of $\gamma(T)$ involves an infimum over labelled nets: this means that in order to obtain an upper bound on the supremum of a given Gaussian process, we only need to exhibit one particular labelled net for which the quantity in the definition of $\gamma(T)$ is small. In essence, this is what we have been doing in the previous chapter: it is easy to construct labelled nets by piecing together ε-nets at different scales, in which case we

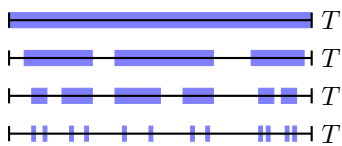
recover the entropy integral of Corollary 5.25 (cf. Problem 6.5). However, to have a sharp understanding of the supremum of a given Gaussian process, we must also obtain a matching lower bound. It is very difficult to obtain lower bounds on $\gamma(T)$, as this would require us to argue that the quantity in the definition of $\gamma(T)$ is large for *every* possible choice of the labelled net.

One should think of a labelled or admissible net, which defines a covering of the set T at many different scales, as a multiscale counterpart to the notion of an ε-net, which defines a covering of the set T at a single scale ε. From this viewpoint, the majorizing measure theorem states that the expected supremum of a Gaussian process over T is equivalent up to universal constants to the *smallest* "size" (in the $\gamma(T)$-sense) of a multiscale covering of T. The classical duality between packing and covering now suggests an interesting idea: is there a corresponding multiscale counterpart to the notion of an ε-packing, so that the supremum of a Gaussian process is equivalent up to the *largest* "size" of a multiscale packing? This is precisely the idea that will be developed in the remainder of this section. Such a dual formulation is precisely what one needs in order to obtain lower bounds on the supremum of a Gaussian process.

It is not difficult to find a good candidate for the notion of multiscale packing. Recall that there was no mystery in the definition of a labelled net: this notion was simply designed to obtain the best possible upper bound on the supremum of a Gaussian process using the super-chaining principle (Proposition 6.21). To obtain a notion of multiscale packing, we apply precisely the same idea in the opposite direction: we design an object that yields the best possible lower bound using the super-Sudakov inequality (Theorem 6.14). To help us with the bookkeeping, let us introduce some useful structures.

Definition 6.37 (Trees). *A T-*tree *is a family $\mathcal{T}$ of nonempty subsets of T such that $T \in \mathcal{T}$, and for all $C, C' \in \mathcal{T}$ either $C \cap C' = \varnothing$, $C \subseteq C'$, or $C' \subseteq C$.*

The definition of a tree is illustrated in the following figure (the base set T is duplicated several times to clarify the positions of the elements of $\mathcal{T}$):



It is not difficult to see that a T-tree can be thought of as a directed tree in the graph-theoretic sense. The root of the tree is T, and the children of a node $c(A)$ and the leaves of the tree $l(\mathcal{T})$ are defined by inclusion in the obvious fashion. For every leaf $A \in l(\mathcal{T})$, we will denote the corresponding branch of the tree as $A_0 \subseteq A_1 \subseteq \ldots$ (starting at the root $A_0 = T$).

An increasing sequence of partitions, such as in the definition of a labelled net, naturally defines a T-tree with the additional property that its leaves cover T. In contrast, in a multiscale notion of packing, we would like the

children of each node in the tree to be well separated. The following notion is specifically designed in order to apply Theorem 6.14.

Definition 6.38 (Packing tree). *A* packing tree $(\mathcal{T}, \varkappa)$ *is a* T*-tree* $\mathcal{T}$ *together with a map* $\varkappa : \mathcal{T} \to \mathbb{Z}$ *such that the following holds for every* $A \notin l(\mathcal{T})$*:*

1. For every $C \in c(A)$*, there exists* $t_C \in T$ *such that* $C \subseteq B(t_C, \alpha^{\varkappa(A)+1})$*.*
2. $d(t_C, t_{C'}) \geq \alpha^{\varkappa(A)}$ *for all* $C, C' \in c(A)$*,* $C \neq C'$*.*

We can now state a dual form of the majorizing measure theorem (where we note that the upper bound holds already when $\{X_t\}_{t\in T}$ is subgaussian.)

Theorem 6.39 (Dual majorizing measure theorem). *Let* $\{X_t\}_{t\in T}$ *be a separable Gaussian process. Then we have for universal constants* c_1, c_2

$$c_1\gamma''(T) \leq \mathbf{E}\left[\sup_{t\in T} X_t\right] \leq c_2\gamma''(T).$$

Here we defined

$$\gamma''(T) = \sup_{(\mathcal{T},\varkappa)} \inf_{A\in l(\mathcal{T})} \sum_{n\geq 0} \alpha^{\varkappa(A_n)} \sqrt{\log |c(A_n)|},$$

where the supremum is taken over all packing trees $(\mathcal{T}, \varkappa)$*.*

While we only formally defined the notion of a packing tree here, this is not the first time that we have enocountered this idea: we essentially constructed a packing tree in the proof of Theorem 6.19. The special feature of the stationary case is that the packing tree is *regular*, so that $\gamma''(T)$ can be expressed in terms of the packing numbers of T. Then the equivalence between $\gamma''(T)$ and the entropy integral follows from the simple duality between packing and covering numbers each scale. Theorem 6.39 could be viewed as a generalization of this idea to the nonstationary setting. This result lies much deeper, however, as we must now run the duality argument in a multiscale fashion.

We now turn to the proof of Theorem 6.39. The lower bound is easy: it follows almost trivially, by design, from iterating the super-Sudakov inequality.

Proof (Lower bound). Given a packing tree $(\mathcal{T}, \varkappa)$, we obtain

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \geq c \inf_{A\in l(\mathcal{T})} \sum_{n\geq 0} \alpha^{\varkappa(A_n)} \sqrt{\log |c(A_n)|}$$

by repeatedly applying Theorem 6.14 starting from the root of the tree (we do not need to worry about the remainder term at the end of the chaining argument as this is a lower bound). Now take the supremum over $(\mathcal{T}, \varkappa)$. □

The interesting part of the proof is the upper bound $\gamma(T) \lesssim \gamma''(T)$. It turns out that we already did almost all the necessary work in the proof of the lower bound in Theorem 6.24, but this is not at all obvious at the moment. Let us therefore first give an abstract statement of what we accomplished there.

Definition 6.40 (Growth functional). *A map $F : 2^T \to \mathbb{R}_+$ is called a* growth functional *if $F(B) \le F(A)$ whenever $B \subseteq A \subseteq T$, and*

$$F(A) \ge c\alpha^n \sqrt{\log |N|} + \min_{s \in N} F(A \cap B(s, \alpha^{n+1}))$$

whenever N is an α^n-packing of $A \subseteq T$ for some $n \in \mathbb{Z}$.

Theorem 6.41 (Partitioning scheme). *$\gamma(T) \le KF(T)$ for any growth functional F (the constant K depends only on c, α in the definition of F).*

Proof. Repeat *verbatim* the proof of the lower bound of Theorem 6.24, replacing the special growth functional $G(A)$ by $F(A)$ throughout. □

The key insight behind Theorem 6.41 is that in order to upper bound $\gamma(T)$ in the proof of the majorizing measure theorem, the only Gaussian property that we used was the super-Sudakov inequality. Thus we can use the same proof to bound $\gamma(T)$ by any other object that satisfies the super-Sudakov inequality. Theorem 6.41 turns out to be perhaps the most important tool in applications of the majorizing measure theorem: while it is exceedingly difficult to construct good labelled nets (or even admissible nets) by hand in any given situation, it is often much more promising to try to guess the form of a growth functional that captures the geometry of the problem. Thus Theorem 6.41 provides a powerful tool to obtain upper bounds on $\gamma(T)$ in different problems (supposing, of course, that the easiest entropy integral bounds from the previous chapter do not suffice). We presently give a simple illustration of this idea by completing the proof of the upper bound in Theorem 6.39.

Proof (Upper bound). It suffices by Theorem 6.27 to show that $\gamma(T) \le K\gamma''(T)$ for a universal constant K. To this end, we will show that γ'' is itself a growth functional, so that the proof is complete by Theorem 6.41.

Fix a set $S \subseteq T$ and an α^n-packing N of S. Let $\varepsilon > 0$, and choose for every $s \in N$ a packing tree $(\mathcal{T}_s, \varkappa_s)$ of $S \cap B(s, \alpha^{n+1})$ such that

$$\inf_{A \in l(\mathcal{T}_s)} \sum_{n \ge 0} \alpha^{\varkappa_s(A_n)} \sqrt{\log |c(A_n)|} \ge \min_{s \in N} \gamma''(S \cap B(s, \alpha^{n+1})) - \varepsilon.$$

Now define a new tree $\mathcal{T} = \{S\} \cup \bigcup_{s \in N} \mathcal{T}_s$, and assign labels $\varkappa(A) = \varkappa_s(A)$ for $A \in \mathcal{T}_s$ and $\varkappa(S) = n$. Then clearly $(\mathcal{T}, \varkappa)$ is a packing tree of S and

$$\begin{aligned}\gamma''(S) &\ge \inf_{A \in l(\mathcal{T})} \sum_{n \ge 0} \alpha^{\varkappa(A_n)} \sqrt{\log |c(A_n)|} \\ &\ge \alpha^n \sqrt{\log |N|} + \min_{s \in N} \gamma''(S \cap B(s, \alpha^{n+1})) - \varepsilon.\end{aligned}$$

Letting $\varepsilon \downarrow 0$ shows that γ'' satisfies the super-Sudakov inequality. As γ'' is clearly increasing $\gamma''(A) \le \gamma''(B)$ for $A \subseteq B$, it is a growth functional. □

Problems

6.8 (Chaining with admissible nets). The formulation of the majorizing measure theorem in terms of admissible nets seems to be somewhat simpler than the formulation in terms of labelled nets, as there are no labels to keep track of. In fact, from the point of view of the *upper* bounds, chaining with admissible nets is even easier than using labelled nets.

a. Give a direct proof, along the lines of Theorem 6.27, of the fact that if $\{X_t\}_{t\in T}$ is a separable subgaussian process on (T,d) then

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \lesssim \gamma'(T).$$

It is in fact also possible to give a direct proof of the lower bound in the majorizing measure theorem in terms of admissible nets. However, this approach is less intuitive than the proof in terms of labelled nets, as we lose the natural symmetry between the upper and lower bounds in the chaining argument.

Let us now consider a less structured variant of the notion of an admissible net. Call $\mathcal{A} = \{\mathcal{A}_n\}_{n\geq 0}$ an *admissible family* if each $\mathcal{A}_n$ individually is a partition of T with $|\mathcal{A}_n| < 2^{2^n}$, but where we do not make the assumption that the sequence of partitions is increasing. Define

$$\gamma_0'(T) := \inf_{\mathcal{A}} \sup_{t\in T} \sum_{n\geq 0} 2^{n/2} \operatorname{diam}(A_n(t)),$$

with the infimum taken over all admissible families $\mathcal{A}$.

b. Show that $\gamma_0'(T) \leq \gamma'(T) \leq K\gamma_0'(T)$ for a universal constant K.

Hint: given an admissible family $\mathcal{A}$, define an increasing sequence of partitions $\mathcal{B}$ by letting $\mathcal{B}_n$ be the partition generated by $\mathcal{A}_0, \ldots, \mathcal{A}_n$.

c. Give a direct proof of the upper bound in terms of entropy numbers

$$\gamma_0'(T) \leq \sum_{n\geq 0} 2^{n/2} e_n(T)$$

that is equivalent to the simple chaining bound in the previous chapter.

6.9 (Separated trees). A *separated tree* $(\mathcal{T},\varkappa)$ is a T-tree $\mathcal{T}$ together with a map $\varkappa : \mathcal{T} \to \mathbb{Z}$ such that for every $A \notin l(\mathcal{T})$, we have $d(C,C') \geq \alpha^{\varkappa(A)}$ and $\varkappa(C) > \varkappa(A)$ for all $C, C' \in c(A)$, $C \neq C'$. Thus a separated tree is a less structured variant of a packing tree where we have no control of the diameters of the elements of a separated tree. Nonetheless, we will see that the quantity

$$\gamma_0''(T) = \sup_{(\mathcal{T},\varkappa)} \inf_{A\in l(\mathcal{T})} \sum_{n\geq 0} \alpha^{\varkappa(A_n)} \sqrt{\log |c(A_n)|},$$

where the supremum is taken here over all separated trees $(\mathcal{T},\varkappa)$, plays an equivalent role to the quantity $\gamma''(T)$. This is not at all obvious, as we cannot apply the super-Sudakov inequality without control on the diameter.

a. Show that $\gamma''(T) \lesssim \gamma_0''(T)$.

b. Show that $\gamma_0''(T) \lesssim \gamma(T)$.

Hint: fix a separated tree $(\mathcal{T}, \varkappa)$ and labelled net $(\mathcal{A}, \ell)$. Now argue as follows starting from the root B_0 of $\mathcal{T}$: as the children $c(B_0)$ are $\alpha^{\varkappa(B_0)}$-separated, each element of $\mathcal{A}_{\varkappa(B_0)+1}$ can intersect at most one element of $c(B_0)$. Thus we can choose $B_1 \in c(B_0)$ and $A_1 \in \mathcal{A}_{\varkappa(B_0)+1}$ with $\ell(A_1) \geq |c(B_0)|$. Now iterate this procedure to select a full branch $B_0, B_1, \ldots$ of $\mathcal{T}$ and a sequence $A_{i+1} \in \mathcal{A}_{\varkappa(B_i)+1}$ with $\ell(A_{i+1}) \geq |c(B_i)|$. Finally, compare the sums that appear in the definitions of $\gamma_0''(T)$ and $\gamma(T)$ for this selection.

c. Conclude that $\gamma''(T) \asymp \gamma_0''(T)$.

6.10 (Ultrametrics). A (finite) *ultrametric space* (U, d) is a (finite) set U together with a metric d on U that satisfies the ultra-triangle inequality

$$d(u, v) \leq \max\{d(u, w), d(v, w)\} \quad \text{for all } u, v, w \in X.$$

Ultrametric spaces play an important role in the geometry of metric spaces, where they play a role analogous to that of Hilbert spaces in functional analysis (any finite ultrametric space can be isometrically embedded in ℓ_2; the proof of this fact is left to the interested reader). They also arise naturally in statistical physics, computer science, and computational biology.

a. Let U be a finite set and $\mathcal{T}$ be a U-tree whose leaves are the singletons $\{u\}$. Fix $\delta : \mathcal{T} \to \mathbb{R}_+$ so that $\delta(\{u\}) = 0$ and $\delta(C) < \delta(A)$ if $C \in c(A)$, and let

$$d(u, v) = \delta(A(u, v)), \qquad A(u, v) = \bigcap\{C \in \mathcal{T} : C \supseteq \{u, v\}\}.$$

Show that (U, d) is an ultrametric space.

b. Let (U, d) be a finite ultrametric space. Show that there is a tree $\mathcal{T}$ and assignment $\delta : \mathcal{T} \to \mathbb{R}_+$ as in part a. such that $d(u, v) = \delta(A(u, v))$.

Hint: show that if (U, d) is ultrametric, then balls $B(u, \varepsilon)$ and $B(v, \varepsilon)$ that do not coincide must be disjoint. Thus $\{B(u, \varepsilon) : u \in U\}$ is a partition.

A finite metric space (U, d) *K-embeds in an ultrametric space* if there is an ultrametric d_{u} on U such that $K^{-1} d_{\mathrm{u}}(u, v) \leq d(u, v) \leq K d_{\mathrm{u}}(u, v)$ for all $u, v \in U$. This idea proves to be intimately related to Gaussian processes.

c. Prove the following formulation of the majorizing measure theorem: there is a universal constant K so that for any separable Gaussian process $\{X_t\}_{t \in T}$, there is a finite subset $U \subseteq T$ that K-embeds in an ultrametric space with

$$\mathbf{E}\left[\sup_{u \in U} X_u\right] \leq \mathbf{E}\left[\sup_{t \in T} X_t\right] \leq K\, \mathbf{E}\left[\sup_{u \in U} X_u\right].$$

Hint: consider a more structured notion of packing tree with the additional requirement that each $A \in \mathcal{T}$ has diameter $\lesssim \alpha^{\varkappa(A)}$. Use a minor modification of Theorem 6.41 to show that Theorem 6.39 still holds for the modified packing tree. Finally, use the packing tree to define a suitable ultrametric.

Notes

§6.1. The inequalities of Slepian-Fernique and Sudakov are classical results on Gaussian processes. The approach starting from Gaussian interpolation (Lemma 6.9) is due to Slepian [73]. We follow Chatterjee in using the convenient approximation of the maximum in the proof of Theorem 6.5 (see [2]). See [95] for more on applications to random matrices (Problems 6.11 and 6.12). The convex geometry proof of Problem 6.13, due to Talagrand [51], makes it possible to extend Sudakov's inequality to non-Gaussian processes [79].

§6.2. The super-Sudakov inequality is due to Talagrand [77]. The alternative proof of Problem 6.1 is taken from [53]. Theorem 6.19 is due to Fernique [35]. As is noted in [49], the super-Sudakov inequality makes it possible to give a particularly transparent proof that is almost entirely analogous to that of the chaining upper bound. Problem 6.4 is inspired by [81].

§6.3. Talagrand's majorizing measure theorem is considered to be notoriously difficult, perhaps because the complicated chaining object that arises here looks so bizarre. I have tried to tell the story in such a way that the result does not appear as a major miracle, but rather as the natural consequence of basic properties of Gaussian variables. In particular, it seems that the symmetry between Corollary 6.20 and Proposition 6.21 is the central idea in the proof; once this has been understood, it should be almost clear why the result must be true. The proof given here and the formulation in terms of labelled nets is the one developed in [77, 78]; the presentation is inspired by [49, 43] (I learned the proof from [49]). Proposition 6.21 appears in [81].

The original proof of the majorizing measure theorem [75] was very complicated, as everything was formulated directly in terms of continuous majorizing measures which are not well suited to chaining. A good exposition of it can be found in [1]. The most recent formulation in terms of admissible nets (section 6.4) is often simpler to use, but a direct proof of the majorizing measure theorem along these lines [88, 89] is in my opinion more mysterious as the natural symmetry between the upper and lower bounds is lost.

The (continuous) upper bound in the majorizing measure theorem as formulated in Problem 6.7 is much older and is due to Fernique [35]. It can even be developed pathwise as a real analysis lemma, see [8].

§6.4. The proof of Theorem 6.27 is based on [83], while the proof of Theorem 6.33 is inspired by [86]. The remaining results in this chapter are taken from [88, 89], where an exhaustive treatment of the generic chaining method and its applications is given (using exclusively the admissible net formulation). A remarkable application of the connection with separated trees can be found in [23]. The formulation in terms of ultrametric spaces (Problem 6.10) is implicit in [75]; see [59] for further developments in this direction.

7

Empirical processes and combinatorics

In the previous chapter, we have developed a detailed understanding of the supremum of a Gaussian process $\{X_t\}_{t\in T}$ by chaining with respect to the natural metric $d(t,s) = \|X_t - X_s\|_2$. While Gaussian processes are important in their own right, in many applications such processes arise only in an indirect manner. Particularly in areas such as statistics and machine learning, the more fundamental object of interest is the *empirical process* $\{G_n(f)\}_{f\in\mathcal{F}}$ over a class of functions $\mathcal{F}$, defined in terms of an i.i.d. sequence $X_1, X_2, \ldots \sim \mu$ as

$$G_n(f) := \sqrt{n}\{\mu_n f - \mu f\}, \qquad \mu_n = \frac{1}{n}\sum_{k=1}^{n} f(X_k).$$

Understanding the supremum of the empirical process determines the rate of convergence of the law or large numbers uniformly over a class of functions $\mathcal{F}$, and thereby the performance of many types of statistical estimators. Similar problems arise at a fundamental level in the geometry of Banach spaces, in combinatorial set theory, and in many other applications.

That empirical processes are closely related to Gaussian processes is expressed by the following immediate consequence of the multivariate CLT.

Lemma 7.1 (Central limit theorem). *For any $f_1, \ldots, f_k \in \mathcal{F}$, we have*

$$(G_n(f_1), \ldots, G_n(f_k)) \Longrightarrow (Z(f_1), \ldots, Z(f_k)) \text{ in distribution} \quad \text{as } n \to \infty,$$

where $\{Z(f)\}_{f\in\mathcal{F}}$ is the Gaussian process with $\mathrm{Cov}[Z(f), Z(g)] = \mathrm{Cov}_\mu[f,g]$.

In view of the central limit theorem, we expect that the empirical process $\{G_n(f)\}_{f\in\mathcal{F}}$ should in some sense behave like the Gaussian process $\{Z(f)\}_{f\in\mathcal{F}}$ when n is sufficiently large. In particular, as the natural metric for the Gaussian process is given by $d(f,g) = \mathrm{Var}_\mu[f-g]^{1/2}$, we might hope to control the supremum of the empirical process by chaining with respect to d. Of course, the empirical process is not actually Gaussian for finite n, but the Azuma-Hoeffding inequality (Lemma 3.6) ensures that the empirical process

is subgaussian with respect to the metric $d_\infty(f,g) = \|f-g\|_\infty$. We can therefore directly control the supremum of the empirical process by chaining with respect to the uniform metric (indeed, we have already seen this approach in action in Example 5.28!) The problem with this approach is that the uniform metric d_∞ can be *much* larger than the $L^2(\mu)$-metric d in many cases, so that we can incur an enormous loss of efficiency in controlling the empirical process as compared to the limiting Gaussian process. Let us give a simple illustration of a setting where this issue arises in a dramatic fashion.

Example 7.2. Let $X_1, X_2, \ldots$ be an i.i.d. sequence of real-valued random variables with distribution μ. By the law of large numbers, the empirical distribution function $F_n(x) = \mu_n(]-\infty, x])$ converges a.s. to the distribution function $F(x) = \mu(]-\infty, x])$ for every $x \in \mathbb{R}$. However, Glivenko and Cantelli proved already in 1933 that the convergence is even *uniform* in x:

$$\|F_n - F\|_\infty \xrightarrow{n\to\infty} 0 \quad \text{a.s.}$$

To understand this phenomenon (as well as the rate of convergence at which this happens), we must understand the supremum of the empirical process

$$\sup_{f\in\mathcal{F}} |\mu_n f - \mu f| = \frac{1}{\sqrt{n}} \sup_{f\in\mathcal{F}} |G_n(f)|$$

over the class of indicators $\mathcal{F} = \{\mathbf{1}_{]-\infty,x]} : x \in \mathbb{R}\}$. Now note that

$$\|\mathbf{1}_{]-\infty,x]} - \mathbf{1}_{]-\infty,x']}\|_\infty = 1 \quad \text{whenever } x \neq x'.$$

Thus evidently $N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) = \infty$ for every $\varepsilon < 1$! In particular, we see that no chaining argument with respect to the uniform metric can ever capture the uniform convergence of the empirical process over the class $\mathcal{F}$, or for that matter over any other infinite class of (indicators of) sets. On the other hand, it is not difficult to see that the covering numbers $N(\mathcal{F}, \|\cdot\|_{L^2(\mu)}, \varepsilon)$ are small, and thus the Gaussian process $\{Z(f)\}_{f\in\mathcal{F}}$ is easily controlled by chaining.

It should be evident from the above discussion that a direct application of the methods that we developed so far to control the suprema of random processes fails to capture the behavior of empirical processes. In order to obtain better control of empirical processes, we must understand in what sense the behavior of such processes is similar to that of the Gaussian limit. In this chapter, we will develop methods to "bring out the Gaussian nature" of empirical processes and to control the resulting inequalities.

7.1 The symmetrization method

One of the most fundamental approaches to bringing out the Gaussian nature of empirical processes is through the method of *symmetrization*. To understand this idea behind this method, let us begin with a (very) informal discussion of "why the central limit theorem works."

Let us fix a bounded function f, and consider the sum $\sum_{k=1}^n \{f(X_k)-\mu f\}$. As this sum contains n terms of order 1 each, this quantity could be as large as $\sim n$ in the worst case. However, the typical situation is quite different: the central limit theorem states that the sum is only of order $\sqrt{n}$ in probability! Of course, the reason for this is clear. In order for the sum to be of order n, most of the terms in the sum must have the same sign so that their contributions add up. But as the terms in the sum are independent and centered, they are highly unlikely to all be of the same sign; to the contrary, there will typically be many terms of opposite sign, so that most of the terms in the sum cancel rather than adding up. This cancellation between terms of different sign accounts for the major reduction in scale from $O(n)$ to only $O(\sqrt{n})$.

The cancellation of terms of different signs proves to be the key mechanism of the central limit theorem: it is the aggregate effect of random signs that leads to Gaussian behavior. The remaining features of the distribution μ are only relevant to the limiting behavior to the extent that they determine the scale of the Gaussian (i.e., its variance). This suggests that in order to bring out the Gaussian nature of the empirical process, we should somehow isolate the random signs in such a manner that we can apply the machinery developed in the previous chapters only to the "Gaussian part" of the empirical process. The method of symmetrization achieves precisely this aim.

Lemma 7.3 (Symmetrization). *Let $X_1,\ldots,X_n$ be i.i.d. random variables in $\mathbb{X}$ with distribution μ, and let $\mathcal{F}$ be a class of functions on $\mathbb{X}$. Then*

$$\begin{aligned}\mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^n\{f(X_k)-\mu f\}\right|\right] &\le \mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^n\varepsilon_k\{f(X_k)-f(Y_k)\}\right|\right]\\ &\le 2\,\mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^n\{f(X_k)-\mu f\}\right|\right],\end{aligned}$$

where $Y_1,\ldots,Y_n$ is an independent copy of $X_1,\ldots,X_n$, and $\varepsilon_1,\ldots,\varepsilon_n$ are i.i.d. symmetric Bernoulli random variables independent of X,Y.

Proof. As $\mu f = \mathbf{E}[f(Y_k)|X_1,\ldots,X_n]$, Jensen's inequality yields

$$\mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^n\{f(X_k)-\mu f\}\right|\right] \le \mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^n\{f(X_k)-f(Y_k)\}\right|\right].$$

But note that $f(X_k)-f(Y_k)$, being a symmetric random variable (hence the name *symmetrization*!), has the same law as $\varepsilon_k\{f(X_k)-f(Y_k)\}$. This implies

$$\mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^n\{f(X_k)-f(Y_k)\}\right|\right] = \mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^n\varepsilon_k\{f(X_k)-f(Y_k)\}\right|\right],$$

which proves the first inequality. The second inequality follows readily using $f(X_k)-f(Y_k) = f(X_k)-\mu f-\{f(Y_k)-\mu f\}$ and the triangle inequality. □

Let us consider what we have achieved. Define the process

$$Z_n(f) = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \varepsilon_k \{f(X_k) - f(Y_k)\}.$$

At first sight, this process seems no more useful than the empirical process $G_n(f)$: the best we can do is still to apply the Azuma-Hoeffding inequality, which shows that $\{Z_n(f)\}_{f\in\mathcal{F}}$ is subgaussian with respect to the uniform norm. However, this is not the right way to bound the supremum of $Z_n(f)$. What we have accomplished here is to isolate the behavior of the signs: the random signs ε_k are independent of the remaining randomness in the problem. We should therefore apply our machinery *conditionally* on X, Y, so that only the "Gaussian part" of the process remains. If we apply the Azuma-Hoeffding inequality conditionally on X, Y, we find that the process $\{Z_n(f)\}_{f\in\mathcal{F}}$ is subgaussian with respect to the *random* metric d_n on $\mathcal{F}$ defined by

$$d_n(f,g) := \left[\frac{1}{n}\sum_{k=1}^{n}\{f(X_k) - g(X_k) - f(Y_k) + g(Y_k)\}^2\right]^{1/2}.$$

To interpret this metric, note that by the law of large numbers

$$\lim_{n\to\infty} d_n(f,g) = \mathbf{E}[\{f(X_1) - g(X_1) - f(Y_1) + g(Y_1)\}^2]^{1/2} = 2\,\mathrm{Var}_\mu[f-g]^{1/2},$$

which is none other (up to a constant factor) than the natural metric d for the limiting Gaussian process $Z(f)$! Thus the symmetrization method isolates precisely in what sense the empirical process approximates the Gaussian process $Z(f)$: by Lemma 7.3, controlling the supremum of the empirical process $G_n(f)$ is equivalent to controlling the supremum of a process that is subgaussian for an empirical approximation to the natural metric of $Z(f)$.

