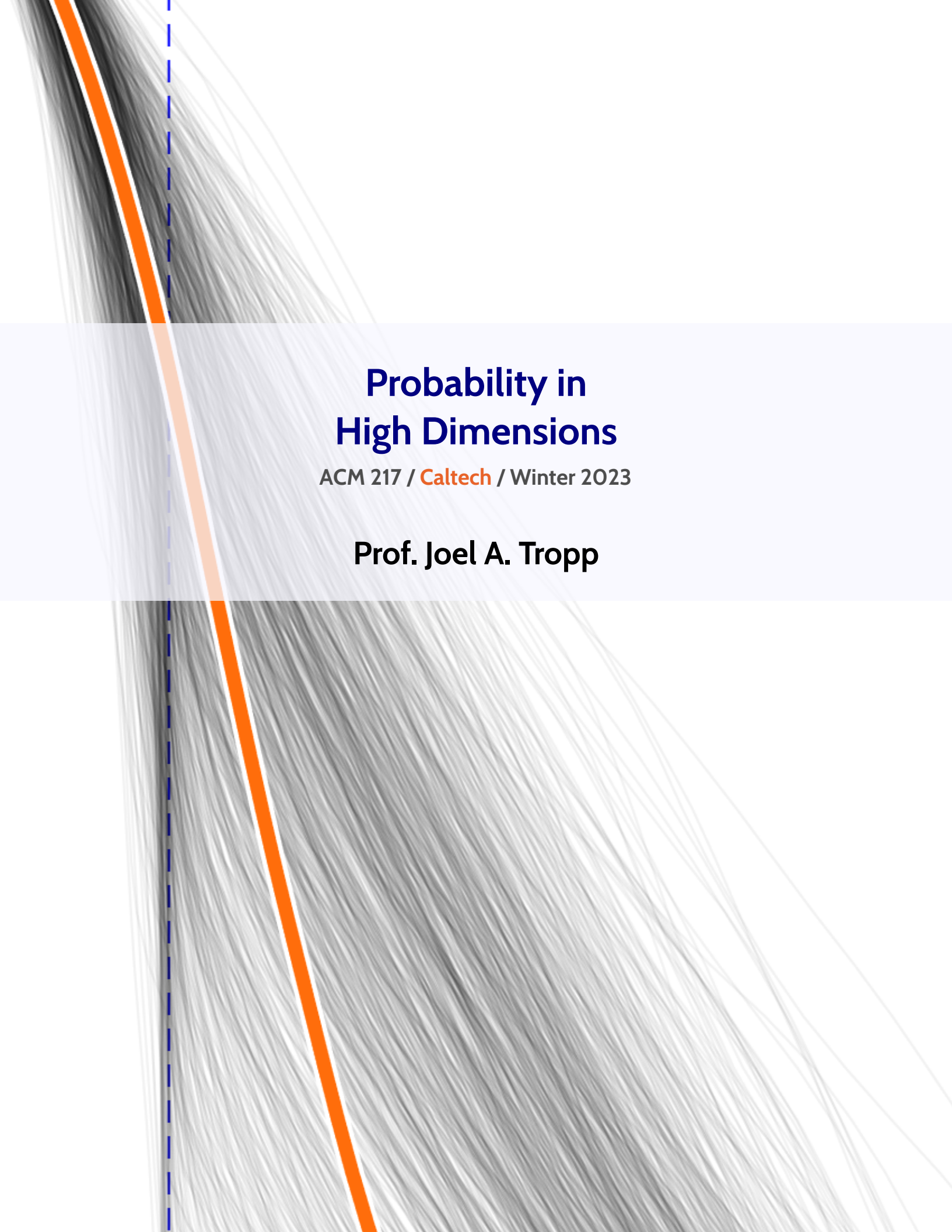


Probability in High Dimensions

ACM 217 / Caltech / Winter 2023

Prof. Joel A. Tropp

Typeset on March 2, 2023

Cite as: Joel A. Tropp, *ACM 217: Probability in High Dimensions*, Caltech CMS Lecture Notes 2021-01, Pasadena, March 2021. Corrected March 2023.