Once the symmetrization argument has been understood, we can apply all the machinery developed in the previous chapters conditionally on X, Y. For example, applying Corollary 5.25 conditionally yields

$$\mathbf{E}\left[\sup_{f\in\mathcal{F}} |G_n(f)|\right] \lesssim \mathbf{E}\left[\int_0^\infty \sqrt{\log N(\mathcal{F}, d_n, \varepsilon)}\, d\varepsilon\right].$$

This is a vast improvement over the analogous bound with $N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)$ that would be obtained by a direct application of Azuma-Hoeffding to $G_n(f)$. At the same time, the fact that we have to deal with a random metric d_n is a nontrivial complication: to control the covering numbers $N(\mathcal{F}, d_n, \varepsilon)$ we must understand the *random geometry* of the metric space $(\mathcal{F}, d_n)$. In the following sections we will develop some tools to deal with this problem.

So far there has been no loss in our estimates except universal constants: Lemma 7.3 has matching upper and lower bounds. In many applications of symmetrization, however, the following bounds prove to be convenient.

Lemma 7.4 (Symmetrization II). *Let $X_1, \ldots, X_n$ be i.i.d. random variables in $\mathbb{X}$ with distribution μ, and let $\mathcal{F}$ be a class of functions on $\mathbb{X}$. Then*

$$\mathbf{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^n\{f(X_k)-\mu f\}\right] \le 2\,\mathbf{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^n \varepsilon_k f(X_k)\right] \le \sqrt{2\pi}\,\mathbf{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^n g_k f(X_k)\right],$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. symmetric Bernoulli random variables and $g_1, \ldots, g_n$ are i.i.d. $N(0,1)$ random variables independent of X.

Remark 7.5. The symmetrization method has its origin in functional analysis, where symmetric Bernoulli random variables are often referred to as Rademacher variables. Thus the first inequality is called Rademacher symmetrization, while the second inequality is called Gaussian symmetrization.

Proof. It follows exactly as in the proof of Lemma 7.3 that

$$\mathbf{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^n\{f(X_k)-\mu f\}\right] \le \mathbf{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^n \varepsilon_k\{f(X_k)-f(Y_k)\}\right].$$

Splitting the supremum yields

$$\mathbf{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^n\{f(X_k)-\mu f\}\right] \le \mathbf{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^n \varepsilon_k f(X_k)\right] + \mathbf{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^n (-\varepsilon_k) f(Y_k)\right].$$

As (ε_k, X_k) has the same distribution as $(-\varepsilon_k, Y_k)$, the first inequality follows. For the second inequality, as $\mathbf{E}[|g_k||\varepsilon_1, \ldots, \varepsilon_n, X_1, \ldots, X_n] = \sqrt{2/\pi}$, we have

$$\sqrt{\frac{2}{\pi}}\,\mathbf{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^n \varepsilon_k f(X_k)\right] = \mathbf{E}\left[\sup_{f\in\mathcal{F}}\mathbf{E}\left[\sum_{k=1}^n \varepsilon_k|g_k|f(X_k)\,\middle|\,\varepsilon, X\right]\right] \le \mathbf{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^n \varepsilon_k|g_k|f(X_k)\right].$$

But as g_k is symmetric, $\varepsilon_k|g_k|$ has the same law as g_k, and we are done. □

Lemma 7.4 has two advantages. First, the natural random metric associated with the symmetrized process is the $L^2(\mu_n)$-metric

$$\|f-g\|_{L^2(\mu_n)} = \left[\frac{1}{n}\sum_{k=1}^n\{f(X_k)-g(X_k)\}^2\right]^{1/2},$$

which is often easier to control than the metric d_n defined above (while the latter is more precise, in most applications Lemma 7.4 suffices). Second, here

we see that we can even control the supremum of the empirical process by the supremum of a true Gaussian process (conditionally on X), not just by a subgaussian process. This is conceptually pleasing, but does not make any difference in most applications: upper bounds using chaining work just as well for Gaussian processes as for subgaussian processes. We have used the Gaussian property much more heavily in deriving lower bounds; however, the Gaussian symmetrization is not necessarily sharp, so that we cannot derive lower bounds in this manner without further work (see, however, Problems 7.1 and 7.2 below for situations where one can implement this idea).

We conclude this section by noting that we can use symmetrization not only to bound the expectated supremum of the empirical process, but also its tail probabilities. The following simple tool provides one way to do this.

Lemma 7.6 (Panchenko). *Let X, Y be random variables such that*

$$\mathbf{E}[\Phi(X)] \le \mathbf{E}[\Phi(Y)]$$

for every increasing convex function Φ. If

$$\mathbf{P}[Y \ge t] \le c_1 e^{-c_2 t^\alpha} \quad \text{for all } t \ge 0$$

for some $c_1, \alpha \ge 1$ and $c_2 > 0$, then

$$\mathbf{P}[X \ge t] \le c_1 e^{1-c_2 t^\alpha} \quad \text{for all } t \ge 0.$$

Proof. As $x \mapsto \Phi(x_+^\alpha)$ is increasing and convex for every $\alpha \ge 1$, it suffices to consider the case $\alpha = 1$. Applying the assumption to $\Phi(x) = (x-t)_+$ yields

$$\int_t^\infty \mathbf{P}[X \ge s]\, ds \le \int_t^\infty \mathbf{P}[Y \ge s]\, ds \le \frac{c_1}{c_2}\, e^{-c_2 t} \quad \text{for all } t \ge 0.$$

Thus we have

$$\mathbf{P}[X \ge t] \le \frac{1}{a} \int_{t-a}^t \mathbf{P}[X \ge s]\, ds \le \frac{e^{c_2 a}}{c_2 a}\, c_1 e^{-c_2 t} \quad \text{for all } t \ge a.$$

Choosing the optimal value $a = 1/c_2$ yields the result for $t \ge 1/c_2$, while the result holds trivially for $t \le 1/c_2$ as then $c_1 e^{1-c_2 t} > 1 \ge \mathbf{P}[X \ge t]$. □

Using this lemma, we readily obtain the following symmetrization bound.

Corollary 7.7 (Symmetrization tail bound). *Suppose that*

$$\mathbf{P}\left[2 \sup_{f \in \mathcal{F}} \sum_{k=1}^n \varepsilon_k f(X_k) \ge K + t\right] \le c_1 e^{-c_2 t^2} \quad \text{for all } t \ge 0$$

for some constants $c_1 \ge 1$ and $c_2, K \ge 0$. Then

$$\mathbf{P}\left[\sup_{f \in \mathcal{F}} \sum_{k=1}^n \{f(X_k) - \mu f\} \ge K + t\right] \le 3 c_1 e^{-c_2 t^2} \quad \text{for all } t \ge 0.$$

Proof. The identical proof to Lemma 7.4 shows that

$$\mathbf{E}\left[\Phi\left(\sup_{f\in\mathcal{F}}\sum_{k=1}^n\{f(X_k)-\mu f\}\right)\right]\le\mathbf{E}\left[\Phi\left(2\sup_{f\in\mathcal{F}}\sum_{k=1}^n\varepsilon_k f(X_k)\right)\right]$$

for any increasing convex function Φ. It remains to apply Lemma 7.6. □

Corollary 7.7 can now be used in conjunction with results such as Theorem 5.29 to obtain tail bounds for the empirical process in terms of chaining.

Problems

7.1 (Rademacher and Gaussian processes). Let $T\subseteq\mathbb{R}^n$. In the proof of Lemma 7.4, we have seen that we can always bound

$$r(T):=\mathbf{E}\left[\sup_{t\in T}\sum_{k=1}^n\varepsilon_k t_k\right]\lesssim\mathbf{E}\left[\sup_{t\in T}\sum_{k=1}^n g_k t_k\right]=:g(T),$$

where $\varepsilon_1,\ldots,\varepsilon_n$ are i.i.d. symmetric Bernoulli and $g_1,\ldots,g_n$ are i.i.d. $N(0,1)$. Unfortunately, the converse inequality does not hold in general.

a. Show for $T=\{t\in\mathbb{R}^n:\|t\|_1\le 1\}$ that $r(T)\sim 1$ and $g(T)\sim\sqrt{2\log n}$.

b. Evidently $r(T)$ can be small for two distinct reasons: either because $g(T)$ is small, or because the ℓ_1-diameter $\sup_{t\in T}\|t\|_1$ is small. Combine these as follows: if $T\subseteq T_1+T_2$ with $g(T_1)\le a$ and $\sup_{t\in T_2}\|t\|_1\le a$, then $r(T)\le 2a$.

A deep result, conjectured by Talagrand and proved by Bednorz and Latała, shows that this idea captures completely the behavior of the Rademacher process: if $r(T)\le a$, then $T\subseteq T_1+T_2$ for some T_1,T_2 such that $g(T_1)\le ca$ and $\sup_{t\in T_2}\|t\|_1\le ca$. This result is proved by a very sophisticated form of the generic chaining method, and is beyond our scope.

In the example of part a., $r(T)$ and $g(T)$ are apart by a factor $\sim\sqrt{\log n}$. It turns out that this is the worst case situation: we always have

$$r(T)\lesssim g(T)\lesssim r(T)\sqrt{\log n}.$$

This could be deduced from Bednorz and Latała, but we give a direct proof.

c. Show that if $|a_1|,\ldots,|a_n|\le 1$, then

$$\mathbf{E}\left[\sup_{t\in T}\sum_{k=1}^n\varepsilon_k t_k a_k\right]\le\mathbf{E}\left[\sup_{t\in T}\sum_{k=1}^n\varepsilon_k t_k\right].$$

Hint: $(a_1,\ldots,a_n)\mapsto\mathbf{E}[\sup_{t\in T}\sum_{k=1}^n\varepsilon_k t_k a_k]$ is convex.

d. Conclude that $g(T)\lesssim r(T)\sqrt{\log n}$.

We have seen above that in general, the supremum of a Rademacher process and a Gaussian process can be far apart. However, in the context of the symmetrization Lemma 7.4, the situation should be much better than in the general case: here the supremum is taken over the random set $T = \{(f(X_1), \ldots, f(X_n)) : f \in \mathcal{F}\}$. Informally speaking, the typical magnitude of the ℓ_1-norm of an element of this set is of order n, so we expect that $r(T)$ can be small only if $g(T)$ is small. Let us try to prove such a result.

e. Provided $\{\varepsilon_1, \ldots, \varepsilon_k\}$, $\{g_1, \ldots, g_k\}$, $\{X_1, \ldots, X_k\}$ are independent, show

$$\begin{aligned}\mathbf{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^n g_k f(X_k)\right] &\le \int_0^\infty \mathbf{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^n \varepsilon_k \mathbf{1}_{|g_k|\ge x} f(X_k)\right] dx \\ &= \int_0^\infty \mathbf{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^{|\{k\le n:|g_k|\ge x\}|} \varepsilon_k f(X_k)\right] dx.\end{aligned}$$

Hint: use $|g_k| = \int_0^\infty \mathbf{1}_{|g_k|\ge x}\, dx$.

f. Let $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ be a concave increasing function. Suppose that

$$\mathbf{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^n \varepsilon_k f(X_k)\right] \le \varphi(n) \quad \text{for all } n \ge 1.$$

Show that

$$\mathbf{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^n g_k f(X_k)\right] \le \int_0^\infty \varphi(n\mathbf{P}[|g_1| \ge x])\, dx \quad \text{for all } n \ge 1.$$

In particular, if we choose $\varphi(n) = cn^\alpha$ for $\frac{1}{2} \le \alpha < 1$, then we can control the Gaussian and Rademacher symmetrizations by the same rate.

7.2 (The Glivenko-Cantelli theorem). Let $X_1, X_2, \ldots$ be i.i.d. random variables with distribution μ on a measurable space $\mathbb{X}$, and let $\mathcal{F}$ be a class of functions on $\mathbb{X}$. For simplicity, we will assume throughout this problem that the class $\mathcal{F}$ is uniformly bounded (and, as we have implicitly assumed throughout these notes, that the suprema we will encounter are measurable). The class of functions $\mathcal{F}$ is said to be *μ-Glivenko-Cantelli* if

$$\mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{k=1}^n \{f(X_k) - \mu f\}\right|\right] \xrightarrow{n\to\infty} 0.$$

Technically speaking, such a class is called weak Glivenko-Cantelli, as opposed to the strong Glivenko-Cantelli property that requires a.s. convergence.

a. Show that the weak Glivenko-Cantelli property implies the strong Glivenko-Cantelli property in the setting of this problem (of uniformly bounded $\mathcal{F}$).

Hint: use a suitable concentration inequality and Borel-Cantelli.

Symmetrization is a key tool to understand Glivenko-Cantelli classes. Let $\varepsilon_1, \varepsilon_2, \ldots$ and $g_1, g_2, \ldots$ be i.i.d. Rademacher and Gaussian variables as usual.

b. Show that $\mathcal{F}$ is Glivenko-Cantelli if and only if

$$\mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{k=1}^n \varepsilon_k f(X_k)\right|\right] \xrightarrow{n\to\infty} 0.$$

Hint: use $|\sum_{k=1}^n \varepsilon_k f(X_k)| \le |\sum_{k=1}^n \varepsilon_k\{f(X_k)-\mu f\}| + \|f\|_\infty|\sum_{k=1}^n \varepsilon_k|$.

c. In the previous problem we discussed a method to reverse the inequality between Rademacher and Gaussian symmetrization. In the present setting it will be useful to prove the following related inequality: for any $M \ge 0$

$$\mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{k=1}^n g_k f(X_k)\right|\right] \le \frac{\|\mathcal{F}\|_\infty}{M} + M\,\mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{k=1}^n \varepsilon_k f(X_k)\right|\right].$$

Hint: insert $1 = \mathbf{1}_{|g_k|\le M} + \mathbf{1}_{|g_k|>M}$ inside the Gaussian symmetrization.

d. Show that $\mathcal{F}$ is Glivenko-Cantelli if and only if

$$\mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{k=1}^n g_k f(X_k)\right|\right] \xrightarrow{n\to\infty} 0.$$

We are now ready to give a necessary and sufficient condition for the Glivenko-Cantelli propery in terms of the random geometry of the set $\mathcal{F}$: we claim that $\mathcal{F}$ is Glivenko-Cantelli if and only if the following condition $(*)$ holds:

$$\frac{\log N(\mathcal{F}, \|\cdot\|_{L^2(\mu_n)}, \varepsilon)}{n} \xrightarrow{n\to\infty} 0 \quad \text{in probability} \quad \text{for every } \varepsilon > 0.$$

Here μ_n denotes the empirical measure of $X_1, \ldots, X_n$.

e. Show that condition $(*)$ is sufficient for the Glivenko-Cantelli property.
Hint: use Lemma 5.7.

f. Show that condition $(*)$ is necessary for the Glivenko-Cantelli property.
Hint: use Sudakov's inequality.

7.3 (Self-normalized sums). Consider independent Gaussian random variables $X_1, \ldots, X_n$ with $\mathbf{E}[X_i] = 0$ and $\mathrm{Var}[X_i] = \sigma_i^2$. Obviously we have

$$\mathbf{P}\left[\sum_{i=1}^n X_i \ge t\left\{\sum_{i=1}^n \sigma_i^2\right\}^{1/2}\right] \le e^{-t^2/2} \quad \text{for all } t \ge 0.$$

Can one obtain similar inequalities when the variables X_i are not Gaussian? By Azuma's inequality (Lemma 3.7), we obtain the same result if X_i is σ_i^2-subgaussian. However, for general random variables, there is no hope to obtain

such an inequality. Indeed, if the variables X_i have heavy tails, for example, then clearly the sum cannot have a Gaussian tail for large t.

Remarkably, there is a method to obtain Gaussian inequalities of this type that works without any tail assumption on the random variables! The key idea is to choose a random normalization that plays the role of the sum of the variances in the Gaussian case. We then say the sum is *self-normalized*.

a. Consider first the simplest case of independent random variables X_i that all have symmetric distributions. Show that

$$\mathbf{P}\left[\sum_{i=1}^n X_i \ge t\left\{\sum_{i=1}^n X_i^2\right\}^{1/2}\right] \le e^{-t^2/2} \quad \text{for all } t \ge 0.$$

Hint: apply Hoeffding's inequality conditionally.

b. Prove the following consequence of Lemma 7.6: if $c_1 \ge 1$, $c_2 > 0$ are constants and X, Y, Z are random variables such that Y is nonnegative and

$$\mathbf{P}[X \ge \sqrt{tY}] \le c_1 e^{-c_2 t} \quad \text{for all } t \ge 0,$$

then

$$\mathbf{P}[\mathbf{E}[X|Z] \ge \sqrt{t\mathbf{E}[Y|Z]}] \le c_1 e^{1-c_2 t} \quad \text{for all } t \ge 0.$$

Hint: use $\sqrt{tY} = \inf_{a>0}\{t/2a + aY/2\}$.

c. Let $X_1, \ldots, X_n$ be any independent random variables with $\mathbf{E}[X_i] = 0$ and $\mathbf{E}[X_i^2] = \sigma_i^2$. Prove the following self-normalized inequality:

$$\mathbf{P}\left[\sum_{i=1}^n X_i \ge t\left\{\sum_{i=1}^n (X_i^2 + \sigma_i^2)\right\}^{1/2}\right] \le e^{1-t^2/2} \quad \text{for all } t \ge 0.$$

7.4 (The contraction principle). Let $g_1, \ldots, g_n$ be i.i.d. $N(0,1)$. Consider

$$\mathbf{E}\left[\sup_{t\in T} \sum_{k=1}^n g_i t_i\right]$$

for a subset $T \subseteq \mathbb{R}^n$. In the best case $T = \{-t, t\}$, the magnitude of this quantity is of order $\sqrt{n}$. We informally view this rate as arising from cancellation of terms in the sum with opposite signs. When the set T is "large," however, this quantity can be much larger than $\sqrt{n}$ as the supremum can cancel some of the signs. For example, in the extreme case that $T = \{-1, 1\}^n$, we can cancel the signs exactly and the above quantity is of order n.

The above dicussion suggests that a class T with "less variability" should lead to a smaller Gaussian supremum. One simple result along these lines is

$$\mathbf{E}\left[\sup_{t\in T} \sum_{k=1}^n g_i |t_i|\right] \le \mathbf{E}\left[\sup_{t\in T} \sum_{k=1}^n g_i t_i\right].$$

This statment is an easy consequence of Slepian's inequality.

a. Prove the above bound.

We now turn our attention to the Rademacher process

$$\mathbf{E}\left[\sup_{t\in T}\sum_{k=1}^{n}\varepsilon_i t_i\right].$$

Is there an analogue for the Rademacher process of the property proved in part a.? It is not immediately clear how to proceed, as there is no Slepian inequality for Rademacher processes (in fact, the absence of such an inequality presents a major challenge in the deeper understanding of such processes!) However, there is a less powerful comparison inequality for Rademacher processes, called the *contraction principle*, that can sometimes play an analogous role to Slepian's inequality in this setting. We develop it presently.

b. Let T be a bounded subset of $\mathbb{R}^2$, and let $\varphi : \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz. Prove

$$\sup_{t\in T}\{t_1+\varphi(t_2)\}+\sup_{t\in T}\{t_1-\varphi(t_2)\}\le\sup_{t\in T}\{t_1+t_2\}+\sup_{t\in T}\{t_1-t_2\}.$$

c. Let $\varphi_i : \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz for $i \le n$. Prove the contraction principle

$$\mathbf{E}\left[\sup_{t\in T}\sum_{k=1}^{n}\varepsilon_i\varphi_i(t_i)\right]\le\mathbf{E}\left[\sup_{t\in T}\sum_{k=1}^{n}\varepsilon_i t_i\right].$$

Hint: apply the previous part conditionally on $\varepsilon_1,\dots,\varepsilon_{i-1},\varepsilon_{i+1},\dots,\varepsilon_n$.

d. Deduce the Rademacher analogue of the above Gaussian inequality:

$$\mathbf{E}\left[\sup_{t\in T}\sum_{k=1}^{n}\varepsilon_i|t_i|\right]\le\mathbf{E}\left[\sup_{t\in T}\sum_{k=1}^{n}\varepsilon_i t_i\right].$$

e. Let $\mathcal{F}$ be a uniformly bounded class of functions with $\|f\|_\infty \le M$ for all $f \in \mathcal{F}$. In various applications, it proves to be important to control the empirical process over the family of *squares* f^2. Show that

$$\mathbf{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^{n}\{f(X_k)^2-\mu(f^2)\}\right]\le 4M\,\mathbf{E}\left[\sup_{f\in\mathcal{F}}\sum_{k=1}^{n}\varepsilon_k f(X_k)\right],$$

so that it is possible to control the empirical process using the covering numbers of $\mathcal{F}$ itself (rather than the covering numbers of $\mathcal{F}^2=\{f^2 : f\in\mathcal{F}\}$ that would arise from a direct application of symmetrization).

Let us note that with a bit more work, we can also deduce a version of the contraction principle that makes it possible to obtain tail bounds by including a convex function as we did for symmetrization in the proof of Corollary 7.7.

7.2 Vapnik-Chervonenkis combinatorics

In the previous section, we saw that we can bound using symmetrization

$$\mathbf{E}\left[\sup_{f\in\mathcal{F}} G_n(f)\right] \lesssim \mathbf{E}\left[\int_0^\infty \sqrt{\log N(\mathcal{F}, \|\cdot\|_{L^2(\mu_n)}, \varepsilon)}\, d\varepsilon\right].$$

This is a vast improvement over the result that we would have obtained by chaining directly using the Azuma-Hoeffding inequality, in which case the covering number would be replaced by the much larger quantity $N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)$. The difficulty in applying the above bound, however, is that we must control the random covering numbers $N(\mathcal{F}, \|\cdot\|_{L^2(\mu_n)}, \varepsilon)$. Unfortunately, it is often difficult to obtain bounds that exploit the specific structure of the random geometry of $(\mathcal{F}, L^2(\mu_n))$. We therefore concentrate on the intermediate problem of controlling the random covering numbers *uniformly*:

$$N(\mathcal{F}, \|\cdot\|_{L^2(\mu_n)}, \varepsilon) \le \|N(\mathcal{F}, \|\cdot\|_{L^2(\mu_n)}, \varepsilon)\|_\infty \le N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon).$$

At first sight, one might expect that uniform control of the random covering numbers would essentially reduce to covering in the uniform norm, as all the structure of the original distribution μ is lost. Surprisingly, this intuition proves to be incorrect: in many cases, the *combinatorial* structure of the class $\mathcal{F}$ makes it possible to control its uniform covering numbers very effectively, while covering in the uniform norm leads to useless bounds. We have seen in Example 7.2 that the latter difficulty already arises in an extreme manner for classes of indicator functions. We therefore begin in this section by investigating this situation: that is, we will assume that $\mathcal{F} = \{\mathbf{1}_C : C \in \mathcal{C}\}$ for a class of sets $\mathcal{C}$. Such problems are of significant interest in their own right in many applications, and also serve to illustrate the ideas that we are about to develop in the simplest possible setting. In the following section, we will extend these ideas to general classes of functions.

As we will be working exclusively with sets in this section, we will simplify our notation by implicitly identifying sets and their indicator functions; in particular, we denote by $(\mathcal{C}, \|\cdot\|)$ the class of sets $\mathcal{C}$ with the metric $\|\mathbf{1}_C - \mathbf{1}_{C'}\|$. Let us begin by recalling the difficulty with using the uniform norm: clearly $\|\mathbf{1}_C - \mathbf{1}_{C'}\|_\infty = 1$ whenever $C \ne C'$, so a moment's reflection will show that

$$N(\mathcal{C}, \|\cdot\|_\infty, \varepsilon) = |\mathcal{C}| \qquad \text{for } \varepsilon < 1.$$

As $|\mathcal{C}| = \infty$ in most cases of interest, this is useless. How can symmetrization beat this limitation? In fact, symmetrization can help us in two distinct ways:

1. The symmetrized bound requires covering only in L^2 rather than L^∞.

2. The symmetrized bound involves only norms supported on the finite set $\operatorname{supp}\mu_n = \{X_1, \ldots, X_n\}$ rather than the entire space $\mathbb{X}$.

The combination of these two ideas will lead to a powerful machinery to control the covering numbers in the symmetrization bound. In order to gain insight into the roles played by each of these ideas, we will begin by disregarding the first point completely and see how far we can get by exploiting only the reduction in complexity provided by the second point. Once this idea has been understood, we will return to the first point and show how it can be exploited to further reduce the complexity of the problem.

In order to exploit the reduction of the underlying space to a finite set, let us bound the random covering numbers in the most naive manner possible:

$$N(\mathcal{C}, \|\cdot\|_{L^2(\mu_n)}, \varepsilon) \le N(\mathcal{C}, \|\cdot\|_{L^\infty(\mu_n)}, \varepsilon) = |\mathcal{C} \cap \{X_1, \ldots, X_n\}|.$$

As $\mathcal{C} \cap \{X_1, \ldots, X_n\}$ consists of subsets of at most n points, the above quantity is bounded by at most 2^n. Thus this naive bound already improves over covering in the uniform norm on the entire space $\mathbb{X}$! Unfortunately, bounding the covering number by 2^n does not lead to any nontrivial result. Indeed, as the diameter of the set $\mathcal{C}$ is bounded by one, we can estimate

$$\mathbf{E}\left[\sup_{C\in\mathcal{C}} G_n(C)\right] \lesssim \mathbf{E}[\sqrt{\log|\mathcal{C} \cap \{X_1, \ldots, X_n\}|}] \lesssim \sqrt{n},$$

which we could have seen immediately from the definition of the empirical process (as $\|G_n(C)\|_\infty \le \sqrt{n}$). Of course, we cannot expect anything better at this level of generality: if $\mathcal{C}$ is the class of all (measurable) subsets of $\mathbb{X}$, then clearly $\sup_{C\in\mathcal{C}} G_n(C) = \sqrt{n}$ for any nonatomic measure μ. In order to obtain nontrivial result, we must exploit the structure of the set $\mathcal{C}$. Remarkably, it turns out that in many cases the quantity $|\mathcal{C} \cap \{X_1, \ldots, X_n\}|$ is *much* smaller than 2^n. Before we attempt to understand this phenomenon in a general setting, let us develop some intuition in two illuminating examples.