Available from https://doi.org/10.7907/mxr0-c422

Cover image: Sample paths of a randomized block Krylov method for estimating the largest eigenvalue of a symmetric matrix.

Contents

Preface

> "I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the law of frequency of error. The law would have been personified by the Greeks if they had known of it. It reigns with serenity and complete self-effacement amidst the wildest confusion. The larger the mob, the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of unreason."
>
> —Sir Francis Galton, 1889

ACM 217 is a second-year graduate course on high-dimensional probability, designed for students in computing and mathematical sciences. We discuss phenomena that emerge from probability models with many degrees of freedom, tools for working with these models, and a selection of applications to computational mathematics.

Course overview

After an introductory lecture that describes the character of high-dimensional probability, the course is divided into three parts. The first part concerns concentration inequalities, a family of results that describe situations when a random variable typically takes values close to its mean. The second part develops bounds for the suprema of random processes, with a focus on Gaussian processes. The third part turns to questions about empirical processes, random processes that arise from sampling points from a population.

These notes

The Winter 2021 edition of ACM 217 is the fourth instantiation of a class that initially focused on concentration inequalities and that has expanded to include other topics in high-dimensional probability. This year, the course was more mathematical than some previous editions, with less attention to tools and applications. This slant may not serve applied students well, and it is possible that future versions of the course will strike a different balance between theory and practice.

These lecture notes document ACM 217 as it was taught in Winter 2021. The notes were transcribed by the students as part of their coursework, and they were edited lightly by the instructor. They are intended as a record for the students who have taken the course. Other readers should beware that this course is neither refined nor especially coherent. *There is no warranty about correctness. Furthermore, these notes have been prepared using many sources and without appropriate scholarly citations.* This version has been updated with corrections identified during the Winter 2023 term.

Prerequisites

The prerequisites for this course are differential and integral calculus (e.g., Caltech Math 1ac), fluency with linear algebra (e.g., ACM 104 and ACM 107), and a thorough grounding in probability (e.g., ACM 116 and ACM 117). Exposure to functional analysis (e.g., ACM 105) is valuable but not essential.

Supplemental textbooks

There is no required textbook for the course. Highly recommended resources include

- [Tro15a] Tropp, *An introduction to matrix concentration inequalities*, 2015.
- [van16] van Handel, *Probability in high dimensions*, 2016.
- [Ver18] Vershynin, *High-dimensional probability*, 2018.

Vershynin's book is an elegant introduction to ideas from high-dimensional probability, focusing on basic technical methods, useful results, and stylized applications. Van Handel's lecture notes are more systematic and mathematical, and they are concerned with general phenomena that emerge from high-dimensional probability models. My monograph treats all the basic matrix concentration inequalities, along with many applications in computational mathematics. *These notes draw extensively from these references.*

Some other relevant surveys and books include

- [Bar05] Barvinok, "Concentration of measure," 2005.
- [BLM13] Boucheron et al., "Concentration inequalities," 2013.
- [FR13] Foucart and Rauhut, *A mathematical introduction to compressed sensing*, 2012.
- [Led01] Ledoux, *The concentration of measure phenomenon*, 2001.
- [LT11] Ledoux and Talagrand, *Probability in Banach spaces*, 1991.
- [Ros11] Ross, "Fundamentals of Stein's method," 2011.
- [Tro17] Tropp, "ACM 217: Lecture notes on concentration inequalities," 2017.
- [Tro19] Tropp, "Matrix concentration and computational linear algebra," 2019.
- [van17] van Handel, "Structured random matrices," 2017.
- [Ver12] Vershynin, "Introduction to the non-asymptotic analysis of random matrices," 2012.
- [Wai19] Wainwright, *High-dimensional statistics*, 2019.

Most of these references are freely available online, or they can be downloaded from the Caltech library.

Acknowledgements

These notes have been transcribed from the lectures by the students in the course: Chi-Fang Chen, Yifan Chen, Anushri Dixit, Ethan Epperly, Hamed Hamze, Hsin-Yuan Huang, Taylan Kargin, Eitan Levin, Jack Li, Serena Liu, Riley Murray, Nicholas H. Nelson, Joe Slote, Roy Wang, Jing Yu, Kevin Yu, Ziyun Zhang. Credit is also due to participants who proposed corrections: Heri Rajaoberison, Sabhrant Sachan, Rob Webber. Many thanks are due for their care and diligence. All remaining errors are the fault of the instructor.

Joel A. Tropp
Steele Family Professor of Applied & Computational Mathematics
California Institute of Technology

jtropp@cms.caltech.edu
http://users.cms.caltech.edu/~jtropp

Pasadena, California
March 2021, March 2023

Notation and Definitions

The notation in this course is standard in probability theory and related fields.

Set theory

The Pascal notation := and =: generates a definition. Sets without any particular internal structure are denoted with sans serif capitals: $\mathsf{A}, \mathsf{B}, \mathsf{E}$. Collections of sets are written in a calligraphic font: $\mathscr{A}, \mathscr{B}, \mathscr{F}$.

The natural numbers $\mathbb{N} := \{1, 2, 3, \dots\}$. Ordered tuples and sequences are written with parentheses, e.g.,

$$(a_1, a_2, a_3, \dots, a_n) \quad \text{or} \quad (a_1, a_2, a_3, \dots)$$

Alternative notations include things like $(a_i : i \in \mathbb{N})$ or $(a_i)_{i \in \mathbb{N}}$ or simply (a_i).

Real analysis

We write $\mathbb{R}$ for the field of real numbers, equipped with the absolute value $|\cdot|$. The extended real numbers $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ are defined with the usual rules of arithmetic and order. In particular, we instate the conventions that $0/0 = 0$ and $0 \cdot \pm\infty = 0$. We use the standard (American) notation for open and closed intervals, e.g.,

$$(a, b) := \{x \in \overline{\mathbb{R}} : a < x < b\} \quad \text{and} \quad [a, b] := \{x \in \overline{\mathbb{R}} : a \le x \le b\}.$$

Occasionally, we will visit the complex field $\mathbb{C}$.

In this course, we enforce the convention that *positive* means ≥ 0 and *negative* means ≤ 0. For example, the positive integers compose the set $\mathbb{Z}_+ := \{0, 1, 2, 3, \dots\}$, and the positive reals compose the set $\mathbb{R}_+ := \{x \in \mathbb{R} : x \ge 0\}$. When required, we may deploy the phrase *strictly positive* to mean > 0 and *strictly negative* to mean < 0. Similarly, *increasing* means "never going down" and *decreasing* means "never going up."

Warning: Positive means ≥ 0! ■

Linear algebra

We usually denote scalars with lowercase Roman (a, b) or Greek (α, β) letters. Lowercase boldface italics $(\boldsymbol{a}, \boldsymbol{b})$ refer to vectors. Uppercase boldface italics $(\boldsymbol{A}, \boldsymbol{B})$ are associated with matrices or linear maps.

The symbol * denotes the (conjugate) transpose of a vector or matrix. The operator tr returns the trace of a square matrix. Nonlinear functions bind before the trace.

Norms and pseudonorms are denoted with double bars: $\|\cdot\|$. We typically add a subscript to refer to a specific norm, such as the Euclidean norm $\|\cdot\|_{\ell_2}$.

Probability

Uppercase italic letters (near the end of the Roman alphabet) usually refer to (real) random variables: W, X, Y, Z. For vector-valued random variables, we typically use lowercase boldface italic: $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$. Matrix-valued random variables will often be denoted with uppercase boldface italic: $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$.