Example 7.8 (The empirical distribution function). Let us revisit the setting of Example 7.2 where $\mathbb{X} = \mathbb{R}$ and $\mathcal{C} = \{]-\infty, x] : x \in \mathbb{R}\}$. Clearly

$$\mathcal{C} \cap \{X_1, \ldots, X_n\} = \{\{X_{(n)}, \ldots, X_{(k)}\} : k = 1, \ldots, n\} \cup \{\varnothing\},$$

where $X_{(1)} \ge \cdots \ge X_{(n)}$ is the decreasing rearrangement of $X_1, \ldots, X_n$. Thus we have shown in this case that $|\mathcal{C} \cap \{X_1, \ldots, X_n\}| \le n + 1$, which is much smaller than 2^n! In particular, this implies the nontrivial result

$$\mathbf{E}\|F_n - F\|_\infty = \frac{1}{\sqrt{n}}\mathbf{E}\left[\sup_{C\in\mathcal{C}} |G_n(C)|\right] \lesssim \sqrt{\frac{\log n}{n}}.$$

It turns out that the rate that we obtained here is not optimal: we lost a logarithmic factor when we bounded the $L^2(\mu_n)$-covering number by the $L^\infty(\mu_n)$-covering number. This inefficiency will be addressed later in this section. Nonetheless, the simple argument given here already suffices to prove the classical Glivenko-Cantelli theorem discussed in Example 7.2 (it is left as an exercise to deduce a.s. convergence from convergence of the mean using McDiarmid's inequality and the Borel-Cantelli lemma).

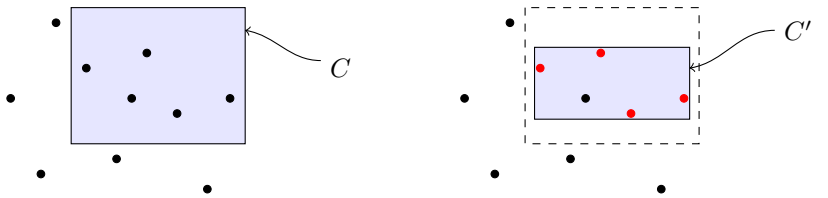
Example 7.9 (Rectangles). Let $\mathbb{X} = \mathbb{R}^2$ and let

$$\mathcal{C} = \{[a,b] \times [c,d] : a \le b,\ c \le d\}$$

be the class of axis-parallel rectangles. We claim that in this case

$$|\mathcal{C} \cap \{X_1, \ldots, X_n\}| \le n^4.$$

To see why this is the case, let us use a simple counting argument. Fix a configuration of points $X_1, \ldots, X_n$. To every rectangle $C \in \mathcal{C}$, we can associate uniquely another rectangle C' that is the smallest rectangle such that $C \cap \{X_1, \ldots, X_n\} = C' \cap \{X_1, \ldots, X_n\}$. This is illustrated in the following figure:



Note that $|\mathcal{C} \cap \{X_1, \ldots, X_n\}|$ is equal to the number of minimal rectangles C'. Each C' can be represented by specifying four points in $\{X_1, \ldots, X_n\}$, one for each edge. Thus there are at most n^4 such possibilities. (To be precise, we must account separately for the case $C = \varnothing$; however, as not every 4-tuple of points defines a valid rectangle, the crude upper bound n^4 is still valid.)

In view of this simple estimate, we can now bound the supremum of the empirical process over rectangles precisely as in the previous example.

It appears in these examples that the quantity $|\mathcal{C} \cap \{X_1, \ldots, X_n\}|$ somehow captures the number of degrees of freedom of the class $\mathcal{C}$. In the first example there was only one parameter $x \in \mathbb{R}$, and the number of sets was $\sim n$. In the second example there were four parameters $a, b, c, d \in \mathbb{R}$, and the number of sets was $\sim n^4$. This is not a coincidence: it is typically the case that a class of sets $\mathcal{C}$ of "dimension" d satisfies $|\mathcal{C} \cap \{X_1, \ldots, X_n\}| \sim n^d$. To understand this phenomenon for general classes of sets, we must understand how to define an intrinsic notion of "dimension" that does not depend on a parametrization. To this end, we introduce a *combinatorial* notion of dimension.

Definition 7.10 (Shattering). *A set $I \subseteq \mathbb{X}$ is said to be* shattered *by $\mathcal{C}$ if $\mathcal{C} \cap I = 2^I$, that is, if for every $J \subseteq I$, there exists $C \in \mathcal{C}$ such that $C \cap I = J$.*

Definition 7.11 (VC-dimension). *The* Vapnik-Chervonenkis dimension *or* VC-dimension *of $\mathcal{C}$ is defined as* $\mathrm{vc}(\mathcal{C}) := \sup\{|I| : I \text{ is shattered by } \mathcal{C}\}$.

In words, $\mathrm{vc}(\mathcal{C})$ is the cardinality of the largest set of points so that we can recover all possible subsets of these points by intersecting with elements of $\mathcal{C}$. Another way to view the VC-dimension $\mathrm{vc}(\mathcal{C})$ is as the largest integer

n such that $|\mathcal{C} \cap \{x_1, \ldots, x_n\}| = 2^n$ for some set of points $x_1, \ldots, x_n \in \mathbb{X}$. If $\mathrm{vc}(\mathcal{C}) = \infty$, then it is quite possible that $|\mathcal{C} \cap \{X_1, \ldots, X_n\}| \sim 2^n$ for all n, and there is nothing nontrivial to be gained from the present approach (at least without exploiting specific properties of the random samples $X_1, \ldots, X_n$). It is not at all obvious at this point, however, that we are any better off in the situation where $\mathrm{vc}(\mathcal{C}) < \infty$: even if $|\mathcal{C} \cap \{x_1, \ldots, x_n\}| < 2^n$ for all points $x_1, \ldots, x_n$, what is preventing us from having, say, $|\mathcal{C} \cap \{x_1, \ldots, x_n\}| \geq 2^{n/2}$? It is a remarkable combinatorial fact that this situation cannot occur: a class of sets with $\mathrm{vc}(\mathcal{C}) = d$ always satisfies $|\mathcal{C} \cap \{x_1, \ldots, x_n\}| \lesssim n^d$.

Lemma 7.12 (Sauer-Shelah). *For all $n \geq 1$ and $x_1, \ldots, x_n \in \mathbb{X}$*

$$|\mathcal{C} \cap \{x_1, \ldots, x_n\}| \leq \sum_{k=0}^{\mathrm{vc}(\mathcal{C})} \binom{n}{k} \leq \left(\frac{en}{\mathrm{vc}(\mathcal{C})}\right)^{\mathrm{vc}(\mathcal{C})}.$$

The proof of Lemma 7.12 is an exercise in combinatorics: we must find an effective way to count the subsets $|\mathcal{C} \cap \{x_1, \ldots, x_n\}|$. We will postpone the proof of this result until the end of this section, so that we can focus our attention on its implications for the control of empirical processes. Before we continue down this road, however, it is instructive to verify the validity of the Sauer-Shelah lemma in the two examples discussed above.

Example 7.13 (The empirical distribution function). In the setting of Example 7.8, it is easily seen that $\mathrm{vc}(\mathcal{C}) = 1$. Indeed, clearly any singleton $\{z\}$ is shattered, as $\{z\} \cap]-\infty, z-1] = \varnothing$ and $\{z\} \cap]-\infty, z] = \{z\}$. On the other hand, no set of two points $\{z_1, z_2\}$ is shattered: after all, if $z_1 < z_2$, then the set $\{z_2\}$ cannot be recovered by intersecting with any set in $\mathcal{C}$.

Example 7.14 (Rectangles). In the setting of Example 7.9, we claim that $\mathrm{vc}(\mathcal{C}) = 4$. It is easy to construct a set of four points that is shattered (for example, $\{(0,1), (0,-1), (1,0), (-1,0)\}$). On the other hand, choose any set I of five points, and let C be the smallest rectangle enclosing I. Then at least four points in I touch the boundary of C. But any rectangle that contains these four points must necessarily also contain the fifth, so I cannot be shattered.

As can be seen in these examples, the VC-dimension of a class of sets is often easy to compute. The beauty of this notion is that shattered sets, which are "witnesses" to high-dimensional behavior, are very rigid objects, and it is therefore often straightforward to rule out their existence in specific situations. The combinatorial principle expressed by the Sauer-Shelah lemma is consequently a powerful tool not just in theory but also in practice.

Let us now return to the control of the empirical process. By combining the Sauer-Shelah lemma with our symmetrization bound, we immediately obtain

$$\sup_\mu \mathbf{E}\left[\sup_{C \in \mathcal{C}} \{\mu_n(C) - \mu(C)\}\right] \lesssim \sqrt{\mathrm{vc}(\mathcal{C})}\sqrt{\frac{\log n}{n}},$$

where the supremum is taken over all probability measures μ on $\mathbb{X}$. This result shows not only that the law of large numbers holds uniformly over classes of sets $\mathcal{C}$ with finite VC-dimension—a far-reaching generalization of the original result of Glivenko and Cantelli discussed in Example 7.2—but we even obtain a bound on the rate of convergence that is completely independent of the distribution of the underlying independent variables! Classes $\mathcal{C}$ that satisfy this property are called *uniform Glivenko-Cantelli classes*.

Remark 7.15. The independence of our bounds of the distribution μ can be both a positive and negative feature. In applications in statistics or machine learning, where only independent samples are available and the underlying distribution μ is unknown, distribution-free estimates make it possible to evaluate the error of statistical procedures without making any assumptions on the data-generating mechanism. On the other hand, it is certainly possible for a class $\mathcal{C}$ to satisfy the μ-Glivenko-Cantelli property for some distributions μ and not for others, and the VC-dimension cannot capture this behavior. In such situations, we cannot ignore the law of the samples $X_1, \ldots, X_n$: the random geometry must be genuinely understood to obtain nontrivial results. We will encounter an example in which this can be done in Problem 7.10.

Despite that we have obtained a decidedly nontrivial result from a direct application of the Sauer-Shelah lemma, it turns out that this result is not sharp: the optimal rate in the uniform law of large numbers for classes of finite VC-dimension is in fact the usual $1/\sqrt{n}$ central limit theorem rate! Thus we have apparently picked up an extra factor of order $\sqrt{\log n}$. This origin of this inefficiency does not lie in the Sauer-Shelah lemma: our combinatorial bound

$$N(\mathcal{C}, \|\cdot\|_{L^\infty(\mu_n)}, \varepsilon) \lesssim n^{\mathrm{vc}(\mathcal{C})}$$

is sharp, as can be seen in Examples 7.8 and 7.9. The problem lies in the very first step of our analysis, where is applied the crude estimate

$$N(\mathcal{C}, \|\cdot\|_{L^2(\mu_n)}, \varepsilon) \le N(\mathcal{C}, \|\cdot\|_{L^\infty(\mu_n)}, \varepsilon).$$

The L^2-covering numbers prove to be *much* smaller than the L^∞-covering numbers: while the latter must necessarily grow with n, the former do not depend on n at all! In fact, it turns out that the space $(\mathcal{C}, \|\cdot\|_{L^2(\mu)})$ has metric dimension $\propto \mathrm{vc}(\mathcal{C})$, uniformly over all probability measures μ.

Theorem 7.16 (Dudley). *There is a universal constant K such that*

$$\sup_\mu N(\mathcal{C}, \|\cdot\|_{L^2(\mu)}, \varepsilon) \le \left(\frac{K}{\varepsilon}\right)^{K\,\mathrm{vc}(\mathcal{C})} \qquad \text{for all } \varepsilon < 1.$$

Where did the dependence on n disappear to? The idea is suprisingly simple. Suppose that $\{C_1, \ldots, C_m\}$ is a maximal ε-packing of $(\mathcal{C}, \|\cdot\|_{L^2(\mu)})$:

that is, $\|\mathbf{1}_{C_i} - \mathbf{1}_{C_j}\|_{L^2(\mu)} > \varepsilon$ for all $i \neq j$. If we draw r random samples from μ, then the law of large numbers ensures that we have

$$\|\mathbf{1}_{C_i} - \mathbf{1}_{C_j}\|_{L^2(\mu)} \approx \|\mathbf{1}_{C_i} - \mathbf{1}_{C_j}\|_{L^2(\mu_r)}.$$

Thus if we choose r large enough, then we can ensure that $\{C_1, \ldots, C_m\}$ is still an $\varepsilon/2$-packing of $(\mathcal{C}, \|\cdot\|_{L^2(\mu_r)})$, and in this case we obtain

$$N(\mathcal{C}, \|\cdot\|_{L^2(\mu)}, \varepsilon) \le N(\mathcal{C}, \|\cdot\|_{L^2(\mu_r)}, \varepsilon/4) \le N(\mathcal{C}, \|\cdot\|_{L^\infty(\mu_r)}, \varepsilon/4) \lesssim r^{\mathrm{vc}(\mathcal{C})}.$$

The key insight is now that the number of samples r that we need to draw in order to ensure that this estimate holds depends only on ε and m—the original sample size n of the empirical process is completely irrelevant! In particular, just as we previously exploited the fact that symmetrization reduces the space $\mathbb{X}$ to a finite set $\{X_1, \ldots, X_n\}$ of cardinality n, we now reduce the size of the space even further by throwing out those points that are not needed to maintain the separation between the sets C_i. The gain obtained from this reduction accounts precisely for the improvement in Theorem 7.16. This idea, called *probabilistic extraction*, is made precise by the following lemma. For future reference, we formulate it for general functions rather than sets (see Problem 7.6 for a somewhat sharper bound that is specific to sets).

Lemma 7.17 (Extraction). *Let $f_1, \ldots, f_m$ be functions on $\mathbb{X}$ such that*

$$\|f_i\|_\infty \le 1, \qquad \|f_i - f_j\|_{L^2(\mu)} > \varepsilon \qquad \text{for all } 1 \le i < j \le m.$$

Then there exist $r \le c\varepsilon^{-4} \log m$ points $x_1, \ldots, x_r \in \mathbb{X}$ such that

$$\|f_i - f_j\|_{L^2(\mu^x)} > \varepsilon/2 \qquad \text{for all } 1 \le i < j \le m,$$

where $\mu^x := \frac{1}{r}\sum_{k=1}^r \delta_{x_k}$ and c is a universal constant.

Proof. Let $X_1, \ldots, X_r \sim \mu$ be i.i.d., and denote by μ_r their empirical measure. Then we can estimate using the Azuma-Hoeffding inequality

$$\mathbf{P}\left[\|f_i - f_j\|^2_{L^2(\mu_r)} \le \frac{\varepsilon^2}{4}\right] \le \mathbf{P}\left[\|f_i - f_j\|^2_{L^2(\mu_r)} \le \|f_i - f_j\|^2_{L^2(\mu)} - \frac{3\varepsilon^2}{4}\right] \le e^{-r\varepsilon^4/15}$$

for every $i \neq j$. A union bound now gives

$$\mathbf{P}\left[\|f_i - f_j\|_{L^2(\mu_r)} > \frac{\varepsilon}{2} \text{ for all } i \neq j\right] \ge 1 - m^2 e^{-r\varepsilon^4/15} > 0$$

for $r > 30\varepsilon^{-4} \log m$, and the result follows readily. □

We can now easily complete the proof of Theorem 7.16.

Proof (Theorem 7.16). Let μ be any probability on $\mathbb{X}$, and let $C_1, \ldots, C_m$ be a maximal ε-packing of $(\mathcal{C}, \|\cdot\|_{L^2(\mu)})$. By Lemma 7.17, there exist $r \le c\varepsilon^{-4} \log m$ points $x_1, \ldots, x_r$ so that $C_1, \ldots, C_m$ is still a packing of $(\mathcal{C}, \|\cdot\|_{L^2(\mu^x)})$. Thus

$$m \le |\mathcal{C} \cap \{x_1, \ldots, x_r\}| \le \left(\frac{er}{\mathrm{vc}(\mathcal{C})}\right)^{\mathrm{vc}(\mathcal{C})} \le \left(\frac{\log m}{\mathrm{vc}(\mathcal{C})} \frac{(ec)^{1/4}}{\varepsilon}\right)^{4\,\mathrm{vc}(\mathcal{C})}$$

by the Sauer-Shelah lemma. But using $\alpha \log m \le m^\alpha$, we obtain

$$m^{1/2} \le \left(\frac{2(ec)^{1/4}}{\varepsilon}\right)^{4\,\mathrm{vc}(\mathcal{C})},$$

and the proof is complete as $m \ge N(\mathcal{C}, \|\cdot\|_{L^2(\mu)}, \varepsilon)$ by Lemma 5.12. □

With the proof of Theorem 7.16 being complete, we have now accomplished what we set out to do at the beginning of this section: we obtained uniform control over the L^2-covering numbers of a class of sets $\mathcal{C}$ in terms of its combinatorial structure. In particular, we can now obtain the optimal rate in the uniform law of large numbers for classes of finite VC-dimension.

Corollary 7.18 (Uniform Glivenko-Cantelli classes). *There is a universal constant L such that for any class $\mathcal{C}$ of measurable subsets of $\mathbb{X}$ and $n \ge 1$*

$$\sup_\mu \mathbf{E}\left[\sup_{C \in \mathcal{C}} |\mu_n(C) - \mu(C)|\right] \le L\sqrt{\frac{\mathrm{vc}(\mathcal{C})}{n}},$$

where the supremum is taken over all probability measures μ on $\mathbb{X}$.

Proof. Using symmetrization and Theorem 7.16 we obtain

$$\begin{aligned}\mathbf{E}\left[\sup_{C \in \mathcal{C}} |\mu_n(C) - \mu(C)|\right] &\le \frac{K'}{\sqrt{n}}\,\mathbf{E}\left[\int_0^1 \sqrt{\log N(\mathcal{C}, \|\cdot\|_{L^2(\mu_n)}, \varepsilon)}\,d\varepsilon\right] \\ &\le \sqrt{\frac{\mathrm{vc}(\mathcal{C})}{n}}\,K'\sqrt{K}\int_0^1 \sqrt{\log\frac{K}{\varepsilon}}\,d\varepsilon,\end{aligned}$$

where K' is the universal constant that arises in Corollary 5.25 and we have used that the diameter of $(\mathcal{C}, \|\cdot\|_{L^2(\mu)})$ is at most one. □

It remains to take care of unfinished business: we must still prove the Sauer-Shelah lemma. The remainder of the section will be devoted to this task. There are in fact a number of different proofs of the Sauer-Shelah lemma, each of which is interesting in its own right. We will develop in some detail a proof that is loosely reminiscent of the lower bound construction in the proof of the majorizing measure theorem. In the case of classes of sets, this proof is somewhat pedantic; the same basic step can be used to give a shorter proof by induction on the dimension (Problem 7.7). However, the ideas that we will develop will prove to be particularly useful in the next section when we attempt to extend the conclusion of Theorem 7.16 to classes of functions.

The conclusion of the Sauer-Shelah lemma is in fact an immediate consequence of the following more precise combinatorial principle.

Theorem 7.19 (Pajor). *For any class $\mathcal{C}$ of subsets of $\mathbb{X}$, we have*

$$|\mathcal{C}| \le |\{I \subseteq \mathbb{X} : I \text{ is shattered by } \mathcal{C}\}|.$$

Let us see why Lemma 7.12 follows.

Proof (Lemma 7.12). By the definition of the VC-dimension, every shattered set I must satisfy $|I| \le \mathrm{vc}(\mathcal{C})$. Thus Theorem 7.19 implies

$$\begin{aligned}|\mathcal{C} \cap \{x_1, \dots, x_n\}| &\le |\{I \subseteq \{x_1, \dots, x_n\} : I \text{ is shattered by } \mathcal{C}\}| \\ &\le |\{I \subseteq \{x_1, \dots, x_n\} : |I| \le \mathrm{vc}(\mathcal{C})\}| = \sum_{k=0}^{\mathrm{vc}(\mathcal{C})} \binom{n}{k}.\end{aligned}$$

The remaining bound in Lemma 7.12 is an elementary consequence of the binomial theorem: for any $d \le n$ we can estimate

$$\left(\frac{d}{n}\right)^d \sum_{k=0}^{d} \binom{n}{k} \le \sum_{k=0}^{d} \binom{n}{k} \left(\frac{d}{n}\right)^k = \left(1 + \frac{d}{n}\right)^n \le e^d.$$

Thus the proof of Lemma 7.12 is hereby complete. □

Remark 7.20. It is not difficult to see that Theorem 7.19 and Lemma 7.12 are sharp. Indeed, consider the class $\mathcal{C} = \{I \subseteq \{1, \dots, n\} : |I| \le d\}$. Then every subset of cardinality d is shattered, and clearly no set of cardinality greater than d can be shattered. Thus $\mathrm{vc}(\mathcal{C}) = d$, and in this example the result of Theorem 7.19 and the first inequality in Lemma 7.12 hold with equality.

We now finally turn to the heart of the matter, which is to prove Theorem 7.19. The essential difficulty that we face is to devise an efficient way to organize our counting of the number of shattered sets. This requires some form of bookkeeping. To this end, we will build a tree (cf. Definition 6.37) of subsets of $\mathcal{C}$—that is, each node of the tree represents a family of sets in $\mathcal{C}$—that encodes information about what points are shattered.

Definition 7.21 (Splitting tree). *Let $\mathcal{C}$ be a class of subsets of $\mathbb{X}$. A $\mathcal{C}$-tree A is called a* splitting tree *if every node $\mathcal{A} \in \mathsf{A}$ that is not a leaf satisfies:*

1. $\mathcal{A}$ has exactly two children $\mathcal{A}_+$ and $\mathcal{A}_-$;
2. There exists $x_{\mathcal{A}} \in \mathbb{X}$ so that $x_{\mathcal{A}} \in C$ for $C \in \mathcal{A}_+$ and $x_{\mathcal{A}} \notin C$ for $C \in \mathcal{A}_-$.

The motivation for this definition is that a set $I = \{x_1, \dots, x_n\}$ is shattered if and only if there exists a splitting tree A with the following properties:

1. A is a complete binary tree of depth n.
2. $\{x_{\mathcal{A}} : \mathcal{A} \in \mathsf{A}\} = \{x_1, \dots, x_n\}$.

Indeed, suppose such a tree exists. Then for any $J \subseteq I$, we can find a set $C \in \mathcal{C}$ such that $C \cap I = J$ (thereby verifying that I is shattered) by using the tree as a lookup table: starting at the root $\mathcal{C}$, traverse down the unique path in the tree such that at every node $\mathcal{A}$, we move to $\mathcal{A}_+$ if $x_{\mathcal{A}} \in J$ and to $\mathcal{A}_-$ otherwise. We end up at a leaf $\mathcal{A}_J$ of the tree, and by construction any $C \in \mathcal{A}_J$ satisfies $C \cap I = J$. Conversely, if I is shattered, then

$$\mathsf{A} = \{\{C \in \mathcal{C} : C \cap \{x_1, \ldots, x_i\} = J\} : 0 \le i \le n,\ J \subseteq \{x_1, \ldots, x_i\}\}$$

evidently defines a splitting tree with the above two special properties.

In view of the above discussion, finding shattered sets is equivalent to finding *complete* splitting trees. The difficulty is that complete splitting trees are hard to find. However, it is very easy to construct a splitting tree without the above special properties by repeatedly splitting each node of the tree into two subsets in a "greedy" fashion starting at the root. The idea behind the proof of Theorem 7.19 is to show that any large splitting tree must contain many subtrees that are complete. This is a simple example of the *Ramsey phenomenon* that arises in many combinatorial problems, which states that that any "large" system must contain large "highly structured" subsystems.

Lemma 7.22. *Let $\mathcal{C}$ be a class of subsets of $\mathbb{X}$. Then for any splitting tree A*

$$|\{\textit{leaves of } \mathsf{A}\}| \le |\{I \subseteq \mathbb{X} : I \textit{ is shattered by } \mathcal{C}\}|.$$

Proof. It is convenient to define for $\mathcal{A} \subseteq \mathcal{C}$

$$\mathcal{S}(\mathcal{A}) := \{I \subseteq \mathbb{X} : I \text{ is shattered by } \mathcal{A}\},$$

where we note that $\varnothing \in \mathcal{S}(\mathcal{A})$ for any $\mathcal{A}$. The key point of the proof is that

$$|\mathcal{S}(\mathcal{A})| \ge |\mathcal{S}(\mathcal{A}_+)| + |\mathcal{S}(\mathcal{A}_-)|$$

holds for every node $\mathcal{A} \in \mathsf{A}$ that is not a leaf. To see this, note that if a set I is shattered by a subfamily of $\mathcal{A}$, then it is shattered by $\mathcal{A}$ as well by definition. Thus the only issue we have to address is that sets I that are shattered *both* by $\mathcal{A}_+$ and $\mathcal{A}_-$ are double-counted in the lower bound. On the other hand, if I is shattered by both $\mathcal{A}_+$ and $\mathcal{A}_-$, then it is easily verified that both I and $I \cup \{x_{\mathcal{A}}\}$ are shattered by $\mathcal{A}$. Thus the claim is valid. To complete the proof, it remains to iterate the above inequality starting from the root. This yields

$$|\mathcal{S}(\mathcal{C})| \ge \sum_{\mathcal{A} \text{ is a leaf}} |\mathcal{S}(A)| \ge |\{\text{leaves of } \mathsf{A}\}|,$$

where we have used that $|\mathcal{S}(\mathcal{A})| \ge 1$ (because $\varnothing \in \mathcal{S}(\mathcal{A})$). □

To complete the proof of Theorem 7.19, it remains to construct a splitting tree with $|\mathcal{C}|$ leaves. But this is trivial: the most naive construction works.

Lemma 7.23. *For any class of sets $\mathcal{C}$, there exists a splitting tree* A *with*

$$|\{\textit{leaves of } \mathsf{A}\}| = |\mathcal{C}|.$$

Proof. It is trivial that for any subset $\mathcal{A} \subseteq \mathcal{C}$ with $|\mathcal{A}| \geq 2$, we can choose $x_{\mathcal{A}} \in \mathbb{X}$ such that $\mathcal{A}_+ = \{C \in \mathcal{A} : x_{\mathcal{A}} \in C\}$ and $\mathcal{A}_- = \{C \in \mathcal{A} : x_{\mathcal{A}} \notin C\}$ are nonempty: indeed, it suffices to choose any $x_{\mathcal{A}} \in C \triangle C'$ for distinct elements $C, C' \in \mathcal{A}$. Thus we can grow a splitting tree by starting at the root $\mathcal{C}$ and repeatedly splitting the leaves of the tree into two subsets until all the leaves are singletons. As we have not thrown out any elements of $\mathcal{C}$, the leaves form a partition of $\mathcal{C}$, and as each leaf is a singleton the conclusion follows. □

Combining Lemmas 7.22 and 7.23 concludes the proof of Theorem 7.19.

Problems

7.5 (Computing the VC-dimension). The aim of this problem is to compute the VC-dimension of various classes of sets $\mathcal{C}$. We begin with a simple observation that is useful in many geometric situations.

a. Let $\mathcal{C}$ be a class of *convex* subsets of $\mathbb{R}^d$. Show that if $I \subset \mathbb{R}^d$ is shattered by $\mathcal{C}$, then every $x \in I$ must be an extreme point of the convex hull $\operatorname{conv} I$.

We now consider several interesting examples of classes of convex sets.

b. Show that $\operatorname{vc}(\mathcal{C}) = 3$ for the class of discs in the plane

$$\mathcal{C} = \{\{x \in \mathbb{R}^2 : \|x - z\| \leq r\} : z \in \mathbb{R}^2,\ r \in \mathbb{R}_+\}.$$

Hint: suppose that $\{x_1, x_2, x_3, x_4\}$ are the corners of a convex polygon, listed in clockwise order. Suppose that $\|x_1 - x_3\| \geq \|x_2 - x_4\|$. Show that no disc can contain x_1, x_3 without containing either x_2 or x_4.

c. Show that $\operatorname{vc}(\mathcal{C}) = d + 1$ for the class of d-dimensional halfspaces

$$\mathcal{C} = \{\{x \in \mathbb{R}^d : \langle z, x \rangle \geq a\} : z \in \mathbb{R}^d,\ a \in \mathbb{R}\}.$$

Hint: consider $\{0, e_1, \ldots, e_d\}$ (where $\{e_i\}$ denotes the unit basis in $\mathbb{R}^d$). On the other hand, for any $\{x_1, \ldots, x_{d+2}\}$, one can find $b \in \mathbb{R}^{d+2} \backslash \{0\}$ such that $b_1 x_1 + \cdots + b_{d+2} x_{d+2} = 0$ and $b_1 + \cdots + b_{d+2} = 0$.

d. Show that $\operatorname{vc}(\mathcal{C}) = 7$ for the class of all triangles

$$\mathcal{C} = \{\operatorname{conv}\{x_1, x_2, x_3\} : x_1, x_2, x_3 \in \mathbb{R}^2\}.$$

Hint: consider 7 points lying on a circle. On the other hand, let $\{x_1, \ldots, x_8\}$ be the corners of a convex polygon, listed in clockwise order. Show that no triangle can contain x_1, x_3, x_5, x_7 but exclude x_2, x_4, x_6, x_8, as every pair x_i, x_{i+2} must be separated from x_{i+1} by an edge of the triangle.

Let us note that triangles are naturally described by 6 parameters, while the VC-dimension is 7. Similarly, halfspaces can be described by d parameters (as we may assume without loss of generality that $\|z\| = 1$), while the VC-dimension is $d+1$. Thus it is not always the case that the VC-dimension of a parametrized family of sets equals the number of parameters.

The following construction provides a useful method to generate classes of sets with small VC-dimension that can have complicated structure.

e. Let $\mathbb{X}$ be any set, and let $g_1, \ldots, g_d : \mathbb{X} \to \mathbb{R}$ be arbitrary functions. Show that $\mathrm{vc}(\mathcal{C}) \le d$ if we define the class of upper level sets

$$\mathcal{C} = \{\{x \in \mathbb{X} : b_1 g_1(x) + \cdots + b_d g_d(x) \ge 0\} : b_1, \ldots, b_d \in \mathbb{R}\}.$$

Use this to give another proof of the VC-dimension of discs in part b.

Finally, we note that even "nice" sets can have infinite VC-dimension.

f. Show that $\mathrm{vc}(\mathcal{C}) = \infty$ for

$$\mathcal{C} = \{C \subset \mathbb{R}^2 : C \text{ is compact and convex}\}.$$

Hint: consider n points on a circle.

7.6 (A sharper uniform covering bound). Theorem 7.16, as we have stated it, implies that the metric dimension of $(\mathcal{C}, \|\cdot\|_{L^2(\mu)})$ is at most $K\,\mathrm{vc}(\mathcal{C})$ uniformly over all probability measures μ. The constant K that we obtained in not sharp. The reason for this is that we have used a very general probabilistic extraction principle in the form of Lemma 7.17. For classes of sets, we can get away with a more elementary approach that leads to a better constant.

The problem with Lemma 7.17 is that it insists that the ε-packing $\{C_1, \ldots, C_m\}$ in $L^2(\mu)$ remains a $\varepsilon/2$-packing in $L^2(\mu^x)$. This strong separation will be needed when we extend to classes of functions in the next section. Here, however, we are only interested in counting $|\mathcal{C} \cap \{x_1, \ldots, x_r\}| \ge m$. Therefore, to ensure that this is the case, we only need to ensure that the sets $C_1, \ldots, C_m$ remain distinct when they are intersected with $\{x_1, \ldots, x_r\}$.

a. Let $X_1, \ldots, X_r \sim \mu$ be i.i.d. Show that

$$\mathbf{P}[C \cap \{X_1, \ldots, X_r\} = C' \cap \{X_1, \ldots, X_r\}] = \bigl\{1 - \|\mathbf{1}_C - \mathbf{1}_{C'}\|^2_{L^2(\mu)}\bigr\}^r.$$

b. Conclude that if $C_1, \ldots, C_m$ is an ε-packing in $L^2(\mu)$, then

$$\mathbf{P}[C_i \cap \{X_1, \ldots, X_r\} \ne C_j \cap \{X_1, \ldots, X_r\} \text{ for all } i \ne j] \ge 1 - m^2\{1-\varepsilon^2\}^r > 0$$

for $r > 2\varepsilon^{-2} \log m$ (compare with $r \gtrsim \varepsilon^{-4} \log m$ in Lemma 7.17!)

c. Deduce the following improved form of Theorem 7.16:

$$\sup_\mu N(\mathcal{C}, \|\cdot\|_{L^2(\mu)}, \varepsilon) \le \left(\frac{K_\delta}{\varepsilon}\right)^{(2+\delta)\,\mathrm{vc}(\mathcal{C})} \qquad \text{for all } \varepsilon < 1,\ \delta > 0,$$

where K_δ is a universal constant that depends on δ.

d. The last bound is sharp in the following sense. Let $\mathcal{C} = \{I \subset \mathbb{N} : |I| \le d\}$, so that $\mathrm{vc}(\mathcal{C}) = d$. Show that for a universal constant K'_δ depending on δ

$$\sup_\mu N(\mathcal{C}, \|\cdot\|_{L^2(\mu)}, \varepsilon) \ge \left(\frac{K'_\delta}{\varepsilon}\right)^{(2-\delta)\,\mathrm{vc}(\mathcal{C})} \qquad \text{for all } \varepsilon < 1,\ \delta > 0.$$

Hint: consider probability measures $\mu(\{n\}) \propto n^{-(1+\alpha)}$.

Evidently $2\,\mathrm{vc}(\mathcal{C})$ is the optimal value of the exponent in the behavior of the uniform covering numbers of a class of sets. In the above bounds, however, $K_\delta \to \infty$ and $K'_\delta \to 0$ as $\delta \to 0$. A delicate analysis due to Haussler shows that it is in fact possible to attain the exponent $2\,\mathrm{vc}(\mathcal{C})$ with a finite constant.

7.7 (A short induction proof of Pajor's theorem). Our proof of Theorem 7.19 introduced splitting trees as a bookkeeping device. The insight gained from this idea will pay off in the next section. In the case of sets, however, one can rewrite the proof in a much more efficient manner without any reference to splitting trees. This yields perhaps the shortest and cleanest approach.

a. Suppose that the conclusion of Theorem 7.19 holds for any class $\mathcal{C}$ of subsets of $\mathbb{X}$ with $|\mathbb{X}| = m$. Show that the conclusion also follows when $|\mathbb{X}| = m+1$.

Hint: let $|\mathbb{X}| = m+1$ and choose any $x \in \mathbb{X}$. Define $\mathcal{C}_+ = \{C \in \mathcal{C} : x \in C\}$ and $\mathcal{C}_- = \{C \in \mathcal{C} : x \notin C\}$, and apply the basic argument of Lemma 7.22.

b. Conclude the proof of Theorem 7.19 by induction on $|\mathbb{X}|$.

Let us emphasize that this proof is essentially identical to the proof we have given. Here we have simply merged the construction of the splitting tree with the proof of Lemma 7.22, so that no additional bookkeeping is needed.

7.8 (A rearrangement proof of Pajor's theorem). The goal of this problem is to give an entirely different proof of Theorem 7.19 in the spirit of extremal combinatorics. This elegant method is useful in many other problems.

Let us begin by gaining some intuition. A class $\mathcal{C}$ of subsets of a set $\mathbb{X}$ is called *hereditary* if $C \in \mathcal{C}$ implies $C' \in \mathcal{C}$ for all $C' \subseteq C$.

a. Show that Theorem 7.19 holds with equality for hereditary $\mathcal{C}$.

Evidently hereditary classes are extremal with respect to shattering. The idea we will now pursue is that an arbitrary class $\mathcal{C}$ can be transformed into a heredetary class without changing its cardinality or increasing the number of shattered sets. This will be done by a form of rearrangement (in analogy with the proof of the classical isoperimetric inequality by Steiner symmetrization).

Consider a class $\mathcal{C}$ of subsets of a *finite* set $\mathbb{X}$. The basic step that we consider is as follows. Given a point $x \in \mathbb{X}$, define $\mathscr{S}_x\mathcal{C} = \{\mathscr{S}_x C : C \in \mathcal{C}\}$ such that $\mathscr{S}_x C = C\backslash\{x\}$ if $C\backslash\{x\} \notin \mathcal{C}$, and $\mathscr{S}_x C = C$ otherwise. This operation is called *shifting*: it tries to "remove the holes" in the class $\mathcal{C}$ that prevent it from being hereditary, one coordinate at a time. Let us investigate its consequences.

b. Show that $|\mathscr{S}_x\mathcal{C}| = |\mathcal{C}|$.

c. Show that if $I \subseteq \mathbb{X}$ is shattered by $\mathscr{S}_x\mathcal{C}$, then it is also shattered by $\mathcal{C}$.

d. Show that if $\mathscr{S}_x\mathcal{C} = \mathcal{C}$ for all $x \in \mathbb{X}$, then $\mathcal{C}$ is hereditary.

e. Now starting from any class $\mathcal{C}$, repeatedly apply the operation $\mathscr{S}_x$ by cycling through the points $x \in \mathbb{X}$. Show that the transformed set $\mathscr{S}_{x_q} \cdots \mathscr{S}_{x_1}\mathcal{C}$ becomes hereditary after a finite number q of such operations.

f. Show that the conclusion of Theorem 7.19 follows readily (while we assumed here that $\mathbb{X}$ is finite, argue that this entails no loss of generality).

7.9 (Necessity of finite VC-dimension). We have seen that classes $\mathcal{C}$ with $\mathrm{vc}(\mathcal{C}) < \infty$ have many nice properties. In particular, such classes admit *distribution-free* bounds. The aim of this problem is to show that the condition $\mathrm{vc}(\mathcal{C}) < \infty$ is often also necessary to obtain distribution-free results.

Let us begin by considering the uniform covering number. We have seen

$$\mathrm{vc}(\mathcal{C}) < \infty \quad \text{implies} \quad \sup_\mu N(\mathcal{C}, \|\cdot\|_{L^2(\mu)}, \varepsilon) < \infty$$

by Theorem 7.16. Let us show, conversely, that for $\varepsilon < 1/2$

$$\mathrm{vc}(\mathcal{C}) = \infty \quad \text{implies} \quad \sup_\mu N(\mathcal{C}, \|\cdot\|_{L^2(\mu)}, \varepsilon) = \infty.$$

a. Prove the following basic result.

Lemma 7.24 (Gilbert-Varshamov). *Let $\mathcal{C} = 2^{\mathbb{X}}$ be the class of all subsets of $\mathbb{X} = \{1, \ldots, n\}$ and let $d(C, D) = |C \triangle D|$. Then $N(\mathcal{C}, d, n/4) \geq e^{n/8}$.*

Hint: use a "volume argument" with the uniform measure on $\mathcal{C}$ in the role of the volume, and use Azuma-Hoeffding to estimate the volume of d-balls.

b. Conclude that $\mathrm{vc}(\mathcal{C}) = \infty$ implies $\sup_\mu N(\mathcal{C}, \|\cdot\|_{L^2(\mu)}, \varepsilon) = \infty$ for $\varepsilon < 1/2$.
Hint: let μ be the uniform distribution on a shattered set $I \subseteq \mathbb{X}$.

Let us now consider the uniform law of large numbers. We have seen that

$$\mathrm{vc}(\mathcal{C}) < \infty \quad \text{implies} \quad \limsup_{n\to\infty} \sup_\mu \mathbf{E}\left[\sup_{C\in\mathcal{C}} |\mu_n(C) - \mu(C)|\right] = 0$$

by Corollary 7.18. Let us show, conversely, that

$$\mathrm{vc}(\mathcal{C}) = \infty \quad \text{implies} \quad \liminf_{n\to\infty} \sup_\mu \mathbf{E}\left[\sup_{C\in\mathcal{C}} |\mu_n(C) - \mu(C)|\right] > 0.$$

Thus $\mathrm{vc}(\mathcal{C}) < \infty$ is sufficient and necessary to obtain a distribution-free rate in the uniform law of large numbers (the uniform Glivenko-Cantelli property).

c. Let $\varepsilon_1, \ldots, \varepsilon_n$ be i.i.d. symmetric Bernoulli. Show that $\mathrm{vc}(\mathcal{C}) = \infty$ implies

$$\sup_\mu \mathbf{E}\left[\sup_{C\in\mathcal{C}} \left|\frac{1}{n}\sum_{k=1}^n \varepsilon_k \mathbf{1}_C(X_k)\right|\right] \geq \frac{1}{2}.$$

Hint: let μ be the uniform distribution on a shattered set of cardinality $N \gg n$, and show that $\{X_1, \ldots, X_n\}$ is shattered with high probability.

d. Conclude that the uniform Glivenko-Cantelli property fails if $\mathrm{vc}(\mathcal{C}) = \infty$.

Hint: see Problem 7.2.

Finally, we argue that the distribution-free rate obtained in Corollary 7.18 is even *quantitatively* correct up to universal constants. That is, let us show that

$$\begin{aligned} K\sqrt{\mathrm{vc}(\mathcal{C})} &\leq \liminf_{n\to\infty} \sup_\mu \mathbf{E}\left[\sup_{C\in\mathcal{C}} \sqrt{n}|\mu_n(C) - \mu(C)|\right] \\ &\leq \limsup_{n\to\infty} \sup_\mu \mathbf{E}\left[\sup_{C\in\mathcal{C}} \sqrt{n}|\mu_n(C) - \mu(C)|\right] \leq L\sqrt{\mathrm{vc}(\mathcal{C})}. \end{aligned}$$

In view of Corollary 7.18, we must only prove the lower bound.

e. Denote by $\{Z_\mu(C)\}_{C\in\mathcal{C}}$ be the centered Gaussian process whose covariance function is given by $\mathrm{Cov}[Z_\mu(C), Z_\mu(C')] = \mathrm{Cov}_\mu[\mathbf{1}_C, \mathbf{1}_{C'}]$. Show that

$$\liminf_{n\to\infty} \sup_\mu \mathbf{E}\left[\sup_{C\in\mathcal{C}} \sqrt{n}|\mu_n(C) - \mu(C)|\right] \geq \sup_\mu \mathbf{E}\left[\sup_{C\in\mathcal{C}} |Z_\mu(C)|\right].$$

f. Show that the right-hand side in the last inequality is $\gtrsim \sqrt{\mathrm{vc}(\mathcal{C})}$.

Hint: choose μ to be uniformly distributed on a shattered set I, and represent $Z_\mu(C) = |I|^{-1/2} \sum_{x\in I} g_x\{\mathbf{1}_C(x) - \mu(C)\}$ with $\{g_x\}_{x\in I}$ i.i.d. $N(0,1)$.

7.10 (Glivenko-Cantelli theorem and convex sets). We have seen in the previous problem that $\mathrm{vc}(\mathcal{C}) < \infty$ is necessary and sufficient in order for the law of large numbers to hold uniformly over $\mathcal{C}$ with a distribution-free rate. However, when $\mathrm{vc}(\mathcal{C}) = \infty$, it can still be the case that the law of large numbers holds uniformly over $\mathcal{C}$ for any given distribution μ. We characterized such classes in Problem 7.2 in terms of a random entropy condition. It turns out that in the case of sets, the entropy condition can be replaced by a random combinatorial condition: $\mathcal{C}$ is a μ-Glivenko-Cantelli class if and only if

$$\frac{\mathrm{vc}(\mathcal{C} \cap \{X_1, \ldots, X_n\})}{n} \xrightarrow{n\to\infty} 0 \quad \text{in probability,}$$

where $X_1, X_2, \ldots$ is an i.i.d. sequence of variables with distribution μ. Note that this condition can clearly hold even when $\mathrm{vc}(\mathcal{C}) = \infty$.

a. Show that the above condition implies the μ-Glivenko-Cantelli property.

Hint: use the random entropy condition of Problem 7.2.

b. Show that the μ-Glivenko-Cantelli property implies the above condition.

Hint: start with the symmetrized formulation from Problem 7.2, and use that $\mathbf{E}[\sup_{t\in T}\sum_{k\in I}\varepsilon_k t_k] \geq \mathbf{E}[\sup_{t\in T}\sum_{k\in J}\varepsilon_k t_k]$ when $J \subseteq I$.

The advantage of the combinatorial formulation is that shattered sets are very rigid structures that are often easy to detect. Nonetheless, in the present setting we must understand what *random* combinatorial structures can arise in a sample $X_1, \ldots, X_n$ from a given distribution μ, which may not be a trivial matter. Let us develop in detail one example in which this can be done.

Let $\mathcal{C}$ be the class of all compact and convex subsets of $\mathbb{X} = [0,1]^d$ (we can easily extend the following arguments to the case $\mathbb{X} = \mathbb{R}^d$ by a straightforward truncation, but this provides no additional insight). It was shown in Problem 7.5 above that $\mathrm{vc}(\mathcal{C}) = \infty$. Nonetheless, we will show that $\mathcal{C}$ is μ-Glivenko-Cantelli whenever μ has a density with respect to Lebesgue measure.

c. Find an example of a measure μ such that $\mathcal{C}$ fails to be μ-Glivenko-Cantelli. Thus the assumption that μ has a density is not superfluous.

d. Show that a set I is shattered by $\mathcal{C}$ if and only if none of the points $x \in I$ is a convex combination of the others $I\backslash\{x\}$ (that is, I is in *convex position*).

e. Show that if μ has a density with respect to Lebesgue measure, then the boundary ∂C of every convex set $C \in \mathcal{C}$ has zero measure $\mu(\partial C) = 0$.

Hint: if $0 \in \operatorname{int} C$, then $\partial C \subset (1+\varepsilon)C\backslash(1-\varepsilon)C$.

The heuristic idea behind the proof is now as follows. By the combinatorial formulation developed in the first part of this problem, we must show that among n random points $X_1, \ldots, X_n$, the maximal size of a subset that is in convex position is sublinear in n. Suppose, to the contrary, that there is a subset $I \subseteq \{X_1, \ldots, X_n\}$ with $|I| \geq \alpha n$ that is in convex position. Then the boundary of the convex set $C = \operatorname{conv} I$ has empirical measure $\mu_n(\partial C) \geq \alpha$. If we could argue $\mu_n(\partial C) \approx \mu(\partial C)$ for all $C \in \mathcal{C}$, we would have a contradiction. At first sight, it seems like this got us nowhere: we must now prove that the class $\partial\mathcal{C}$ of boundaries of convex sets is μ-Glivenko-Cantelli! But the latter problem can be addressed by exploiting the geometry of convex sets.

f. Let $\mathcal{X}_m$ be the partition of $\mathbb{X} = [0,1]^d$ into m^d cubes of side length $1/m$. Define the discretized boundary $\partial_m C = \bigcup\{B \in \mathcal{X}_m : B \cap \partial C \neq \varnothing\}$. Prove

$$\limsup_{n\to\infty} \sup_{C\in\mathcal{C}} \mu_n(\partial C) \leq \inf_{m\geq 1} \sup_{C\in\mathcal{C}} \mu(\partial_m C).$$

g. Clearly $\inf_{m\geq 1}\mu(\partial_m C) = \mu(\partial C) = 0$, but we need this conclusion to hold uniformly over $C \in \mathcal{C}$. Show that if μ is the Lebesgue measure on $\mathbb{X}$, then

$$\sup_{C\in\mathcal{C}} \mu(\partial_{3^m} C) \leq (1-3^{-d})^m \quad \text{for all } m \geq 1.$$

Hint: for $m = 1$, the partition $\mathcal{X}_3$ consists of one cube in the center of $\mathbb{X}$ surrounded by $3^d - 1$ cubes along the sides of $\mathbb{X}$. Show that if all the cubes

along the sides contain a point in ∂C, then the middle cube cannot intersect ∂C. Thus $\mu(\partial_3 C) \le (1 - 3^{-d})\mu(\partial_1 C)$. Now iterate this argument.

h. Deduce that if μ has a density with respect to Lebesgue measure, then

$$\inf_{m \ge 1} \sup_{C \in \mathcal{C}} \mu(\partial_m C) = 0.$$

i. Conclude that the combinatorial condition formulated at the beginning of this problem holds for $\mathcal{C}$ whenever μ has a density with respect to Lebesgue measure by carefully making precise the reasoning given above.

7.11 (Kolmogorov, Smirnov, and Donsker). Let $X_1, X_2, \ldots$ be i.i.d. real-valued variables with distribution function $F(x) = \mu(]-\infty, x])$, and define the empirical distribution function $F_n(x) = \mu_n(]-\infty, x])$. The classical Glivenko-Cantelli theorem states that $\|F_n - F\|_\infty \to 0$. By Corollary 7.18, the convergence even takes place at the central limit theorem rate $\|F_n - F\|_\infty \lesssim n^{-1/2}$. We might therefore wonder whether one can go one step further and show that $\sqrt{n}\|F_n - F\|_\infty$ converges weakly to some limiting distribution.

a. Let $G_n(x) := \sqrt{n}\{F_n(x) - F(x)\}$. Show that for any $x_1, \ldots, x_k \in \mathbb{R}$

$$(G_n(x_1), \ldots, G_n(x_k)) \Longrightarrow (B(F(x_1)), \ldots, B(F(x_k))) \quad \text{in distribution.}$$

Here $\{B(t)\}_{t\in[0,1]}$ is the *Brownian bridge* defined by $B(t) = W(t) - tW(1)$, where $\{W(t)\}_{t\in[0,1]}$ is standard Brownian motion.

In view of this computation, it is natural to conjecture that $\sqrt{n}\|F_n - F\|_\infty$ converges in distribution to $\|B\|_\infty$, the supremum of a Brownian bridge (note that this limiting distribution does not depend on the law μ!) This is indeed the case, as was proved by Kolmogorov and Smirnov in the 1930s, and is of siginificant importance in classical nonparametric statistics.

It is obvious from the central limit theorem that if $I \subset \mathbb{R}$ is a finite set, then $\max_{x\in I} \sqrt{n}|F_n(x) - F(x)|$ converges in distribution to $\max_{x\in I} |B(F(x))|$. It is not at all clear, however, that this is still the case for $I = \mathbb{R}$. To prove this, we must establish that $\sqrt{n}\|F_n - F\|_\infty$ can be approximated uniformly in n by $\max_{x\in I} \sqrt{n}|F_n(x) - F(x)|$ for sufficiently large finite sets I. It is here that the empirical process machinery that we have developed enters the picture.

b. Let $Q \subseteq \mathbb{R}^2$. Show that

$$\mathbf{E}\left[\sup_{(x,x')\in Q} |G_n(x) - G_n(x')|\right] \lesssim \mathbf{E}\left[\omega\left(\sup_{(x,x')\in Q} |F_n(x) - F_n(x')|\right)\right],$$

where $\omega(u) := \int_0^{\sqrt{u}} \sqrt{\log \frac{1}{\varepsilon}}\, d\varepsilon \lesssim \sqrt{u \log(1/u)}$.

c. Let $Q_\delta = \{(x, x') : |F(x) - F(x')| \le \delta\}$. Prove *asymptotic equicontinuity*

$$\lim_{\delta\downarrow 0} \limsup_{n\to\infty} \mathbf{E}\left[\sup_{(x,x')\in Q_\delta} |G_n(x) - G_n(x')|\right] = 0.$$

d. Show that there exist finite sets $I_k \subset \mathbb{R}$ such that

$$\lim_{k\to\infty} \limsup_{n\to\infty} \mathbf{E}\left[\sqrt{n}\left|\|F_n - F\|_\infty - \max_{x\in I_k} |F_n(x) - F(x)|\right|\right] = 0,$$

and conclude that

$$\sqrt{n}\|F_n - F\|_\infty \Longrightarrow \|B\|_\infty \quad \text{in distribution.}$$

From the asymptotic equicontinuity result obtained above, we can in fact derive a much more general statement of the idea that the empirical process G_k converges weakly to the Brownian bridge $B \circ F$. This result, originally due to Donsker, can be viewed as a *uniform* central limit theorem.

e. View the empirical process $x \mapsto G_n(x)$ as a random path with values in $\mathcal{L}^\infty(\mathbb{R})$. Show that for any functional $\mathbf{H} : \mathcal{L}^\infty(\mathbb{R}) \to \mathbb{R}$ that is Lipschitz in the sense $|\mathbf{H}[G] - \mathbf{H}[G']| \le L\|G - G'\|_\infty$ for all $G, G' \in \mathcal{L}^\infty(\mathbb{R})$, we have

$$\mathbf{E}[\mathbf{H}[G_n]] \to \mathbf{E}[\mathbf{H}[B \circ F]] \quad \text{as } n \to \infty.$$

(Assume for simplicity that $\mathbf{H}[G_n]$ and $\mathbf{H}[B \circ F]$ are measurable, though this is neither obvious nor always true; measurability issues of this kind arise often in the development of uniform central limit theorems.)

While we have considered the example of empirical distribution functions for sake of illustration, uniform central limit theorems can be developed in considerable generality. A class of functions $\mathcal{F}$ for which the empirical process satisfies the central limit theorem in $\mathcal{L}^\infty(\mathcal{F})$ is called a *Donsker class*. The characterization of such classes, as well as closely related questions concerning central limit theorems in Banach spaces, have historically motivated the development of many of the tools that are used to control empirical processes.

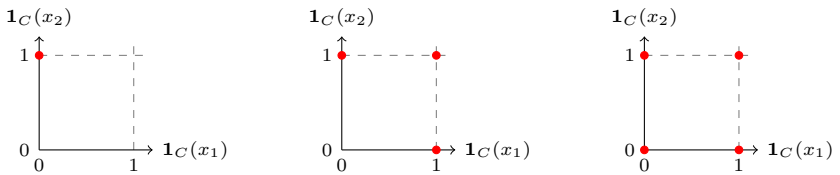
7.3 Combinatorial dimension and uniform covering

In the previous section we developed, in the special case of classes of sets, a combinatorial method to control uniformly the random covering numbers that appear in symmetrization bounds. In a sense, is not surprising that combinatorics enters the picture in this setting: as the empirical measure μ_n that arises in the symmetrization process is supported on a finite set, it is natural that our bounds for classes of sets will essentially reduce to the combinatorial problem of counting induced subsets. Whether such ideas are still useful in the general setting of classes of functions is far from clear at this point: even when restricted to a finite set, a class of functions is still a continuous object (with a potentially nontrivial geometric structure) and is not, *a priori*, combinatorial in nature. Nonetheless, the theory of previous section admits a very natural generalization to classes of functions, which we develop presently.

To gain some intuition, let us begin by reconsidering a class of sets $\mathcal{C}$ in terms of the corresponding class of indicator functions $\mathcal{F} = \{\mathbf{1}_C : C \in \mathcal{C}\}$. As indicator functions only take the values zero and one, the restricted class

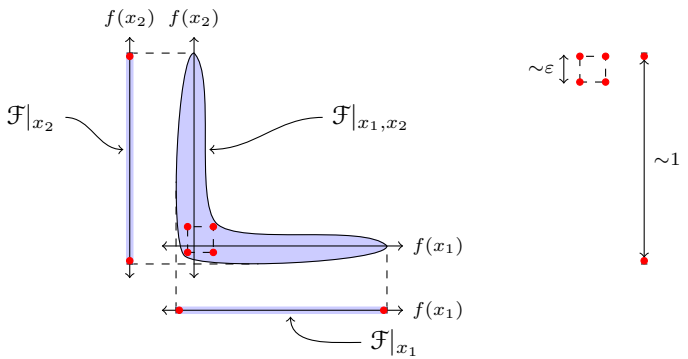
$$\mathcal{F}|_{x_1,\ldots,x_n} = \{(f(x_1),\ldots,f(x_n)) : f \in \mathcal{F}\} \subseteq \mathbb{R}^n$$

is a subset of the hypercube $\{0,1\}^n$. In particular, a set $\{x_1,\ldots,x_n\}$ is said to be shattered by $\mathcal{C}$ precisely when $\mathcal{F}|_{x_1,\ldots,x_n} = \{0,1\}^n$ is the full hypercube. Thus we can interpret $\mathrm{vc}(\mathcal{C})$ geometrically as the largest dimension of a hypercube that is contained in a coordinate projection of $\mathcal{C}$. This idea is illustrated in the following figure for different classes of subsets of $\{x_1, x_2\}$:



$\mathrm{vc}(\mathcal{C}) = 0$ $\mathrm{vc}(\mathcal{C}) = 1$ $\mathrm{vc}(\mathcal{C}) = 2$

In contrast to the special case of indicator functions, for a general class of functions $\mathcal{F}$ the restricted class $\mathcal{F}|_{x_1,\ldots,x_n}$ can be an arbitrary subset of $\mathbb{R}^n$. In analogy with the combinatorial theory of the previous section, we might try to define the VC-dimension of $\mathcal{F}$ as the largest dimension of a cube that is contained in a coordinate projection of $\mathcal{F}$. However, unlike in the case of indicator functions, there is some ambiguity in this definition in the general setting: the notion of dimension we obtain will depend on the *size* of the cubes that we consider and not just on their dimension. For example, it is perfectly possible that $\mathcal{F}$ contains only low-dimensional cubes of the form $\{0,1\}^n$, but contains high-dimensional cubes of the form $\{0,\varepsilon\}^n$ for $\varepsilon \ll 1$. To emphasize this point, consider a simple example that is illustrated in the following figure:



The projection $\mathcal{F}|_{x_1,x_2}$ contains a cube of size at most $\sim \varepsilon$, but each of the projections $\mathcal{F}|_{x_1}$ and $\mathcal{F}|_{x_2}$ contain a cube of size ~ 1. The set evidently contains

no cubes of size $\gg 1$. Thus the dimension of the set depends on the *scale* at which we are viewing it: it is zero-dimensional at very large scales (it looks like a point), it is one-dimensional at scale ~ 1 (it looks like the letter L), and it is two-dimensional at scale $\sim \varepsilon$ (where we see the "fatness" of the set). If the class $\mathcal{F}$ is defined on other points $x_3, x_4, \ldots$ as well, then the set can be higher-dimensional still when viewed as smaller scales. The dependence of the dimension on scale is not a drawback of this approach, but a genuine phenomenon: in extending the theory of the previous section to the general setting, we must introduce a *scale-sensitive* notion of dimension in order to capture the structure of the set from the point of view of covering numbers. In the remainder of this section we will make these ideas precise.

Let us begin by making precise what we mean by the statement that a coordinate projection of $\mathcal{F}$ contains a cube. The requirement that $\mathcal{F}_{x_1,\ldots,x_n}$ actually contains a copy of some hypercube $\{0,\varepsilon\}^n$ is too stringent: for example, if $\mathcal{F}_{x_1,\ldots,x_n}$ were itself a tiny perturbation of a hypercube (e.g., perturb each corner of the hypercube randomly), then it would not contain any hypercube but the dimension should not be much affected. Instead, we introduce a slightly more flexible generalization of the notion of a shattered set.

Definition 7.25 (ε-shattering). *Let $I \subseteq \mathbb{X}$ and $h \in \mathbb{R}^I$. The pair (I,h) is said to be* ε-shattered *by $\mathcal{F}$ if for every $J \subseteq I$, there exists $f \in \mathcal{F}$ such that*

$$f(x) \le h(x) \quad \text{for } x \in J, \qquad f(x) \ge h(x) + \varepsilon \quad \text{for } x \in I \backslash J.$$

The set $I \subseteq \mathbb{X}$ is said to be ε-shattered if (I,h) is ε-shattered for some $h \in \mathbb{R}^I$.

If the inequalities $f(x) \le h(x)$ and $f(x) \ge h(x) + \varepsilon$ in the definition of an ε-shattered set were replaced by equalities, then the definition would reduce to the statement that $\mathcal{F}|_I \supseteq h + \{0,\varepsilon\}^{|I|}$, that is, that the coordinate projection of $\mathcal{F}$ on I contains a (translate of the) hypercube $\{0,\varepsilon\}^{|I|}$. When the class $\mathcal{F}$ is convex these two definitions are even equivalent, see Problem 7.13. However, in the general setting, the notion of ε-shattering as defined above provides a suitable implementation of the idea that $\mathcal{F}$ contains a combinatorial structure that is "larger" than a hypercube $\{0,\varepsilon\}^{|I|}$ in the appropriate sense.

Having defined a notion of shattering for function classes, we can analogously extend the definition of VC-dimension for a given scale $\varepsilon > 0$.

Definition 7.26 (Combinatorial dimension). *The* combinatorial dimension *of $\mathcal{F}$ at scale ε is defined as* $\mathrm{vc}(\mathcal{F},\varepsilon) := \sup\{|I| : I \text{ is } \varepsilon\text{-shattered by } \mathcal{F}\}$.

Remark 7.27. $\mathrm{vc}(\mathcal{F},\varepsilon)$ is known under various different names, including *scale-sensitive dimension* or the somewhat lipectomous *fat-shattering dimension.* Note that, by its definition, $\mathrm{vc}(\mathcal{F},\varepsilon)$ is increasing as $\varepsilon \downarrow 0$.

To illustrate this notion, let us consider two useful examples.

Example 7.28 (Vector spaces). Let $\mathbb{X}$ be any set and let $f_1, \ldots, f_d : \mathbb{X} \to \mathbb{R}$ be linearly independent functions. Consider the linear class of functions

$$\mathcal{F} = \{a_1 f_1 + \cdots + a_d f_d : a_1, \ldots, a_d \in \mathbb{R}\}.$$

We claim that the combinatorial dimension of $\mathcal{F}$ is given by

$$\mathrm{vc}(\mathcal{F}, \varepsilon) = d \quad \text{for all } \varepsilon > 0.$$

Thus in this case, the dimension of $\mathcal{F}$ does not depend on the scale ε.

Let us first show that $\mathrm{vc}(\mathcal{F}, \varepsilon) \geq d$. By linear independence, we can choose $x_1, \ldots, x_d \in \mathbb{X}$ so that the matrix M with $M_{ij} = f_j(x_i)$ is nonsingular. Then for any $b \in \mathbb{R}^d$, we can find $f \in \mathcal{F}$ such that $f(x_i) = b_i$ for all i: just choose

$$f = \sum_{i=1}^{d} a_i f_i \qquad \text{with} \qquad a = M^{-1} b.$$

It follows immediately that $\{x_1, \ldots, x_d\}$ is ε-shattered.

It remains to show that $\mathrm{vc}(\mathcal{F}, \varepsilon) \leq d$. Suppose there exists an ε-shattered set $I = \{x_1, \ldots, x_{d+1}\}$. The matrix M defined above is now a $(d+1) \times d$ matrix, so there exists a vector $z \in \mathbb{R}^{d+1} \backslash \{0\}$ such that $z^* M = 0$. Thus

$$\sum_{i=1}^{d+1} z_i f(x_i) = 0 \quad \text{for all } f \in \mathcal{F}.$$

As I is ε-shattered, however, we can choose $f_+, f_- \in \mathcal{F}$ so that $f_\pm(x_i) \leq h_i$ for $\operatorname{sign} z_i = \mp 1$ and $f_\pm(x_i) \geq h_i + \varepsilon$ otherwise. Then $f = f_+ - f_- \in \mathcal{F}$ satisfies

$$\sum_{i=1}^{d+1} z_i f(x_i) \geq \varepsilon \sum_{i=1}^{d+1} |z_i| > 0,$$

which entails a contradiction. Thus $\{x_1, \ldots, x_{d+1}\}$ cannot be ε-shattered.

Example 7.29 (Functions of bounded variation). Recall that the total variation of a function $f : \mathbb{R} \to \mathbb{R}$ is defined in the following manner:

$$\|f\|_{\mathrm{var}} := \sup_n \sup_{x_1 < \cdots < x_n} \sum_{k=1}^{n-1} |f(x_{k+1}) - f(x_k)|.$$

Let us consider the class of functions of bounded variation

$$\mathcal{F} = \{f : \mathbb{R} \to \mathbb{R} : \|f\|_{\mathrm{var}} \leq V\}.$$

There are many functions of bounded variation: examples include bounded increasing functions and Lipschitz functions with compact support.

We are going to show that the combinatorial dimension of $\mathcal{F}$ satisfies

$$\mathrm{vc}(\mathcal{F},\varepsilon) = 1 + \left\lfloor \frac{V}{\varepsilon} \right\rfloor \quad \text{for all } \varepsilon > 0.$$

Thus, unlike in the previous example, the class $\mathcal{F}$ is genuinely infinite-dimensional: the combinatorial dimension diverges as $\varepsilon \downarrow 0$. Nonetheless, at every fixed scale the class is finite-dimensional, which is precisely what will be needed to estimate the uniform covering numbers below.

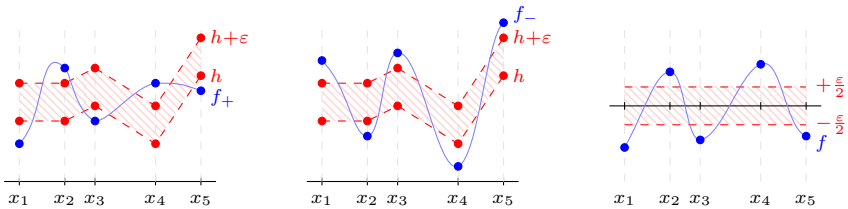
Consider $I = \{x_1, \ldots, x_n\} \subset \mathbb{R}$ with $x_1 < \cdots < x_n$. Suppose that I is ε-shattered by $\mathcal{F}$. Then we can find $h \in \mathbb{R}^I$ and $f_+, f_- \in \mathcal{F}$ such that

$$\begin{aligned} f_+(x_i) &\le h(x_i) \quad \text{for odd } i, & f_+(x_i) &\ge h(x_i) + \varepsilon \quad \text{for even } i, \\ f_-(x_i) &\le h(x_i) \quad \text{for even } i, & f_-(x_i) &\ge h(x_i) + \varepsilon \quad \text{for odd } i. \end{aligned}$$

In particular, $f = \frac{1}{2}\{f_+ - f_-\} \in \mathcal{F}$ satisfies

$$f(x_i) \le -\frac{\varepsilon}{2} \quad \text{for odd } i, \qquad f(x_i) \ge \frac{\varepsilon}{2} \quad \text{for even } i.$$

This construction is illustrated in the following figure.



By construction, we can now estimate

$$(n-1)\varepsilon \le \sum_{k=1}^{n-1} |f(x_{k+1}) - f(x_k)| \le \|f\|_{\mathrm{var}} \le \frac{\|f_+\|_{\mathrm{var}} + \|f_-\|_{\mathrm{var}}}{2} \le V,$$

and thus the cardinality of our shattered set must satisfy $n \le 1 + V/\varepsilon$. As the combinatorial dimension is integer, this evidently implies $\mathrm{vc}(\mathcal{F},\varepsilon) \le 1 + \lfloor V/\varepsilon \rfloor$.

Now let $x_1 < \ldots < x_n$ with $n = 1 + \lfloor V/\varepsilon \rfloor$ be arbitrary. Define

$$f_J(x) = \begin{cases} \varepsilon \mathbf{1}_{x_1 \notin J} & \text{for } x \in \,]-\infty, x_2[, \\ \varepsilon \mathbf{1}_{x_i \notin J} & \text{for } x \in [x_i, x_{i+1}[, \ 1 < i < n, \\ \varepsilon \mathbf{1}_{x_n \notin J} & \text{for } x \in [x_n, \infty[\end{cases}$$

for every $J \subseteq \{x_1, \ldots, x_n\}$. Then $\|f_J\|_{\mathrm{var}} \le (n-1)\varepsilon \le V$, so $f_J \in \mathcal{F}$. Moreover, by construction, $f_J(x_i) = 0$ if $x_i \in J$ and $f_J(x_i) = \varepsilon$ if $x_i \notin J$. Thus *any* set of cardinality n is ε-shattered by $\mathcal{F}$, so we have proved $\mathrm{vc}(\mathcal{F},\varepsilon) = 1 + \lfloor V/\varepsilon \rfloor$.

In view of the above discussion and examples, the combinatorial dimension $\mathrm{vc}(\mathcal{F},\varepsilon)$ is evidently a natural analogue in the general setting of the VC-dimension of a class of sets. However, the real power of this notion lies not in

its definition, but in the fact that it can be used to bound uniform covering numbers in direct analogy to the theory developed in the previous section. This is made precise by the following generalization of Theorem 7.16.

Theorem 7.30 (Mendelson-Vershynin). *Let $\mathcal{F}$ be a class of functions on $\mathbb{X}$ that is uniformly bounded $\sup_{f\in\mathcal{F}} \|f\|_\infty \le 1$. Then we have*

$$\sup_\mu N(\mathcal{F}, \|\cdot\|_{L^2(\mu)}, \varepsilon) \le \left(\frac{K}{\varepsilon}\right)^{K\,\mathrm{vc}(\mathcal{F},\varepsilon/K)} \qquad \text{for all } \varepsilon > 0,$$

where K is a universal constant.

Note that Theorem 7.30 is indeed a generalization of Theorem 7.16: if $\mathcal{F} = \{\mathbf{1}_C : C \in \mathcal{C}\}$, then it is easily seen that $\mathrm{vc}(\mathcal{F}, \varepsilon) = \mathrm{vc}(\mathcal{C})$ for all $\varepsilon < 1$, and thus we recover Theorem 7.16. On the other hand, unlike in the case of sets, Theorem 7.30 can bound the covering numbers of classes of functions with infinite metric dimension: for example, if we consider the class

$$\mathcal{F} = \{f : \mathbb{R} \to [-1, 1] : \|f\|_{\mathrm{var}} \le V\},$$

then Theorem 7.30 yields

$$\sup_\mu N(\mathcal{F}, \|\cdot\|_{L^2(\mu)}, \varepsilon) \le e^{\frac{KV}{\varepsilon}\log\frac{K}{\varepsilon}},$$

so this bound on the covering numbers even grows superexponentially in $1/\varepsilon$ (we will see in the next section that the optimal bound in this example is only exponential in $1/\varepsilon$; however, the above bound suffices in most applications).

We now turn to the proof of Theorem 7.30. The main steps in the proof are precisely the same as in Theorem 7.16. We will first use probabilistic extraction to reduce the original continuous problem to a combinatorial problem; we already phrased the extraction Lemma 7.17 in terms of functions, so that no additional work is needed. Then, we will use a combinatorial principle to resolve the finite problem. The main challenge in the general setting is to prove a counterpart of Pajor's Theorem 7.19 that counts ε-shattered sets (I, h). Let us begin by giving a precise statement of the requisite result.

Definition 7.31 (ε-cube). *A pair (I, h) is called a* ε-cube *in $\mathcal{F}$ if $I \subseteq \mathbb{X}$, $h \in (\varepsilon\mathbb{Z})^I$, and the pair (I, h) is ε-shattered by $\mathcal{F}$.*

Thus an ε-cube is simply an ε-shattered pair (I, h) such that the values of $h(x)$ are integer multiples of ε. The reason for the latter restriction is to ensure that the problem of counting ε-cubes is a combinatorial one: if $|\mathbb{X}| < \infty$ and $\|f\|_\infty \le 1$ for all $f \in \mathcal{F}$, then there are only a finite number of possibilities for I and h. The following result is a form of Pajor's Theorem 7.19 for ε-cubes.

Theorem 7.32. *Let $\mathcal{F}$ be a class of functions and let μ be a probability on $\mathbb{X}$. Then for any $\mathcal{G} \subseteq \mathcal{F}$ that is a $c\varepsilon$-packing of $(\mathcal{F}, \|\cdot\|_{L^2(\mu)})$, we have*

$$|\mathcal{G}|^{1/2} \le |\{(I,h) : (I,h) \text{ is an } \varepsilon\text{-cube}\}|.$$

Here c is a universal constant.

Note that even in the special case of indicator functions, Theorem 7.32 yields a somewhat weaker result than Theorem 7.19. While these two results and their proofs are very much in the same spirit, there is a genuinely new difficulty that arises in the setting of functions that must be overcome by Theorem 7.32 and that accounts for the difference between the two results. To understand the problem, note that for indicator functions $\mathbf{1}_C(x) \neq \mathbf{1}_D(x)$ necessarily implies $\mathbf{1}_C(x) \le 0$ and $\mathbf{1}_D(x) \ge 1$ or vice versa, so a shattered set is automatically 1-shattered. On the other hand, for arbitrary functions $f(x) \neq g(x)$ does not imply $f(x) \le h$ and $g(x) \ge h + \varepsilon$ or vice versa, as is needed in the definition of ε-shattering. In the process of counting ε-shattered sets we will necessarily have to throw out some of the functions in $\mathcal{G}$ that happen to take values in the forbidden regions $[h, h+\varepsilon]$, and the key difficulty in the proof is to ensure that we do not discard too many of these functions. The assumption that $\mathcal{G}$ is a $c\varepsilon$-packing of $(\mathcal{F}, \|\cdot\|_{L^2(\mu)})$ is needed to ensure that we can find coordinates on which there are *many* functions in $\mathcal{G}$ that do not take values in $[h, h+\varepsilon]$. On the other hand, after throwing out the "bad" functions we will only be able to ensure that we have $|\mathcal{G}|^{1/2}$ functions left over, which accounts for the difference between the conclusions of Theorems 7.32 and 7.19. These ideas will be made precise in the proof.

Before proving Theorem 7.32, however, let us first complete the proof of Theorem 7.30 as we now have all the necessary ingredients to do so. We begin by formulating an analogue of the Sauer-Shelah lemma in the present setting.

Corollary 7.33. *Let $\mathcal{F}$ be a class of functions on a finite set $\mathbb{X}$ with $\|f\|_\infty \le 1$ for all $f \in \mathcal{F}$. Then for any probability μ and $c\varepsilon$-packing $\mathcal{G}$ of $(\mathcal{F}, \|\cdot\|_{L^2(\mu)})$*

$$|\mathcal{G}|^{1/2} \le \sum_{k=0}^{\mathrm{vc}(\mathcal{F},\varepsilon)} \binom{|\mathbb{X}|}{k} \left(\frac{2}{\varepsilon}\right)^k \le \left(\frac{2e|\mathbb{X}|}{\varepsilon\,\mathrm{vc}(\mathcal{F},\varepsilon)}\right)^{\mathrm{vc}(\mathcal{F},\varepsilon)}.$$

Proof. If (I,h) is an ε-cube, then $h(x)$ is an integer multiple of ε and we must have $-1 \le h(x) < 1$ as $\|f\|_\infty \le 1$ for all $f \in \mathcal{F}$. Thus, for a given $I \subseteq \mathbb{X}$, there can be at most $(\frac{2}{\varepsilon})^{|I|}$ ε-cubes (I,h). There are consequently at most $\binom{|\mathbb{X}|}{k}(\frac{2}{\varepsilon})^k$ ε-cubes (I,h) with $|I| = k$. By definition, however, any ε-cube (I,h) must have $|I| \le \mathrm{vc}(\mathcal{F},\varepsilon)$. Thus the first inequality follows from Theorem 7.32, while the second inequality follows as in the proof of Lemma 7.12. □

We can now complete the proof of Theorem 7.30.

Proof (Theorem 7.30). Let μ be any probability on $\mathbb{X}$, and let $\mathcal{G} = \{f_1, \ldots, f_m\}$ be a maximal ε-packing of $(\mathcal{F}, \|\cdot\|_{L^2(\mu)})$. By Lemma 7.17, there exist $r < c\varepsilon^{-4} \log m$ points $x_1, \ldots, x_r$ such that $\mathcal{G}$ is an $\varepsilon/2$-packing of $\mu^x = \frac{1}{r}\sum_{k=1}^r \delta_{x_k}$. Using Corollary 7.33 and arguing as in the proof of Theorem 7.16 yields

$$m^{1/2} \le \left(\frac{\log m}{\mathrm{vc}(\mathcal{F},\varepsilon/2c)}\frac{4ec}{\varepsilon^5}\right)^{\mathrm{vc}(\mathcal{F},\varepsilon/2c)} \le m^{1/4}\left(\frac{4(4ec)^{1/5}}{\varepsilon}\right)^{5\,\mathrm{vc}(\mathcal{F},\varepsilon/2c)}.$$

As $N(\mathcal{F},\|\cdot\|_{L^2(\mu)},\varepsilon) \le m$, the proof is readily completed. □

The remainder of this section is devoted to the proof of Theorem 7.32. Let us first recall how we proved the analogous result for classes of sets: first, we introduced a structure, called a splitting tree, to help us count shattered sets. A shattered set corresponds to a *complete* splitting tree, but these are hard to find. Instead, we proved a sort of Ramsey principle: any splitting tree contains at least as many complete subtrees as the number of leaves in the tree. For a class of sets $\mathcal{C}$, it was trivial to construct a splitting tree with $|\mathcal{C}|$ leaves in a greedy fashion, and thus the result followed.

We will follow exactly the same approach in the proof of Theorem 7.32. Let us begin by defining the analogue of a splitting tree in the present setting.

Definition 7.34 (ε-splitting tree). *Let $\mathcal{F}$ be a class of functions on $\mathbb{X}$. A $\mathcal{F}$-tree A is called an* ε-splitting tree *if every $\mathcal{A} \in \mathsf{A}$ that is not a leaf satisfies:*

1. $\mathcal{A}$ has exactly two children $\mathcal{A}_+$ and $\mathcal{A}_-$;

2. There exist $x_{\mathcal{A}} \in \mathbb{X}$ and $h_{\mathcal{A}} \in \varepsilon\mathbb{Z}$ such that

$$f(x_{\mathcal{A}}) \le h_{\mathcal{A}} \quad \text{for } f \in \mathcal{A}_-, \qquad f(x_{\mathcal{A}}) \ge h_{\mathcal{A}} + \varepsilon \quad \text{for } f \in \mathcal{A}_+.$$

In exact analogy to the previous section (cf. Definition 7.21 and the discussion thereafter), an ε-cube corresponds to a complete ε-splitting tree, while any ε-splitting tree contains at least as many complete subtrees as leaves.

Lemma 7.35. *Let $\mathcal{F}$ be a class of functions on $\mathbb{X}$. For any ε-splitting tree A*

$$|\{\text{leaves of } \mathsf{A}\}| \le |\{(I,h) : (I,h) \text{ is an } \varepsilon\text{-cube}\}|.$$

Proof. The proof is identical to that of Lemma 7.22. □

It only remains to construct an ε-splitting tree. While this was trivial in the case of sets, it is here that the difficulties arise in the general setting.

Let us recall in more detail how we constructed a splitting tree for a class $\mathcal{F}$ of indicator functions of sets. Let $\mathcal{A} = \{\mathbf{1}_C : C \in \mathcal{C}\}$ be a class of indicators. Note that for indicator functions, $\mathbf{1}_C \ne \mathbf{1}_D$ necessarily implies that $\mathbf{1}_C(x) = 0$ and $\mathbf{1}_D(x) = 1$, or vice versa, for some $x \in \mathbb{X}$. Therefore, as long as $\mathcal{A}$ is not a singleton, we can *partition* $\mathcal{A}$ into two nonempty sets $\mathcal{A}_+ = \{\mathbf{1}_C \in \mathcal{A} : \mathbf{1}_C(x) = 1\}$ and $\mathcal{A}_- = \{\mathbf{1}_C \in \mathcal{A} : \mathbf{1}_C(x) = 0\}$. We can now repeatedly apply this construction, starting at the root $\mathcal{F}$, until all of the leaves of the resulting tree are singletons. The key point of this construction is that nothing was lost in the process, so the leaves of the tree must form a partition of $\mathcal{F}$. But each leaf is a singleton, so there are $|\mathcal{F}|$ leaves.

Let us now attempt to apply the same idea to a general class of functions $\mathcal{F}$. Consider a set of functions $\mathcal{A} \subseteq \mathcal{F}$ that is not a singleton. Unlike in the case of indicators, $f \neq g$ does not imply that $f(x) \leq h$ and $g(x) \geq h + \varepsilon$, or vice versa, for some $x \in \mathbb{X}$ and $h \in \mathbb{R}$, as is needed for the construction of the children of $\mathcal{A}$. Thus we must assume *some* form of separation between the elements of $\mathcal{A}$. The minimal assumption we could impose is that $\mathcal{A}$ is an ε-packing of $(\mathcal{F}, \|\cdot\|_\infty)$: this would ensure that $\|f - g\|_\infty \geq \varepsilon$, and thus the above conclusion would follow. Therefore, if we introduce this assumption, then both $\mathcal{A}_+ = \{f \in \mathcal{A} : f(x) \leq h\}$ and $\mathcal{A}_- = \{f \in \mathcal{A} : f(x) \geq h + \varepsilon\}$ are nonempty and satisfy the definition of an ε-splitting tree. However, $\mathcal{A}_+$ and $\mathcal{A}_-$ no longer form a partition of $\mathcal{A}$: it is very likely that some of the functions in $\mathcal{A}$ happen to take values in the "forbidden" region $[h, h + \varepsilon]$, and these functions must be thrown out in the construction of the tree. The key problem that we face is that we do not know *how many* functions we throw out, and thus we have no control over the number of leaves in the tree.

To surmount this problem, it is essential to find a coordinate x and level h at which we can split the set $\mathcal{A}$ without discarding too many functions. This is precisely the content of the following result. The price we pay is that the assumption that $\mathcal{A}$ is a packing in $(\mathcal{F}, \|\cdot\|_\infty)$ is too weak to make this happen: we need the stronger assumption that $\mathcal{A}$ is a packing in $(\mathcal{F}, \|\cdot\|_{L^2(\mu)})$.

Proposition 7.36 (Controlled splitting). *Let $\mathcal{F}$ be a class of functions and μ be a probability on $\mathbb{X}$. Let $\mathcal{A}$ be a $c\varepsilon$-packing of $(\mathcal{F}, \|\cdot\|_{L^2(\mu)})$ with $|\mathcal{A}| \geq 2$. Then there exist $x \in \mathbb{X}$ and $h \in \varepsilon\mathbb{Z}$ such that the sets*

$$\mathcal{A}_- = \{f \in \mathcal{A} : f(x) \leq h\}, \qquad \mathcal{A}_+ = \{f \in \mathcal{A} : f(x) \geq h + \varepsilon\}$$

satisfy $|\mathcal{A}_+|^{1/2} + |\mathcal{A}_-|^{1/2} > |\mathcal{A}|^{1/2}$.

Proof. The idea is quite simple. Let us choose two *random* elements $a, a' \in \mathcal{A}$ drawn uniformly and independently. By assumption $\|a - a'\|_{L^2(\mu)} \geq c\varepsilon$ as long as $a \neq a'$, which happens with probability $1 - \frac{1}{|\mathcal{A}|} \geq \frac{1}{2}$. Thus

$$\frac{c^2\varepsilon^2}{2} \leq \left(1 - \frac{1}{|\mathcal{A}|}\right) c^2\varepsilon^2 \leq \mathbf{E}\|a - a'\|^2_{L^2(\mu)} = \int \mathbf{E}|a(x) - a'(x)|^2 \, \mu(dx).$$

Thus we can certainly choose $x \in \mathbb{X}$ such that

$$\frac{c^2\varepsilon^2}{2} \leq \mathbf{E}|a(x) - a'(x)|^2 = 2\,\mathrm{Var}[a(x)].$$

We now want to find $h \in \varepsilon\mathbb{Z}$ such that

$$\mathbf{P}[a(x) \leq h]^{1/2} + \mathbf{P}[a(x) \geq h + \varepsilon]^{1/2} > 1.$$

Indeed, as we have $\mathbf{P}[a(x) \leq h] = \frac{|\mathcal{A}_-|}{|\mathcal{A}|}$ and $\mathbf{P}[a(x) \geq h + \varepsilon] = \frac{|\mathcal{A}_+|}{|\mathcal{A}|}$, the proof would evidently be complete once we can find such an h.

At this point, it seems the proof should reduce to a general probabilistic principle: if $\mathrm{Var}[X] \geq C^2\varepsilon^2$ for $C \gg 1$, then it should not be possible that most of the probability mass of X is concentrated in an interval of size $\leq \varepsilon$. This is precisely the statement of the following result to be proved below.

Lemma 7.37. *There is a universal constant C such that if $\mathrm{Var}[X] \geq C^2\varepsilon^2$, then there exists $b \in \mathbb{R}$ such that $\mathbf{P}[X \leq b]^{1/2} + \mathbf{P}[X \geq b + \varepsilon]^{1/2} > 1$.*

The only remaining issue is that Lemma 7.37 yields $b \in \mathbb{R}$, while we need $h \in \varepsilon\mathbb{Z}$. This is easily resolved, however. Choose the universal constant $c = 4C$. As we have $\mathrm{Var}[a(x)] \geq C^2(2\varepsilon)^2$, Lemma 7.37 yields $b \in \mathbb{R}$ such that

$$\mathbf{P}[a(x) \leq b]^{1/2} + \mathbf{P}[a(x) \geq b + 2\varepsilon]^{1/2} > 1.$$

Now choose h to be the value of b rounded upwards to the nearest multiple of ε. Then $b \leq h \leq b + \varepsilon$, and the proof is readily completed. □

It remains to prove the small deviation principle used above.

Proof (Lemma 7.37). We prove the contrapositive. Suppose the conclusion fails, that is, that $\mathbf{P}[X \leq b]^{1/2} + \mathbf{P}[X \geq b + \varepsilon]^{1/2} \leq 1$ for all $b \in \mathbb{R}$. Then

$$\mathbf{P}[X > b + \varepsilon] \leq \mathbf{P}[X > b]^2, \quad \mathbf{P}[X < b] \leq \mathbf{P}[X < b + \varepsilon]^2 \quad \text{for all } b \in \mathbb{R},$$

where we used $\mathbf{P}[X \leq b] \leq \mathbf{P}[X \leq b]^{1/2}$ ($\mathbf{P}[X \geq b + \varepsilon] \leq \mathbf{P}[X \geq b + \varepsilon]^{1/2}$) in the first (second) inequality. Let $M = \mathrm{med}(X)$ be the median of X. Iterating these inequalities starting from $\mathbf{P}[X > M] \leq \frac{1}{2}$ ($\mathbf{P}[X < M] \leq \frac{1}{2}$) yields

$$\mathbf{P}[X > M + k\varepsilon] \leq 2^{-2^k}, \quad \mathbf{P}[X < M - k\varepsilon] \leq 2^{-2^k} \quad \text{for all } k \in \mathbb{N}.$$

Thus the random variable X has *very* thin tail probabilties. But a random variable with thin tails certainly cannot have large variance: to be precise,

$$\mathrm{Var}[X] \leq \mathbf{E}[(X - M)^2] = \sum_{k=0}^{\infty} \int_{k\varepsilon}^{(k+1)\varepsilon} 2t\, \mathbf{P}[|X - M| > t]\, dt < C^2\varepsilon^2$$

with $C^2 = \sum_{k=0}^{\infty} 4(k+1)2^{-2^k}$. Thus the contrapositive is proved. □

With Proposition 7.36 in hand, we can now construct a large ε-splitting tree in a greedy fashion in the same manner as we did in the case of sets.

Corollary 7.38. *Let $\mathcal{F}$ be a class of functions and μ be a probability on $\mathbb{X}$. Let $\mathcal{G}$ be a $c\varepsilon$-packing of $(\mathcal{F}, \|\cdot\|_{L^2(\mu)})$. There exists a ε-splitting tree A with*

$$|\{\textit{leaves of } \mathsf{A}\}| \geq |\mathcal{G}|^{1/2}.$$

Proof. Grow the ε-splitting tree A by starting with $\mathcal{G}$ as the root and repeatedly splitting the leaves of the tree into two subsets using Proposition 7.36 until all leaves are singletons. By construction, we have $|\mathcal{A}_+|^{1/2} + |\mathcal{A}_-|^{1/2} > |\mathcal{A}|^{1/2}$ for every $\mathcal{A} \in \mathsf{A}$. Iterating this bound starting at the root gives

$$|\mathcal{G}|^{1/2} < \sum_{\mathcal{A} \text{ is a leaf}} |\mathcal{A}|^{1/2} = |\{\text{leaves in } \mathsf{A}\}|,$$

and the proof is complete. □

Combining Lemma 7.35 and Corollary 7.38 yields Theorem 7.32.

Remark 7.39. There is nothing special about the power $|\mathcal{G}|^{1/2}$ in Theorem 7.32: the statement remains valid if $|\mathcal{G}|^{1/2}$ is replaced by $|\mathcal{G}|^{1-\alpha}$ for any $0 < \alpha < 1$ at the expense of changing the value of the universal constant c. To see this, note that the origin of the power $\frac{1}{2}$ is in Lemma 7.37, where the precise value of the power is however entirely irrelevant in the proof. We have stated the above results in terms of $|\mathcal{G}|^{1/2}$ merely to avoid notational distractions (the value of the power ultimately affects only the constants in Theorem 7.30).

Problems

7.12 (VC-subgraph classes and pseudodimension). There is a simple method to extend the bound of Theorem 7.16 for classes of sets to classes of functions without introducing the notion of combinatorial dimension. Given a class of functions $\mathcal{F}$ on a set $\mathbb{X}$, define an associated class of sets $\mathcal{C}_\mathcal{F}$ as

$$\mathcal{C}_\mathcal{F} := \{C \subseteq \mathbb{X} \times \mathbb{R} : C = \{(x,t) : t < f(x)\},\ f \in \mathcal{F}\}.$$

That is, $\mathcal{C}_\mathcal{F}$ is the class of subgraphs of functions in $\mathcal{F}$. We now define the *pseudodimension* $\mathrm{vc}(\mathcal{F})$ as the VC-dimension $\mathrm{vc}(\mathcal{C}_\mathcal{F})$ of the subgraphs.

a. Deduce directly from Theorem 7.16 that if $\mathcal{F}$ is a class of functions such that $\|f\|_\infty \le 1$ for all $f \in \mathcal{F}$, then there is a universal constant K such that

$$\sup_\mu N(\mathcal{C}, \|\cdot\|_{L^2(\mu)}, \varepsilon) \le \left(\frac{K}{\varepsilon}\right)^{K\,\mathrm{vc}(\mathcal{F})} \qquad \text{for all } \varepsilon < 1.$$

Hint: consider $(\mathcal{C}_\mathcal{F}, \|\cdot\|_{L^2(\mu\otimes\lambda)})$ with λ the uniform distribution on $[-1,1]$.

b. Show that the linear class $\mathcal{F}$ in Example 7.28 satisfies $\mathrm{vc}(\mathcal{F}) < \infty$, but that the bounded variation class in Example 7.29 satisfies $\mathrm{vc}(\mathcal{F}) = \infty$.

At first sight, pseudodimension and combinatorial dimension seem to yield two distinct methods to bound the uniform covering numbers of function classes. However, this is not the case: the result of part a. is none other than a special case of Theorem 7.30 for classes of finite metric dimension.

c. Show that $\mathrm{vc}(\mathcal{F}) = \sup_{\varepsilon>0} \mathrm{vc}(\mathcal{F},\varepsilon)$, and conclude that the result of part a. follows as a special case of Theorem 7.30.

7.13 (Combinatorial dimension of convex classes). The notion of combinatorial dimension is designed to be meaningful for any class of functions $\mathcal{F}$. If we assume that the class $\mathcal{F}$ is *convex*, however, the combinatorial dimension can be given a simple geometric interpretation: $(\mathcal{F},\varepsilon)$ is the largest dimension of a cube of side length ε that is contained in a coordinate projection of $\mathcal{F}$.

a. Suppose that $\mathcal{F}$ is convex. Show that

$$(I,h) \text{ is } \varepsilon\text{-shattered} \quad \text{if and only if} \quad \mathcal{F}|_I \supseteq h + [0,\varepsilon]^I.$$

Hint: assume the conclusion is false; use the separating hyperplane theorem and reason as in Example 7.28 to generate a contradiction.

b. Suppose that $\mathcal{F}$ is convex and symmetric. Show that

$$I \text{ is } \varepsilon\text{-shattered} \quad \text{if and only if} \quad \mathcal{F}|_I \supseteq [-\tfrac{\varepsilon}{2},\tfrac{\varepsilon}{2}]^I.$$

Hint: reason as in Example 7.29.

If $\mathcal{F}$ is not convex, one might expect that (I,h) is ε-shattered if and only if the convex hull of $\mathcal{F}$ contains a cube $\mathrm{conv}\,\mathcal{F}|_I \supseteq h + [0,\varepsilon]^I$. This is not true, however: $\mathrm{conv}\,\mathcal{F}$ can have many more shattered sets than $\mathcal{F}$ itself.

c. Let $\mathcal{F} = \{\mathbf{1}_{\{i\}} : i \in \mathbb{N}\}$ be a class of indicator functions on $\mathbb{N}$. Show that $\mathrm{vc}(\mathcal{F},\varepsilon) = 1$ for all $\varepsilon < 1$, but that $\mathrm{vc}(\mathrm{conv}\,\mathcal{F},\varepsilon)$ diverges as $\varepsilon \downarrow 0$. Thus the convex hull of a finite-dimensional class can even be infinite-dimensional.

This example raises a basic question: when $\mathcal{F}$ is not convex, what can be said about the combinatorial dimension of the convex hull $\mathrm{vc}(\mathrm{conv}\,\mathcal{F},\varepsilon)$ in terms of $\mathrm{vc}(\mathcal{F},\varepsilon)$? Surprisingly, Theorem 7.30 can help us address this question.

d. If $\{x_1,\ldots,x_n\} \subseteq \mathbb{X}$ is ε-shattered by $\mathcal{F}$ and $g_1,\ldots,g_n \sim$ i.i.d. $N(0,1)$, prove

$$\ell_I(\mathcal{F}) := \mathbf{E}\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^n g_i f(x_i)\right] \gtrsim \varepsilon n.$$

Hint: replace $f(x_i)$ by $f(x_i) - h_i - \frac{\varepsilon}{2}$ in the definition of $\ell_I(\mathcal{F})$, and choose the functions f to cancel the signs of the Gaussian variables g_i.

e. Suppose that $\|f\|_\infty \le 1$ for all $f \in \mathcal{F}$. Show that for any $\delta > 0$

$$\begin{aligned}\ell_I(\mathcal{F}) &\lesssim n\delta + \sqrt{n}\int_\delta^2 \sqrt{K\,\mathrm{vc}(\mathcal{F},t/K)\log(K/t)}\,dt \\ &\lesssim n\delta + \sqrt{n\,\mathrm{vc}(\mathcal{F},\delta/K)}.\end{aligned}$$

Hint: recall Theorem 5.31.

f. Let $\mathcal{F}$ be a class of functions such that $\|f\|_\infty \le 1$ for all $f \in \mathcal{F}$. Show that

$$\mathrm{vc}(\mathrm{conv}\,\mathcal{F}, L\varepsilon) \le \frac{\mathrm{vc}(\mathcal{F},\varepsilon)}{\varepsilon^2} \qquad \text{for all } \varepsilon > 0,$$

where L is a universal constant.

Hint: show that $\ell_I(\mathcal{F}) = \ell_I(\mathrm{conv}\,\mathcal{F})$ and combine the previous two parts.

7.14 (Elton's theorem). The notion of combinatorial dimension has its origin not in probability theory but in geometric functional analysis. Let us use the machinery we have developed to prove a classic result in this direction.

Let $(B, \|\cdot\|_B)$ be a Banach space. We are interested in the question whether the finite-dimensional Banach space ℓ_1^n embeds into B: that is, whether one can find vectors $x_1, \ldots, x_n \in B$ whose linear span is isomorphic to ℓ_1^n in the sense that there exist constants C_1, C_2 (independent of n) such that

$$C_1 \sum_{i=1}^n |a_i| \le \left\| \sum_{i=1}^n a_i x_i \right\|_B \le C_2 \sum_{i=1}^n |a_i| \quad \text{for all } a \in \mathbb{R}^n.$$

The upper bound is trivial: if we choose any $x_1, \ldots, x_n$ in the unit ball of B (i.e., $\|x_i\|_B \le 1$) then the upper bound holds for $C_2 = 1$ by the triangle inequality. The difficulty is to understand what spaces B admit a lower bound.

If the lower bound holds, then we obtain as a special case that

$$\| \pm x_1 \pm \cdots \pm x_n \|_B \ge C_1 n$$

for all possible choices of signs; when this is the case, we say that ℓ_1^n *sign-embeds* into B. The converse is far from clear, however: if ℓ_1^n sign-embeds into B, does this already imply a full embedding as defined above?

Elton's theorem provides an answer to this question. In fact, Elton only makes the weaker assumption that the sign-embedding holds "on average" in the sense that there exist $x_1, \ldots, x_n$ in the unit ball of B and $\delta > 0$ such that

$$\mathbf{E}\left\| \sum_{i=1}^n \varepsilon_i x_i \right\|_B \ge \delta n,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. symmetric Bernoulli variables (random signs). Under this assumption, we will prove the following quantitative form of Elton's theorem: there is a subset $I \subseteq \{1, \ldots, n\}$ of cardinality $|I| \ge c\delta^2 n$ such that

$$c\delta \sum_{i \in I} |a_i| \le \left\| \sum_{i \in I} a_i x_i \right\|_B \le \sum_{i \in I} |a_i| \quad \text{for all } a \in \mathbb{R}^n,$$

where c is a universal constant. Thus the existence of a random sign-embedding of ℓ_1^n with dimension n and constant δ implies the existence of an embedding of $\ell_1^{n'}$ with dimension $n' \gtrsim n$ and constant $\gtrsim \delta$.

a. Let B_1^* be the unit ball in the dual space of B, and define

$$\mathcal{F} = \{f : \{x_1, \ldots, x_n\} \to \mathbb{R} : f(x) = \langle y, x \rangle,\ y \in B_1^*\}.$$

Show that $\{x_i : i \in I\}$ is 2ε-shattered by $\mathcal{F}$ if and only if

$$\left\| \sum_{i \in I} a_i x_i \right\|_B = \sup_{f \in \mathcal{F}} \sum_{i \in I} a_i f(x_i) \geq \varepsilon \sum_{i \in I} |a_i| \quad \text{for all } a \in \mathbb{R}^n.$$

Hint: use the ideas from the first part of Problem 7.13.

b. Show that for all $\varepsilon > 0$

$$\mathbf{E} \left\| \sum_{i=1}^n \varepsilon_i x_i \right\|_B \lesssim n\varepsilon + \sqrt{n \operatorname{vc}(\mathcal{F}, \varepsilon/K)}.$$

Hint: argue as in the second part of Problem 7.13.

c. Complete the proof of Elton's theorem in the form stated above.

7.4 The iteration method

We have developed in the previous section a powerful combinatorial bound on the uniform covering numbers of classes of functions. This bound suffices in many cases to obtain distribution-free control of the supremum of empirical processes. It is of significant interest, however, to understand how sharp such bounds are in general: does combinatorial dimension capture completely the size of the uniform covering numbers? To gain some insight into this question, let us begin by developing a simple lower bound.

Lemma 7.40. *Let $\mathcal{F}$ be a class of functions on $\mathbb{X}$ that is uniformly bounded $\sup_{f \in \mathcal{F}} \|f\|_\infty \leq 1$. Then for universal constants C, c and all $\varepsilon > 0$*

$$\tfrac{1}{8} \operatorname{vc}(\mathcal{F}, 4\varepsilon) \leq \log \sup_\mu N(\mathcal{F}, \|\cdot\|_{L^2(\mu)}, \varepsilon) \leq C \operatorname{vc}(\mathcal{F}, c\varepsilon) \log\left(\frac{C}{\varepsilon}\right).$$

Proof. The upper bound is Theorem 7.30. To prove the lower bound, let (I, h) be a 4ε-shattered pair with $|I| = \operatorname{vc}(\mathcal{F}, 4\varepsilon)$, and let μ be the uniform distribution on I. The proof follows once we show $\log N(\mathcal{F}, \|\cdot\|_{L^2(\mu)}, \varepsilon) \geq |I|/8$.

The establish this claim, choose for every $J \subseteq I$ a function $f_J \in \mathcal{F}$ such that $f_J(x) \leq h(x)$ for $x \in J$ and $f_J(x) \geq h(x) + 4\varepsilon$ for $x \in I \backslash J$. Then $\|f_J - f_{J'}\|_{L^2(\mu)} \geq 4\varepsilon \sqrt{|I|^{-1} |J \triangle J'|}$ for every $J, J' \subseteq I$. By Lemma 7.24, there exists a family $\mathcal{J}$ of subsets of I with $|\mathcal{J}| \geq e^{|I|/8}$ such that $|J \triangle J'| \geq |I|/4$ for every $J, J' \in \mathcal{J}$, $J \neq J'$. Then $\{f_J : J \in \mathcal{J}\}$ is evidently a 2ε-packing of $\mathcal{F}$, and the claim follows by the duality between packing and covering. □

Lemma 7.40 suggests that our combinatorial bounds are not far from being sharp: up to universal constants, the lower and upper bounds in Lemma 7.40 differ only by a logarithmic factor $\sim \log(1/\varepsilon)$. The immediate question that arises at this point is whether we can close the gap between the upper and lower bounds: perhaps an improved upper bound can eliminate the logarithmic factor, or perhaps an improved lower bound can add an additional logarithmic factor? Unfortunately, no improvement of this kind is possible: the logarithmic factor is sharp for some classes $\mathcal{F}$ but not for others.

Example 7.41. Let $\mathbb{X} = \mathbb{N}$ and $\mathcal{F} = \{\mathbf{1}_{\{i\}} : i \in \mathbb{N}\}$. Then $\mathrm{vc}(\mathcal{F}, \varepsilon) = 1$ for all $0 < \varepsilon \le 1$. On the other hand, if μ is the uniform distribution on $\mathbb{N} \cap [1, 1/8\varepsilon^2]$, then we have $\|\mathbf{1}_{\{i\}} - \mathbf{1}_{\{j\}}\|_{L^2(\mu)} > 2\varepsilon$ for all $i, j \le \lceil 1/8\varepsilon^2 \rceil$, $i \ne j$, which implies $\log N(\mathcal{F}, \|\cdot\|_{L^2(\mu)}, \varepsilon) \gtrsim \log(1/\varepsilon)$ by duality of packing and covering. Thus in this case the logarithmic factor in the upper bound of Lemma 7.40 is sharp.

Example 7.42. Let $\mathbb{X} = [0, 1]$ and $\mathcal{F} = \{f \in \mathrm{Lip}(\mathbb{X}) : 0 \le f \le 1\}$. It is easily shown as in Example 7.29 that $\mathrm{vc}(\mathcal{F}, \varepsilon) = 1 + \lfloor 1/\varepsilon \rfloor$ for all $0 < \varepsilon \le 1$ (the upper bound follows immediately from Example 7.29; for the lower bound, repeating the proof in Example 7.29 with piecewise linear functions f_J shows that $I = \{k\varepsilon : 0 \le k \le \lfloor 1/\varepsilon \rfloor\}$ is ε-shattered). On the other hand, we have proved in Lemma 5.16 that $\log N(\mathcal{F}, \|\cdot\|_{L^2(\mu)}, \varepsilon) \lesssim 1/\varepsilon$ for every probability measure μ. Thus in this case the lower bound in Lemma 7.40 is sharp, while the upper bound contains an unnecessary logarithmic factor.

For what classes must the logarithmic factor to appear and when it is unnecessary? In the remainder of this section, we will develop a method that will make it possible in many cases to resolve the mystery of the logarithmic factor. In concrete applications this will often not yield a major improvement: the logarithmic factor tends to be innocuous except in borderline cases. Nonetheless, a better understanding of uniform covering bounds can lead to sharper results in certain problems, and deepens our fundamental understanding of the connections between covering numbers and combinatorial dimension. More importantly, the *iteration method* that we will develop for this purpose is of significant interest in its own right, and can be used to great effect in many other problems (see, for example, Problem 7.17 below).

In order to understand how one might eliminate the logarithmic factor, let us begin with an elementary observation. While this might not be entirely obvious at first sight, the bound of Theorem 7.30 depends on two distinct scales: on the one hand, we are covering the class $\mathcal{F}$ by balls of radius ε; on the other hand, we have assumed that the class $\mathcal{F}$ is itself uniformly bounded by $\sup_{f \in \mathcal{F}} \|f\|_\infty \le 1$. If we were to assume instead that $\sup_{f \in \mathcal{F}} \|f\|_\infty \le a$, then applying Theorem 7.30 to the scaled class $\mathcal{F}/a$ readily yields

$$\log N(\mathcal{F}, \|\cdot\|_{L^2(\mu)}, \varepsilon) \le C\, \mathrm{vc}(\mathcal{F}, c\varepsilon) \log\left(\frac{Ca}{\varepsilon}\right)$$

for every $\varepsilon > 0$ and every probability measure μ. Thus the logarithmic factor in Lemma 7.40 does not depend on ε, but rather on the ratio ε/a between the

scale of the cover and the size of the class $\mathcal{F}$. The logarithmic factor would disappear entirely if $a \lesssim \varepsilon$, but this is not adequate: the size of the class $\mathcal{F}$ is fixed, while we are interested in the behavior of the covering numbers as $\varepsilon \downarrow 0$. Nonetheless, we will be able to exploit the fact the we have better covering number bounds for classes with controlled size to systematically improve our covering number bounds for arbitrary classes. This is the idea behind the iteration method, which we develop presently in a general setting.

Let (T, d) be a metric space, and suppose that can bound the covering number of any ball $B(t, 2\varepsilon)$ of radius 2ε by balls of radius ε as follows:

$$\log N(T \cap B(t, 2\varepsilon), d, \varepsilon) \le \varphi(\varepsilon).$$

We would like to obtain a bound on the covering number $N(T, d, \varepsilon)$ of the entire set T. To this end, let us first cover T by $N(T, d, 2\varepsilon)$ balls of radius 2ε, and then cover each of these balls by balls of radius ε. Then evidently the union of the latter balls is a cover of T by balls of radius ε, and there are at most $e^{\varphi(\varepsilon)} N(T, d, 2\varepsilon)$ such balls. We have therefore shown that

$$\log N(T, d, \varepsilon) \le \varphi(\varepsilon) + \log N(T, d, 2\varepsilon).$$

We can now iterate this bound to obtain

$$\log N(T, d, \varepsilon) \le \sum_{k=0}^{\infty} \varphi(2^k \varepsilon)$$

(note that if T has finite diameter, then $\log N(T, d, 2^k \varepsilon) = 0$ for k sufficiently large and the remainder term in the iteration vanishes; while if T has infinite diameter, then $\varphi(\varepsilon) \ge \log 2$ for all $\varepsilon > 0$ and the inequality holds trivially).

Despite its simplicity, this procedure already explains the difference between Examples 7.41 and 7.42. Let us assume for the moment that we can apply the above iteration method with $\varphi(\varepsilon) \lesssim \mathrm{vc}(\mathcal{F}, c\varepsilon)$ (this is not entirely obvious at this point, but this idea will be made precise in the remainder of this section). In Example 7.42, we have $\varphi(\varepsilon) \lesssim 1/\varepsilon$, so

$$\log N(T, d, \varepsilon) \lesssim \frac{1}{\varepsilon} \sum_{k=0}^{\infty} 2^{-k} \lesssim \frac{1}{\varepsilon}.$$

Thus we have eliminated the logarithmic term in Lemma 7.40! On the other hand, in Example 7.41 we have $\varphi(\varepsilon) \lesssim 1$ and $\mathrm{vc}(\mathcal{F}, \varepsilon) = 0$ for $\varepsilon > 1$, so that

$$\log N(T, d, \varepsilon) \le \sum_{k=0}^{\log(1/c\varepsilon)} \varphi(2^k \varepsilon) \lesssim \log\left(\frac{1}{c\varepsilon}\right).$$

Thus in this case the logarithmic term in Lemma 7.40 remains in place. This computation explains much of the mystery of the logarithmic term: the lower bound in Lemma 7.40 is sharp for infinite-dimensional classes for which the combinatorial dimension $\mathrm{vc}(\mathcal{F}, \varepsilon)$ is at least polynomial in $1/\varepsilon$, while the upper bound is sharp for finite-dimensional classes when $\mathrm{vc}(\mathcal{F}, \varepsilon)$ is constant.

Remark 7.43. The iteration method should be understood as the direct analogue for covering numbers of the chaining method. In the chaining method, we aim to obtain a bound on the supremum of a general random process on T starting from such a bound for the special case where the cardinality $|T|$ is controlled. To this end, we approximate the supremum of a general process by the supremum over a finite set plus a remainder term that is of the same form as the original supremum, and iterate this bound until the remainder term is eliminated. In a completely analogous manner, the iteration method allows to obtain a bound on the covering number of the set T starting from such a bound for the special case where the diameter of T is controlled. Even if we can directly estimate $N(T,d,\varepsilon)$ as in Lemma 7.40, iteration systematically improves this bound by exploiting the control on the diameter at each scale.

The above discussion contains the key idea that will be developed in the sequel. Unfortunately, we cannot immediately apply the above computation to obtain bounds in terms of combinatorial dimension. In order to apply the simple iteration method developed above, we would require that

$$\log N(\mathcal{F},\|\cdot\|_{L^2(\mu)},\varepsilon) \le C\,\mathrm{vc}(\mathcal{F},c\varepsilon)\log\left(\frac{Ca}{\varepsilon}\right)$$

for all $\varepsilon>0$ whenever $\sup_{f\in\mathcal{F}}\|f\|_{L^2(\mu)}\le a$. However, we have only proved such a bound when $\sup_{f\in\mathcal{F}}\|f\|_\infty\le a$, which does not suffice. Indeed, using the latter bound, the first step of the iteration method would yield

$$\log N(\mathcal{F},\|\cdot\|_{L^2(\mu)},\varepsilon) \le C\log(2C)\,\mathrm{vc}(\mathcal{F},c\varepsilon) + \log N(\mathcal{F},\|\cdot\|_\infty,2\varepsilon),$$

but then no control of the remainder term is possible as the L^∞-covering numbers are generally infinite (as is the case, for example, for classes of sets). On the other hand, we did not use the uniform bound $\sup_{f\in\mathcal{F}}\|f\|_\infty\le a$ in the proof of Theorem 7.30 in a very sharp manner, so that one might hope that an improvement of the proof would show that the conclusion of Theorem 7.30 remains valid under the assumption $\sup_{f\in\mathcal{F}}\|f\|_{L^2(\mu)}\le a$. Unfortunately, this also cannot be the case, as the following simple example demonstrates.

Example 7.44. Let $\mathbb{X}=[0,1]$ and let μ be the uniform distribution on $\mathbb{X}$. Let

$$\mathcal{F}_a = \{\mathbf{1}_{[a,b]} : \|\mathbf{1}_{[a,b]}\|_{L^2(\mu)}\le a\}.$$

It is a trivial exercise to show that $\mathrm{vc}(\mathcal{F}_a,\varepsilon)=2$ for all $0<\varepsilon\le 1$.

On the other hand, let $C_k=[(k-1)\varepsilon^2,k\varepsilon^2]$. As $\|\mathbf{1}_{C_k}\|_{L^2(\mu)}=\varepsilon$ and $\|\mathbf{1}_{C_k}-\mathbf{1}_{C_l}\|_{L^2(\mu)}=2^{1/2}\varepsilon$ for all $1\le k,l\le\lfloor\varepsilon^{-2}\rfloor$, $k\ne l$, we can estimate

$$N(\mathcal{F}_\varepsilon,\|\cdot\|_{L^2(\mu)},2^{-1/2}\varepsilon)\ge\lfloor\varepsilon^{-2}\rfloor$$

by the duality of covering and packing. Thus it is not possible to replace the assumption $\sup_f\|f\|_\infty\le 1$ by $\sup_f\|f\|_{L^2(\mu)}\le 1$ in Theorem 7.30, as that would imply that $N(\mathcal{F}_\varepsilon,\|\cdot\|_{L^2(\mu)},2^{-1/2}\varepsilon)$ can be bounded uniformly in ε.

Despite this discouraging example, things are not quite as bad as they seem. While it is not possible to replace $\sup_f \|f\|_\infty \le 1$ by $\sup_f \|f\|_{L^2(\mu)} \le 1$ in Theorem 7.30, we will show that a significant improvement is still possible: it suffices to assume $\sup_f \|f\|_{L^p(\mu)} \le 1$ for any $p > 2$! In fact, we will prove a more general result that is essential for implementing the iteration method.

Theorem 7.45 (Rudelson-Vershynin). *Let $\mathcal{F}$ be a class of functions on $\mathbb{X}$ and let $p \ge 2$. Suppose that $\sup_{f\in\mathcal{F}} \|f\|_{L^{2p}(\mu)} \le a$ for some probability μ. Then*

$$\log N(\mathcal{F}, \|\cdot\|_{L^p(\mu)}, \varepsilon) \le Cp^2 \,\mathrm{vc}(\mathcal{F}, c\varepsilon) \log\left(\frac{a}{c\varepsilon}\right) \quad \text{for all } 0 < \varepsilon < a,$$

where C, c are universal constants.

Remark 7.46. There is nothing special about the bound $\sup_f \|f\|_{L^{2p}(\mu)} \le a$: the same proof will go through if $\sup_f \|f\|_{L^{\beta p}(\mu)} \le a$ for any $\beta > 1$, provided that we replace the constants C, c by $C_\beta = C\beta/(\beta-1)$ and $c_\beta = c(\beta-1) \wedge c$, cf. Problem 7.15. As we will only need to apply this result for a fixed value of β, however, we have fixed $\beta = 2$ above for notational convenience.

Theorem 7.45 is all we need to apply the iteration method. The idea is exactly the same as in the simple iteration method discussed above: the only new feature is that we must use a different L^p-norm in every stage of the iteration in order to eliminate the logarithmic factor. Before we turn to the proof of Theorem 7.45, let us explore the consequences of this idea.

Corollary 7.47 (Iteration). *Let $\mathcal{F}$ be a class of functions on $\mathbb{X}$. Then*

$$\log \sup_\mu N(\mathcal{F}, \|\cdot\|_{L^2(\mu)}, \varepsilon) \le 4C \log(\alpha/c) \sum_{k=0}^{\infty} 4^k \,\mathrm{vc}(\mathcal{F}, c\alpha^k\varepsilon)$$

for any $\alpha > 1$, where C, c are universal constants.

Proof. Fix a probability measure μ, and let $p \ge 1$ and $\varepsilon > 0$. Define $B_p(f,\varepsilon) = \{g : \|g - f\|_{L^p(\mu)} \le \varepsilon\}$. By covering $\mathcal{F}$ by L^{2p}-balls of radius $\alpha\varepsilon$, and then covering each of these balls by L^p-balls of radius ε, we can estimate

$$N(\mathcal{F}, \|\cdot\|_{L^p(\mu)}, \varepsilon) \le \sup_{f\in\mathcal{F}} N(\mathcal{F} \cap B_{2p}(f, \alpha\varepsilon), \|\cdot\|_{L^p(\mu)}, \varepsilon)\, N(\mathcal{F}, \|\cdot\|_{L^{2p}(\mu)}, \alpha\varepsilon).$$

Applying Theorem 7.45 to $\{\mathcal{F} - f\} \cap B_{2p}(0, \alpha\varepsilon)$ yields

$$\log N(\mathcal{F}, \|\cdot\|_{L^p(\mu)}, \varepsilon) \le C \log(\alpha/c)\, p^2 \,\mathrm{vc}(\mathcal{F}, c\varepsilon) + \log N(\mathcal{F}, \|\cdot\|_{L^{2p}(\mu)}, \alpha\varepsilon).$$

Iterating this bound starting at $p = 2$ readily yields the result, provided that the remainder term $\log N(\mathcal{F}, \|\cdot\|_{L^{2^{n+1}}(\mu)}, \alpha^n\varepsilon)$ vanishes as $n \to \infty$.

To see this, note that if $\sup_{f,g\in\mathcal{F}} \|f - g\|_\infty = \infty$, then $\mathrm{vc}(\mathcal{F}, \varepsilon) \ge 1$ for all $\varepsilon > 0$ and thus the iteration bound holds trivially. On ther other hand, if $\sup_{f,g\in\mathcal{F}} \|f - g\|_\infty < \infty$, then $N(\mathcal{F}, \|\cdot\|_{L^{2^{n+1}}(\mu)}, \alpha^n\varepsilon) \le N(\mathcal{F}, \|\cdot\|_\infty, \alpha^n\varepsilon) = 1$ for all n sufficiently large and thus the remainder term converges to zero. □

Using Corollary 7.47, we can readily understand when the lower bound in Lemma 7.40 is sharp: this is always the case for classes whose combinatorial dimension is at least polynomial. This yields a sharp bound, up to universal constants, for most infinite-dimensional classes of practical interest.

Corollary 7.48 (Infinite-dimensional classes). *Let $\mathcal{F}$ be a class of functions on $\mathbb{X}$. Suppose there is a function $\xi : \mathbb{R}_+ \to \mathbb{R}_+$ and $\alpha > 1$ such that*

$$\mathrm{vc}(\mathcal{F}, \varepsilon) \le \xi(\varepsilon), \qquad \xi(\alpha\varepsilon) \le \xi(\varepsilon)/8 \qquad \text{for all } \varepsilon > 0.$$

Then

$$\log \sup_\mu N(\mathcal{F}, \|\cdot\|_{L^2(\mu)}, \varepsilon) \le 8C \log(\alpha/c)\, \xi(c\varepsilon) \quad \text{for all } \varepsilon > 0.$$

In particular, if $\mathrm{vc}(\mathcal{F}, \varepsilon)$ is comparable to $\xi(\varepsilon)$ in the sense that

$$\xi(\varepsilon/K) \lesssim \mathrm{vc}(\mathcal{F}, \varepsilon) \lesssim \xi(\varepsilon) \quad \text{for all } \varepsilon > 0$$

holds for some constant K, then

$$\mathrm{vc}(\mathcal{F}, 4\varepsilon) \lesssim \log \sup_\mu N(\mathcal{F}, \|\cdot\|_{L^2(\mu)}, \varepsilon) \lesssim \mathrm{vc}(\mathcal{F}, Kc\varepsilon) \quad \text{for all } \varepsilon > 0.$$

Proof. The upper bound follows immediately from Corollary 7.47 and the property $\xi(\alpha^k\varepsilon) \le 8^{-k}\xi(\varepsilon)$. The lower bound follows from Lemma 7.40. □

In applications to empirical processes, we are typically interested not in $N(\mathcal{F}, \|\cdot\|_{L^2(\mu)}, \varepsilon)$ in its own right, but rather in the chaining bound that arises from symmetrization. Applying Theorem 7.30 yields the upper bound

$$\int_0^\infty \sup_\mu \sqrt{\log N(\mathcal{F}, \|\cdot\|_{L^2(\mu)}, \varepsilon)}\, d\varepsilon \lesssim \int_0^\infty \sqrt{\mathrm{vc}(\mathcal{F}, \varepsilon) \log(1/\varepsilon)}\, d\varepsilon,$$

and we have seen that the logarithmic factor can be removed for most infinite-dimensional classes. Surprisingly, however, the latter assumption is not needed: the logarithmic factor can *always* be removed in the entropy integral without any further assumptions! While this is a remarkable result, it should not come as a great surprise: we have essentially already used the iteration method in the proof of Theorem 6.19 in the same manner.

Corollary 7.49 (Entropy integral and combinatorial dimension). *Let $\mathcal{F}$ be a class of functions on $\mathbb{X}$. Then we have*

$$\int_0^\infty \sup_\mu \sqrt{\log N(\mathcal{F}, \|\cdot\|_{L^2(\mu)}, \varepsilon)}\, d\varepsilon \asymp \int_0^\infty \sqrt{\mathrm{vc}(\mathcal{F}, \varepsilon)}\, d\varepsilon.$$

Proof. The lower bound follow immediately from Lemma 7.40. For the upper bound, note that we have by Corollary 7.47 with $\alpha = 4$

$$\sup_\mu \sqrt{\log N(\mathcal{F}, \|\cdot\|_{L^2(\mu)}, \varepsilon)} \lesssim \sum_{k=0}^\infty 2^k \sqrt{\mathrm{vc}(\mathcal{F}, c4^k\varepsilon)}.$$

Integrating both sides and a simple change of variables yields the proof. □

The remainder of this section is devoted to the proof of Theorem 7.45. Somewhat surprisingly, the difficulty of the proof does not lie in the combinatorial aspect of the problem, which is where most of our efforts were spent in the previous sections: the combinatorial part of the proof follows essentially along the same lines as in the proof of Theorem 7.30. As will become clear in due course, the real difficulty of Theorem 7.45 is that the probabilistic extraction principle provided by Lemma 7.17 is no longer adequate when we only assume that the class is bounded in L^p rather than in L^∞.

Let us begin, however, with the combinatorial part of the proof. Following the proof of Theorem 7.30, we first obtain an analogue of Theorem 7.32.

Theorem 7.50. *Let $\mathcal{F}$ be a class of functions and let μ be a probability on $\mathbb{X}$. Then for any $\mathcal{G} \subseteq \mathcal{F}$ that is a $c\varepsilon$-packing of $(\mathcal{F}, \|\cdot\|_{L^p(\mu)})$ for $p \ge 2$, we have*

$$|\mathcal{G}|^{1/p} \le |\{(I,h) : (I,h) \text{ is an } \varepsilon\text{-cube}\}|.$$

Here c is a universal constant.

The proof is almost identical to that of Theorem 7.32, and we only sketch the necessary changes. We first extend Lemma 7.37. It is not at all surprising that this is possible: we left a lot of room in the proof of Lemma 7.37.

Lemma 7.51. *There is a universal constant C so that if $\mathbf{E}[|X - \mathrm{med}(X)|^p] \ge C^p\varepsilon^p$ for some $p \ge 2$, then $\mathbf{P}[X \le b]^{1/p} + \mathbf{P}[X \ge b+\varepsilon]^{1/p} > 1$ for some $b \in \mathbb{R}$.*

Proof. Suppose that the conclusion fails. Then it follows that

$$\mathbf{P}[|X - \mathrm{med}(X)| > k\varepsilon] \le 2^{1-p^k} \quad \text{for all } k \in \mathbb{N}$$

as in the proof of Lemma 7.37. Therefore

$$\mathbf{E}[|X - \mathrm{med}(X)|^p] = \sum_{k=0}^{\infty} \int_{k\varepsilon}^{(k+1)\varepsilon} pt^{p-1}\, \mathbf{P}[|X - \mathrm{med}(X)| > t]\, dt < C^p\varepsilon^p,$$

where we used $\{2p\sum_{k=0}^{\infty}(k+1)^{p-1}2^{-p^k}\}^{1/p} \le 2e\{1+\sum_{k=1}^{\infty}(k+1)2^{-2^{k-1}}\} =: C$ as $p \ge 2$. Thus we proved the contrapositive of the result. □

Proof (Theorem 7.50). We must only prove an analogue of Proposition 7.36: the remainder of the proof is identical to that of Theorem 7.32.

To this end, let $\mathcal{A}$ be a $c\varepsilon$-packing of $(\mathcal{F}, \|\cdot\|_{L^p(\mu)})$ with $|\mathcal{A}| \ge 2$, and draw random elements $a, a' \in \mathcal{A}$ uniformly and independently. Then

$$\frac{c^p\varepsilon^p}{2} \le \mathbf{E}\|a - a'\|_{L^p(\mu)}^p = \int \mathbf{E}|a(x) - a'(x)|^p\, \mu(dx).$$

Thus there exists $x \in \mathbb{X}$ such that

$$\frac{c^p\varepsilon^p}{2} \le \mathbf{E}|a(x) - a'(x)|^p \le 2^p\, \mathbf{E}|a(x) - \mathrm{med}(a(x))|^p,$$

where we used the triangle inequality $|a-a'| \le |a-\mathrm{med}(a)| + |a'-\mathrm{med}(a')|$ and convexity $(x+y)^p \le 2^{p-1}(x^p+y^p)$. We can now apply Lemma 7.51, and the remainder of the proof is identical to that of Proposition 7.36. □

Next, we prove an analogue of Corollary 7.33 in the present setting. The main difficulty here is that we no longer have boundedness of the class in L^∞ but only in L^{2p}. At this stage, however, this is only a minimal inconvenience: even boundedness in L^1 suffices, and the proof is an exercise in counting.

Corollary 7.52. *Let $\mathcal{F}$ be a class of functions on a finite set $\mathbb{X}$, and let μ be the uniform distribution on $\mathbb{X}$. Suppose that $\|f\|_{L^1(\mu)} \le a$ for all $f \in \mathcal{F}$. Then for any $p \ge 2$ and $c\varepsilon$-packing $\mathcal{G}$ of $(\mathcal{F}, \|\cdot\|_{L^p(\mu)})$ with $\varepsilon < a$, we have*

$$|\mathcal{G}|^{1/p} \le \left(\frac{4e^2 a|\mathbb{X}|}{\varepsilon\,\mathrm{vc}(\mathcal{F},\varepsilon)}\right)^{2\,\mathrm{vc}(\mathcal{F},\varepsilon)}.$$

Proof. First, we claim that if (I,h) is an ε-cube, then $\sum_{x\in I}|h(x)| \le a|\mathbb{X}|$. Indeed, as (I,h) is ε-shattered, we can find $f \in \mathcal{F}$ such that $f(x) \le h(x)$ if $h(x) < 0$ and $f(x) \ge h(x)+\varepsilon$ if $h(x) \ge 0$. This implies, in particular, that $|h(x)| \le |f(x)|$ for $x \in I$, and thus the claim follows from $\|f\|_{L^1(\mu)} \le a$.

Now note that given a fixed set $I \subseteq \mathbb{X}$ with $|I| = k$, we have

$$\begin{aligned}
&|\{h \in (\varepsilon\mathbb{Z})^I : \textstyle\sum_{x\in I}|h(x)| \le a|\mathbb{X}|\}| \\
&\le 2^k |\{m_1,\ldots,m_k \in \mathbb{Z}_+ : \textstyle\sum_{i=1}^k m_i \le a|\mathbb{X}|/\varepsilon\}| \\
&= 2^k |\{m_1,\ldots,m_k \in \mathbb{N} : \textstyle\sum_{i=1}^k m_i \le a|\mathbb{X}|/\varepsilon + k\}|.
\end{aligned}$$

As $r_u = \sum_{i=1}^u m_i$ defines a one-to-one correspondence between sequences of integers $m_1,\ldots,m_k \ge 1$ such that $\sum_{i=1}^k m_i \le N$ and increasing sequences of integers $1 \le r_1 < \cdots < r_k \le N$ (of which there are $\binom{N}{k}$), we obtain

$$|\{h : (I,h) \text{ is an } \varepsilon\text{-cube}\}| \le 2^k \binom{\lfloor a|\mathbb{X}|/\varepsilon\rfloor + k}{k} \le \left(\frac{4ea}{\varepsilon}\right)^k \binom{|\mathbb{X}|}{k},$$

where we used $(\frac{N}{k})^k \le \binom{N}{k} \le (\frac{eN}{k})^k$ in the second inequality. Therefore

$$|\{(I,h) \text{ is an } \varepsilon\text{-cube}\}| \le \sum_{k=0}^{\mathrm{vc}(\mathcal{F},\varepsilon)} \binom{|\mathbb{X}|}{k}^2 \left(\frac{4ea}{\varepsilon}\right)^k \le \left[\sum_{k=0}^{\mathrm{vc}(\mathcal{F},\varepsilon)} \binom{|\mathbb{X}|}{k}\left(\frac{4ea}{\varepsilon}\right)^k\right]^2.$$

The right-hand side can be estimated as in the proof of Lemma 7.12, and the proof is completed by applying Theorem 7.50. □

The combinatorial part of the proof is now complete, and all that remains is to apply a probabilistic extraction principle. It is not obvious how to do this, however, as Lemma 7.17 uses uniform boundedness $\sup_f \|f\|_\infty \le 1$ in a fundamental manner. To see why, note that in order for the extraction principle

to yield a nontrivial bound in conjunction with Corollary 7.52, the number of samples r in the extraction principle can be at most (poly)logarithmic in the size of the packing. In Lemma 7.17, the uniform boundedness assumption ensures that the random norm $\|f_i - f_j\|^2_{L^2(\mu_r)}$ is a subgaussian random variable, so that a logarithmic number of samples suffices by a simple union bound. If we only have control of the form $\sup_f \|f\|_q \le 1$ for some $q < \infty$, however, the best we can hope for is a polynomial tail probability for $\|f_i - f_j\|^p_{L^p(\mu_r)}$, and thus a simple union bound gives a polynomial rather than logarithmic number of samples. This does not suffice to conclude the proof.

We must therefore develop a more sophisticated extraction principle. The key idea that makes this possible is that, instead of working directly with the L^p norms $\|f_i - f_j\|_{L^p(\mu)}$, we should focus attention on the tail probabilities $\mu(|f_i - f_j| \ge t)$. The following simple lemma shows how this can be done.

Lemma 7.53. *Let g be a measurable function on the measure space $(\mathbb{X}, \mu)$. If $\|g\|_{L^p(\mu)} > \varepsilon$, then for any $\alpha > 1$ there exists $t \ge 0$ so that*

$$t^{\alpha p}\mu(|g| > t) > \left(\frac{\alpha - 1}{\alpha}\right)^{\alpha} \varepsilon^{\alpha p}.$$

Conversely, if $\|g\|_{L^p(\mu)} \le \varepsilon$, then $t^p\mu(|g| > t) \le \varepsilon^p$ for all $t \ge 0$.

Proof. Suppose that $\mu(|h| > t) \le t^{-\alpha p}$ for all $t \ge 0$. Then we can estimate

$$\|h\|^p_{L^p(\mu)} \le 1 + \int_1^\infty p t^{p-1} \mu(|h| > t)\, dt \le \frac{\alpha}{\alpha - 1}.$$

Inserting $h = (\frac{\alpha}{\alpha-1})^{1/p} g/\varepsilon$ readily yields the contrapositive of the first assertion. The second assertion is immediate from Chebyshev's inequality. □

The key advantage of working with tail probabilities rather than L^p norms is that the empirical measure $\mu_r(|f_i - f_j| \ge t)$ is subgaussian, and we can therefore use a simple union bound to control the empirical tail probabilities using a number of samples that is only logarithmic in the size of the packing. On the other hand, Lemma 7.53 shows that separation in L^p yields a tail bound of order $t^{-p'}$ only if we are willing to lose slightly in the exponent $p' > p$. This explains why it is essential for dimension-free control of L^p-covering numbers that the class is $L^{p'}$-bounded for $p' > p$. Once this idea has been understood, it is not difficult to work out the details.

Proposition 7.54 (Weak extraction). *Let $p \ge 1$, $a > \varepsilon > 0$, $m \ge 2$, and let μ be a probability measure on $\mathbb{X}$. If $f_1, \ldots, f_m$ are functions on $\mathbb{X}$ such that*

$$\|f_i\|_{L^{2p}(\mu)} \le a, \qquad \|f_i - f_j\|_{L^p(\mu)} > \varepsilon \qquad \text{for all } 1 \le i < j \le m,$$

then there exist $r \le C(2a/\varepsilon)^{12p} \log m$ points $x_1, \ldots, x_r \in \mathbb{X}$ and a subset $J \subseteq \{1, \ldots, m\}$ of cardinality $|J| \ge m/2$ such that

$$\|f_i\|_{L^{2p}(\mu^x)} \le 2a, \qquad \|f_i - f_j\|_{L^{3p/2}(\mu^x)} > \varepsilon/9 \qquad \text{for all } i, j \in J,\ i \ne j,$$

where $\mu^x := \frac{1}{r}\sum_{k=1}^r \delta_{x_k}$ and C is a universal constant.

Proof. Let $X_1, \ldots, X_r \sim \mu$ be i.i.d., and denote by μ_r their empirical measure. We begin by controlling the $L^{2p}(\mu_r)$-norm of the functions f_i. Note that

$$\mathbf{P}[\|f_i\|_{L^{2p}(\mu_r)} > 2a] \le \frac{\|f_i\|^{2p}_{L^{2p}(\mu)}}{(2a)^{2p}} \le \frac{1}{4}$$

by Chebyshev's inequality. We therefore have

$$\mathbf{E}|\{i : \|f_i\|_{L^{2p}(\mu_r)} \le 2a\}| = \sum_{i=1}^{m} \mathbf{P}[\|f_i\|_{L^{2p}(\mu_r)} \le 2a] \ge \frac{3m}{4}.$$

Using $\mathbf{E}[Z] < u + \|Z\|_\infty \mathbf{P}[Z \ge u]$, we can estimate

$$\mathbf{P}\left[|\{i : \|f_i\|_{L^{2p}(\mu_r)} \le 2a\}| \ge \frac{m}{2}\right] > \frac{1}{4}.$$

Thus with probability more than one quarter, at least half of the functions f_i remain bounded as $\|f_i\|_{L^{2p}(\mu_r)} \le 2a$ under the empirical measure.

We now turn to controlling the separation between the functions f_i. Applying Lemma 7.53 with $\alpha = 3/2$, we choose $t_{ij} > 0$ for every $i < j$ so that

$$3^{-3/2}\left(\frac{\varepsilon}{t_{ij}}\right)^{3p/2} \le \mu(|f_i - f_j| > t_{ij}) \le \left(\frac{2a}{t_{ij}}\right)^{2p}.$$

Rearranging yields $(\varepsilon/t_{ij})^{3p/2} > 3^{-9/2}(\varepsilon/2a)^{6p}$. We can therefore estimate using the Azuma-Hoeffding inequality

$$\begin{aligned}
&\mathbf{P}\Big[t_{ij}^{3p/2}\mu_r(|f_i - f_j| > t_{ij}) \le 3^{-2}\varepsilon^{3p/2}\Big] \\
&\le \mathbf{P}\Big[t_{ij}^{3p/2}\mu_r(|f_i - f_j| > t_{ij}) \le t_{ij}^{3p/2}\mu(|f_i - f_j| > t_{ij}) - 3^{-3}\varepsilon^{3p/2}\Big] \\
&\le e^{-r3^{-15}(\varepsilon/2a)^{12p}}.
\end{aligned}$$

A union bound now gives

$$\mathbf{P}\Big[t_{ij}^{3p/2}\mu_r(|f_i - f_j| > t_{ij}) > 3^{-2}\varepsilon^{3p/2}\ \forall\, i < j\Big] \ge 1 - m^2 e^{-r3^{-15}(\varepsilon/2a)^{12p}} > \frac{3}{4}$$

for $r \gtrsim (2a/\varepsilon)^{12p} \log m$. In particular, Lemma 7.53 implies that

$$\mathbf{P}\Big[\|f_i - f_j\|_{L^{3p/2}(\mu_r)} > \varepsilon/9 \text{ for all } i < j\Big] > \frac{3}{4}$$

for $r \gtrsim (2a/\varepsilon)^{12p} \log m$. Thus with probability more that three quarters, all functions f_i are separated by $\varepsilon/9$ in $L^{3p/2}(\mu_r)$ under the empirical measure.

Now note that the sum of the probabilities of the events on which boundedness and separation hold under the empirical measure exceeds one if we choose $r = \lfloor C(2a/\varepsilon)^{12p} \log m \rfloor$ for a sufficiently large universal constant C. Thus these events cannot be disjoint, and we can select a sample $x_1, \ldots, x_r$ in their intersection. The conclusion of the proof follows readily. □

We now have all the ingredients to complete the proof of Theorem 7.45.

Proof (Theorem 7.45). Let $f_1, \ldots, f_m \in \mathcal{F}$ be a ε-packing of $(\mathcal{F}, \|\cdot\|_{L^p(\mu)})$ of cardinality $m \geq N(\mathcal{F}, \|\cdot\|_{L^p(\mu)}, \varepsilon)$. By Proposition 7.54 there exist $r \leq C(2a/\varepsilon)^{12p} \log m$ points $x_1, \ldots, x_r \in \mathbb{X}$ and $f_1, \ldots, f_l \in \mathcal{F}$ with $l \geq m/2$ such that $\|f_i\|_{L^{2p}(\mu_r)} \leq 2a$ and $\|f_i - f_j\|_{L^{3p/2}(\mu_r)} \geq \varepsilon/9$ for all $1 \leq i < j \leq l$, where μ_r is the uniform distribution on $x_1, \ldots, x_r$. By Corollary 7.52, we have

$$m \leq \left(\frac{Ka}{\varepsilon}\right)^{39p^2 \operatorname{vc}(\mathcal{F}, \varepsilon/9c)} \left(\frac{\log m}{6p \operatorname{vc}(\mathcal{F}, \varepsilon/9c)}\right)^{3p \operatorname{vc}(\mathcal{F}, \varepsilon/9c)}$$

for a universal constant K. Using $\alpha \log m \leq m^\alpha$ and rearranging, this yields

$$N(\mathcal{F}, \|\cdot\|_{L^p(\mu)}, \varepsilon) \leq m \leq \left(\frac{Ka}{\varepsilon}\right)^{78p^2 \operatorname{vc}(\mathcal{F}, \varepsilon/9c)}.$$

This completes the proof. □

Problems

7.15 (Improved uniform covering bounds). In order to keep the notation minimal, we made some arbitrary choices in the statement and proof of Theorem 7.45. By carefully keeping track of the constants in the proof, extend Theorem 7.45 to bound $N(\mathcal{F}, \|\cdot\|_{L^p(\mu)}, \varepsilon)$ under the assumption $\sup_f \|f\|_{L^{\beta p}(\mu)} \leq a$ for any $p \geq 1$ and $\beta > 1$ as indicated in Remark 7.46.

7.16 (L^∞-covering numbers and combinatorial dimension). Throughout this chapter, we have obtained dimension-free estimates for L^p-covering numbers with $p < \infty$. One cannot expect to obtain dimension-free L^∞-covering numbers, however. For example, when $\mathcal{F}$ is a class of indicator functions on a finite set $\mathbb{X}$, then $N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) = |\mathcal{F}|$ for all $0 < \varepsilon < 1$ and thus any nontrivial L^∞-covering number bound must depend on $|\mathbb{X}|$. While this dependence is in general exponential, the Sauer-Shelah Lemma 7.12 states that the L^∞-covering numbers grow only polynomially in $|\mathbb{X}|$ for VC-classes of sets. It is natural to ask whether this is also true for general function classes.

a. Let $\mathbb{X}$ be a finite set and let μ be the uniform distribution on $\mathbb{X}$. Show that $e^{-1}\|f\|_\infty \leq \|f\|_{L^{\log|\mathbb{X}|}(\mu)} \leq \|f\|_\infty$ for every function f on $\mathbb{X}$.

b. Deduce from Corollary 7.52 that if $\mathcal{F}$ is a class of functions on a finite set $\mathbb{X}$ such that $\|f\|_\infty \leq 1$ for all $f \in \mathcal{F}$, then for universal constants c, C

$$\log N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \leq 2 \operatorname{vc}(\mathcal{F}, c\varepsilon) \log|\mathbb{X}| \log\left(\frac{C|\mathbb{X}|}{\varepsilon \operatorname{vc}(\mathcal{F}, c\varepsilon)}\right).$$

For classes of sets $\mathcal{C}$ the Sauer-Shelah lemma implies $\log N(\mathcal{C}, \|\cdot\|_\infty, \varepsilon) \lesssim \log|\mathbb{X}|$, while we have obtained above a bound of order $\log^2|\mathbb{X}|$ for arbitrary function classes $\mathcal{F}$. It is not known whether a polynomial bound is possible in the general setting. However, we can achieve nearly polynomial scaling by improving the above bound to $\log^{1+\delta}|\mathbb{X}|$ for any $\delta > 0$.

c. The small deviation result of Lemma 7.51 is not the most efficient. Show that the conclusion can be improved to $\mathbf{P}[X \le b]^{1/p^\delta} + \mathbf{P}[X \ge b+\varepsilon]^{1/p^\delta} > 1$ for any $\delta > 0$, with the constant C depending on δ but not on p.

d. Prove a general bound of order $\log N(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \lesssim \log^{1+\delta}|\mathbb{X}|$.

e. Similarly, the scaling $\propto p^2$ of the bound of Theorem 7.45 is not the best possible. Show that the scaling can be improved to $\propto p^{1+\delta}$ for any $\delta > 0$.

7.17 (Iteration and Sudakov's inequality). We have systematically developed upper and lower bounds for the suprema of random processes in terms of covering numbers. An implicit motivation for these results is that it is often easier to bound the covering numbers of a set T than to bound directly a random process defined on T. However, these results prove to be useful also in the converse direction: there are situations where a direct estimate on the supremum of a random process can be used to obtain nontrivial bounds for covering numbers that are otherwise hard to compute.

The simplest result that can be used for this purpose is Sudakov's inequality. Let $T \subseteq B(0,1)$ be a subset of the Euclidean unit ball in $\mathbb{R}^n$, and

$$X_t := \sum_{i=1}^n g_i t_i, \qquad \omega(\varepsilon) := \sup_{s\in T} \mathbf{E}\left[\sup_{t\in T\cap B(s,\varepsilon)} X_t\right]$$

where $g_1, \ldots, g_n$ are i.i.d. $N(0,1)$. Note that $\{X_t\}_{t\in T}$ is a Gaussian process whose natural distance d is the Euclidean distance. We can therefore estimate

$$\log N(T, d, \varepsilon) \lesssim \frac{\omega(1)^2}{\varepsilon^2}.$$

How good is this bound? Unfortunately, it leaves something to be desired.

a. Let $T = B(0,1)$ be the Euclidean unit ball. Show that Sudakov's inequality yields at best $\log N(T, d, \varepsilon) \lesssim n/\varepsilon^2$. On the other hand, show that in fact $\log N(T, d, \varepsilon) \asymp n\log(1/\varepsilon)$, which is far better than is predicted by Sudakov.

It is not too surprising that Sudakov's inequality fails to capture the correct behavior of the covering numbers even in the simplest possible example: $\omega(1) < \infty$ can hold even for infinite-dimensional classes, and thus we cannot predict correctly the behavior of the covering numbers on the basis of this quantity only. On the other hand, the local modulus of continuity $\omega(\varepsilon)$ contains much more information. It can be exploited using an iteration argument.

b. Show that for any $\varepsilon > 0$

$$\log N(T, d, \varepsilon) \lesssim \sum_{k=0}^{\infty} \mathbf{1}_{2^k\varepsilon<1} \frac{\omega(2^{k+1}\varepsilon)^2}{(2^k\varepsilon)^2} \lesssim \int_{\varepsilon}^{2} \frac{\omega(2x)^2}{x^3} dx.$$

c. Show that if $T = B(0, 1)$ is the Euclidean unit ball, then $\omega(x) \le x\sqrt{n}$ and thus iteration yields a covering number estimate of the correct order.

Notes

§7.1. The symmetrization method, which has its origin in probability in Banach spaces, has been a fundamental part of empirical process theory following the influential work of Giné and Zinn [40]. A slightly different form of symmetrization was already used by Vapnik and Chervonenkis [92]. Lemma 7.6 is due to Panchenko [64]. The characterization of Bernoulli processes mentioned in Problem 7.1 was proved by Bednorz and Latała [9] (see also [89] for an exposition). The simple contraction method used in Problem 7.1 is classical [51], while the "inverse" Gaussian symmetrization method is based on [67]. Problem 7.2 is based on [40] (the result developed here dates back to [93]). See also [82] for more precise characterizations of the Glivenko-Cantelli property. Much more on self-normalized processes (Problem 7.3) can be found in [65]. The contraction principle of Problem 7.4 can be found in [51].

§7.2. The notion of VC-dimension and its application to the Glivenko-Cantelli problem were developed by Vapnik and Chervonenkis [92]. The Sauer-Shelah lemma was proved by Sauer in answer to a question posed by Erdős [72]; an infinite version of it appeared in work on mathematical logic by Shelah. Theorem 7.16 is due to Dudley [29]. Uniform Glivenko-Cantelli classes were studied systematically by Dudley, Giné and Zinn [32] and Alon et al. [4]. Pajor's formulation of the Sauer-Shelah lemma is from [63]. The somewhat pedantic proof we have given here (based on [60]) is intended to prepare the reader for the next section. Classical proofs are developed in Problems 7.7 and 7.8. The formulation of the Glivenko-Cantelli theorem in Problem 7.10 is due to Steele [74]; the example of convex sets follows the treatment in [68]. Problem 7.11 gives a very brief introduction to the topic of uniform central limit theorems that has historically motivated many developments in empirical process theory; textbook treatments can be found in [30, 91].

§7.3. The notion of combinatorial dimension has its origin in Banach space theory. It was used implicitly by Elton [34] following the development of an infinite counterpart of this idea by Rosenthal [69] to characterize Banach spaces that embed ℓ_1 (see [45] for the probabilistic significance of the latter notion). A first result along the lines of Theorem 7.30, but with much worse scaling, is due to Pajor [63]. Theorem 7.30, due to Mendelson and Vershynin

[60], is essentially the best possible. The much simpler notion of VC-subgraph classes (Problem 7.12) appeared independently, cf. [68]. Problem 7.13 is taken from [61], while the approach of Problem 7.14 follows [60].

§7.4. The lower bound in Lemma 7.40 is from [87]. The iteration method is often used in Banach space theory; see, for example, [5] for an interesting application. Example 7.44 is inspired by the example given in [6, Lemma 4.9]. Theorem 7.45 and its use as an iteration principle are due to Rudelson and Vershynin [70], and we follow a simplified version of their proof. L^∞-covering bounds in terms of combinatorial dimension (Problem 7.16) were first obtained in [4] with a worse scaling. Problem 7.17 is inspired by [15].

References

1. Adler, R.J.: An introduction to continuity, extrema, and related topics for general Gaussian processes. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 12. Institute of Mathematical Statistics, Hayward, CA (1990)
2. Adler, R.J., Taylor, J.E.: Random fields and geometry. Springer Monographs in Mathematics. Springer, New York (2007)
3. Ajtai, M., Komlós, J., Tusnády, G.: On optimal matchings. Combinatorica **4**(4), 259–264 (1984)
4. Alon, N., Ben-David, S., Cesa-Bianchi, N., Haussler, D.: Scale-sensitive dimensions, uniform convergence, and learnability. J. ACM **44**(4), 615–631 (1997)
5. Artstein, S., Milman, V., Szarek, S.J.: Duality of metric entropy. Ann. of Math. (2) **159**(3), 1313–1328 (2004)
6. Assouad, P.: Densité et dimension. Ann. Inst. Fourier (Grenoble) **33**(3), 233–282 (1983)
7. Bakry, D., Gentil, I., Ledoux, M.: Analysis and geometry of Markov diffusion operators, *Grundlehren der Mathematischen Wissenschaften*, vol. 348. Springer, Cham (2014)
8. Bednorz, W.: A theorem on majorizing measures. Ann. Probab. **34**(5), 1771–1781 (2006)
9. Bednorz, W., Latała, R.: On the suprema of Bernoulli processes. C. R. Math. Acad. Sci. Paris **351**(3-4), 131–134 (2013)
10. Bobkov, S., Ledoux, M.: Poincaré's inequalities and Talagrand's concentration phenomenon for the exponential distribution. Probab. Theory Related Fields **107**(3), 383–400 (1997)
11. Bobkov, S.G., Götze, F.: Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. J. Funct. Anal. **163**(1), 1–28 (1999)
12. Bobkov, S.G., Tetali, P.: Modified logarithmic Sobolev inequalities in discrete settings. J. Theoret. Probab. **19**(2), 289–336 (2006)
13. Boucheron, S., Lugosi, G., Massart, P.: Concentration inequalities. Oxford University Press, Oxford (2013)
14. Boucheron, S., Thomas, M.: Concentration inequalities for order statistics. Electron. Commun. Probab. **17**, no. 51, 12 (2012)
15. Bousquet, O., Koltchinskii, V., Panchenko, D.: Some local measures of complexity of convex hulls and generalization bounds. In: J. Kivinen, R.H. Sloan (eds.)

Computational Learning Theory, *Lecture Notes in Computer Science*, vol. 2375, pp. 59–73. Springer Berlin Heidelberg (2002)
16. Brazitikos, S., Giannopoulos, A., Valettas, P., Vritsiou, B.H.: Geometry of isotropic convex bodies, *Mathematical Surveys and Monographs*, vol. 196. American Mathematical Society, Providence, RI (2014)
17. Chafaï, D.: Entropies, convexity, and functional inequalities: on Φ-entropies and Φ-Sobolev inequalities. J. Math. Kyoto Univ. **44**(2), 325–363 (2004)
18. Chatterjee, S.: Superconcentration and related topics. Springer Monographs in Mathematics. Springer, Cham (2014)
19. Chentsov, N.N.: Weak convergence of stochastic processes whose trajectories have no discontinuities of the second kind and the "heuristic" approach to the Kolmogorov-Smirnov tests. Theor. Probab. Appl. **1**, 140–144 (1956)
20. Csiszár, I., Körner, J.: Information theory, second edn. Cambridge University Press, Cambridge (2011)
21. Dembo, A.: Information inequalities and concentration of measure. Ann. Probab. **25**(2), 927–939 (1997)
22. Dembo, A., Zeitouni, O.: Large deviations techniques and applications, *Stochastic Modelling and Applied Probability*, vol. 38. Springer-Verlag, Berlin (2010)
23. Ding, J., Lee, J.R., Peres, Y.: Cover times, blanket times, and majorizing measures. Ann. of Math. (2) **175**(3), 1409–1471 (2012)
24. Djellout, H., Guillin, A., Wu, L.: Transportation cost-information inequalities and applications to random dynamical systems and diffusions. Ann. Probab. **32**(3B), 2702–2732 (2004)
25. Dobrušin, R.L.: Definition of a system of random variables by means of conditional distributions. Teor. Verojatnost. i Primenen. **15**, 469–497 (1970)
26. Dubhashi, D.P., Panconesi, A.: Concentration of measure for the analysis of randomized algorithms. Cambridge University Press, Cambridge (2009)
27. Dudley, R.M.: The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. J. Functional Analysis **1**, 290–330 (1967)
28. Dudley, R.M.: Sample functions of the Gaussian process. Ann. Probability **1**(1), 66–103 (1973)
29. Dudley, R.M.: Central limit theorems for empirical measures. Ann. Probab. **6**(6), 899–929 (1979) (1978)
30. Dudley, R.M.: Uniform central limit theorems, *Cambridge Studies in Advanced Mathematics*, vol. 63. Cambridge University Press, Cambridge (1999)
31. Dudley, R.M.: Real analysis and probability, *Cambridge Studies in Advanced Mathematics*, vol. 74. Cambridge University Press, Cambridge (2002)
32. Dudley, R.M., Giné, E., Zinn, J.: Uniform and universal Glivenko-Cantelli classes. J. Theoret. Probab. **4**(3), 485–510 (1991)
33. Efron, B., Stein, C.: The jackknife estimate of variance. Ann. Statist. **9**(3), 586–596 (1981)
34. Elton, J.: Sign-embeddings of l_1^n. Trans. Amer. Math. Soc. **279**(1), 113–124 (1983)
35. Fernique, X.: Regularité des trajectoires des fonctions aléatoires gaussiennes. In: École d'Été de Probabilités de Saint-Flour, IV-1974, pp. 1–96. Lecture Notes in Math., Vol. 480. Springer, Berlin (1975)
36. Gärtner, B., Matoušek, J.: Understanding and Using Linear Programming. Universitext. Springer-Verlag, Berlin (2007)
37. van de Geer, S.: Oracle inequalities and regularization. In: Lectures on empirical processes, EMS Ser. Lect. Math., pp. 191–252. Eur. Math. Soc., Zürich (2007)

38. van de Geer, S.A.: Applications of empirical process theory, *Cambridge Series in Statistical and Probabilistic Mathematics*, vol. 6. Cambridge University Press, Cambridge (2000)
39. Gibbs, J.W.: Elementary Principles in Statistical Mechanics. Charles Scribner's Sons, New York (1902)
40. Giné, E., Zinn, J.: Some limit theorems for empirical processes. Ann. Probab. **12**(4), 929–998 (1984)
41. Gozlan, N.: A characterization of dimension free concentration in terms of transportation inequalities. Ann. Probab. **37**(6), 2480–2498 (2009)
42. Gross, L.: Logarithmic Sobolev inequalities. Amer. J. Math. **97**(4), 1061–1083 (1975)
43. Guédon, O., Zvavitch, A.: Supremum of a process in terms of trees. In: Geometric aspects of functional analysis, *Lecture Notes in Math.*, vol. 1807, pp. 136–147. Springer, Berlin (2003)
44. Guionnet, A., Zegarlinski, B.: Lectures on logarithmic Sobolev inequalities. In: Séminaire de Probabilités, XXXVI, *Lecture Notes in Math.*, vol. 1801, pp. 1–134. Springer, Berlin (2003)
45. van Handel, R.: The universal Glivenko-Cantelli property. Probab. Theory Related Fields **155**(3-4), 911–934 (2013)
46. Karatzas, I., Shreve, S.E.: Brownian motion and stochastic calculus, *Graduate Texts in Mathematics*, vol. 113, second edn. Springer-Verlag, New York (1991)
47. Kolmogorov, A.N., Tihomirov, V.M.: ε-entropy and ε-capacity of sets in function spaces. Uspehi Mat. Nauk **14**(2 (86)), 3–86 (1959)
48. Ledoux, M.: On Talagrand's deviation inequalities for product measures. ESAIM Probab. Statist. **1**, 63–87 (electronic) (1995/97)
49. Ledoux, M.: Isoperimetry and Gaussian analysis. In: Lectures on probability theory and statistics (Saint-Flour, 1994), *Lecture Notes in Math.*, vol. 1648, pp. 165–294. Springer, Berlin (1996)
50. Ledoux, M.: The concentration of measure phenomenon, *Mathematical Surveys and Monographs*, vol. 89. American Mathematical Society, Providence, RI (2001)
51. Ledoux, M., Talagrand, M.: Probability in Banach spaces, *Ergebnisse der Mathematik und ihrer Grenzgebiete*, vol. 23. Springer-Verlag, Berlin (1991)
52. Liggett, T.M.: Continuous time Markov processes, *Graduate Studies in Mathematics*, vol. 113. American Mathematical Society, Providence, RI (2010)
53. Marcus, M.B., Rosen, J.: Markov processes, Gaussian processes, and local times, *Cambridge Studies in Advanced Mathematics*, vol. 100. Cambridge University Press, Cambridge (2006)
54. Marton, K.: Bounding $\bar{d}$-distance by informational divergence: a method to prove measure concentration. Ann. Probab. **24**(2), 857–866 (1996)
55. Marton, K.: A measure concentration inequality for contracting Markov chains. Geom. Funct. Anal. **6**(3), 556–571 (1996)
56. Matoušek, J.: Lectures on discrete geometry, *Graduate Texts in Mathematics*, vol. 212. Springer-Verlag, New York (2002)
57. Maurey, B.: Some deviation inequalities. Geom. Funct. Anal. **1**(2), 188–197 (1991)
58. McDiarmid, C.: On the method of bounded differences. In: Surveys in combinatorics, 1989 (Norwich, 1989), *London Math. Soc. Lecture Note Ser.*, vol. 141, pp. 148–188. Cambridge Univ. Press, Cambridge (1989)
59. Mendel, M., Naor, A.: Ultrametric skeletons. Proc. Natl. Acad. Sci. USA **110**(48), 19,256–19,262 (2013)

60. Mendelson, S., Vershynin, R.: Entropy and the combinatorial dimension. Invent. Math. **152**(1), 37–55 (2003)
61. Mendelson, S., Vershynin, R.: Remarks on the geometry of coordinate projections in $\mathbb{R}^n$. Israel J. Math. **140**, 203–220 (2004)
62. Milman, V.D., Schechtman, G.: Asymptotic theory of finite-dimensional normed spaces, *Lecture Notes in Mathematics*, vol. 1200. Springer-Verlag, Berlin (1986)
63. Pajor, A.: Sous-espaces l_1^n des espaces de Banach, *Travaux en Cours [Works in Progress]*, vol. 16. Hermann, Paris (1985)
64. Panchenko, D.: Symmetrization approach to concentration inequalities for empirical processes. Ann. Probab. **31**(4), 2068–2081 (2003)
65. de la Peña, V.H., Lai, T.L., Shao, Q.M.: Self-normalized processes. Probability and its Applications (New York). Springer-Verlag, Berlin (2009)
66. Pisier, G.: Some applications of the metric entropy condition to harmonic analysis. In: Banach spaces, harmonic analysis, and probability theory (Storrs, Conn., 1980/1981), *Lecture Notes in Math.*, vol. 995, pp. 123–154. Springer, Berlin (1983)
67. Pisier, G.: Probabilistic methods in the geometry of Banach spaces. In: Probability and analysis (Varenna, 1985), *Lecture Notes in Math.*, vol. 1206, pp. 167–241. Springer, Berlin (1986)
68. Pollard, D.: Convergence of stochastic processes. Springer Series in Statistics. Springer-Verlag, New York (1984)
69. Rosenthal, H.P.: A characterization of Banach spaces containing l^1. Proc. Nat. Acad. Sci. U.S.A. **71**, 2411–2413 (1974)
70. Rudelson, M., Vershynin, R.: Combinatorics of random processes and sections of convex bodies. Ann. of Math. (2) **164**(2), 603–648 (2006)
71. Samson, P.M.: Concentration of measure inequalities for Markov chains and Φ-mixing processes. Ann. Probab. **28**(1), 416–461 (2000)
72. Sauer, N.: On the density of families of sets. J. Combinatorial Theory Ser. A **13**, 145–147 (1972)
73. Slepian, D.: The one-sided barrier problem for Gaussian noise. Bell System Tech. J. **41**, 463–501 (1962)
74. Steele, J.M.: Empirical discrepancies and subadditive processes. Ann. Probability **6**(1), 118–127 (1978)
75. Talagrand, M.: Regularity of Gaussian processes. Acta Math. **159**(1-2), 99–149 (1987)
76. Talagrand, M.: An isoperimetric theorem on the cube and the Kintchine-Kahane inequalities. Proc. Amer. Math. Soc. **104**(3), 905–909 (1988)
77. Talagrand, M.: A simple proof of the majorizing measure theorem. Geom. Funct. Anal. **2**(1), 118–125 (1992)
78. Talagrand, M.: Constructions of majorizing measures, Bernoulli processes and cotype. Geom. Funct. Anal. **4**(6), 660–717 (1994)
79. Talagrand, M.: The supremum of some canonical processes. Amer. J. Math. **116**(2), 283–325 (1994)
80. Talagrand, M.: Concentration of measure and isoperimetric inequalities in product spaces. Inst. Hautes Études Sci. Publ. Math. **81**, 73–205 (1995)
81. Talagrand, M.: Applying a theorem of Fernique. Ann. Inst. H. Poincaré Probab. Statist. **32**(6), 779–799 (1996)
82. Talagrand, M.: The Glivenko-Cantelli problem, ten years later. J. Theoret. Probab. **9**(2), 371–384 (1996)

83. Talagrand, M.: Majorizing measures: the generic chaining. Ann. Probab. **24**(3), 1049–1103 (1996)
84. Talagrand, M.: A new look at independence. Ann. Probab. **24**(1), 1–34 (1996)
85. Talagrand, M.: Transportation cost for Gaussian and other product measures. Geom. Funct. Anal. **6**(3), 587–600 (1996)
86. Talagrand, M.: Majorizing measures without measures. Ann. Probab. **29**(1), 411–417 (2001)
87. Talagrand, M.: Vapnik-Chervonenkis type conditions and uniform Donsker classes of functions. Ann. Probab. **31**(3), 1565–1582 (2003)
88. Talagrand, M.: The generic chaining. Springer Monographs in Mathematics. Springer-Verlag, Berlin (2005)
89. Talagrand, M.: Upper and lower bounds for stochastic processes, *Ergebnisse der Mathematik und ihrer Grenzgebiete*, vol. 60. Springer, Heidelberg (2014)
90. Tsirelson, B.S., Ibragimov, I.A., Sudakov, V.N.: Norms of Gaussian sample functions. In: Proceedings of the Third Japan-USSR Symposium on Probability Theory (Tashkent, 1975), pp. 20–41. Lecture Notes in Math., Vol. 550. Springer, Berlin (1976)
91. van der Vaart, A.W., Wellner, J.A.: Weak convergence and empirical processes. Springer Series in Statistics. Springer-Verlag, New York (1996)
92. Vapnik, V.N., Červonenkis, A.J.: The uniform convergence of frequencies of the appearance of events to their probabilities. Teor. Verojatnost. i Primenen. **16**, 264–279 (1971)
93. Vapnik, V.N., Chervonenkis, A.Y.: Necessary and sufficient conditions for the uniform convergence of empirical means to their true values. Teor. Veroyatnost. i Primenen. **26**(3), 543–563 (1981)
94. Vershik, A.M.: Long history of the Monge-Kantorovich transportation problem. Math. Intelligencer **35**(4), 1–9 (2013)
95. Vershynin, R.: Introduction to the non-asymptotic analysis of random matrices. In: Compressed sensing, pp. 210–268. Cambridge Univ. Press, Cambridge (2012)
96. Viens, F.G., Vizcarra, A.B.: Supremum concentration inequality and modulus of continuity for sub-nth chaos processes. J. Funct. Anal. **248**(1), 1–26 (2007)
97. Villani, C.: Topics in optimal transportation, *Graduate Studies in Mathematics*, vol. 58. American Mathematical Society, Providence, RI (2003)
98. Villani, C.: Optimal transport, *Grundlehren der Mathematischen Wissenschaften*, vol. 338. Springer-Verlag, Berlin (2009)
99. Wu, L.: Poincaré and transportation inequalities for Gibbs measures under the Dobrushin uniqueness condition. Ann. Probab. **34**(5), 1960–1989 (2006)