In some instances, random variables may take values in a Polish space Ω. You will not miss much by assuming that $\Omega = \mathbb{R}$ or $\Omega = \mathbb{R}^n$. To avoid distracting technicalities, we do not always decorate theorems with complete hypotheses. You should always assume that random variables are sufficiently regular for the results to make sense (e.g., having two finite moments or taking values in a Polish space).

A *Polish space* is a complete, separable metric space.

We use small capitals for named distributions. For example, UNIFORM or NORMAL. The symbol $\sim$ means "has the distribution." We write "iid" for *independent and identically distributed*.

The map $\mathbb{P}\{\cdot\}$ returns the probability of an event. The operator $\mathbb{E}[\cdot]$ returns the expectation of a random variable taking values in a linear space. We only include the brackets when it is necessary for clarity, and we impose the convention that nonlinear functions bind before the expectation.

The operator $\mathrm{Var}[\cdot]$ returns the variance of a random variable, while $\mathrm{Cov}(\cdot, \cdot)$ computes the covariance of a pair of random variables. Occasionally, we may write $\mathbb{M}[\cdot]$ for the median.

1. Introduction to HDP

Date: 5 January 2021 **Scribe:** Joel Tropp

Agenda:

1. What is HDP?
2. Applications of HDP
3. Nonasymptotic analysis
4. Concentration
5. Suprema
6. Universality
7. Phase transitions

Probability theory is the study of predictable phenomena that arise from randomness. Although a simple probabilistic experiment has an unpredictable outcome, we can still make confident statements about the aggregate behavior of a large number of experiments. For example, if we flip a fair coin 100 times, we anticipate that it turns up heads roughly half of the time. We would be shocked if this experiment yielded 97 heads, and we would rightly question whether the coin is indeed fair.

Classical limit laws offer detailed information about what happens when we increase the number of experiments without bound. High-dimensional probability (HDP) studies an intermediate regime: models where the number of random variables is large but fixed. HDP also considers phenomena that occur in high-dimensional (linear) spaces. We will see that many of the classical limit laws admit nonasymptotic variants that quantify the precise role of the number of random variables or the dimension. While results in HDP are usually not as lapidary as limit laws, they are far more useful for applications in computational mathematics and statistics—which are finite by nature.

The goal of this course is to develop some of the core principles and mathematics behind HDP, with applications to computational mathematics. The lectures will focus primarily on the mathematics, while application material will be deferred to exercises and problem sets.

1.1 HDP and its applications

To reiterate, high-dimensional probability refers to the study of probability models that involve either a large number of random variables or random variables that take values in a high-dimensional (linear) space. The overarching goal of this field is to obtain quantitative, nonasymptotic statements that provide explicit information about how probabilities depend on problem parameters.

1.1.1 Phenomena in HDP

There are a number of striking probabilistic phenomena that emerge in high dimensions. Some of these behaviors can be regarded as nonasymptotic analogues of the classical limit theorems, while others are novel contributions of the high-dimensional theory.

- **Concentration.** Even for a complicated random variable, it is often true that the typical values are close to the median or expectation. Concentration theory describes when this situation occurs and it quantifies the probability that the random variable exhibits a large deviation from its mean. These results may be viewed as nonasymptotic refinements of the weak law of large numbers.
- **Suprema.** In many applications, we encounter the supremum of a large number of correlated random variables (aka the supremum of a random process). One of the main tasks in HDP is to develop bounds on the expectation of this type of random variable. These results can be viewed as deep improvements over elementary maximal inequalities.

A *random process* is an indexed family of random variables.

- **Universality.** The central limit theorem shows that a standardized sum converges weakly to a normal distribution under minimal assumptions on the distribution of the summands. More general universality results describe other settings where a complicated random variable has negligible dependence on the distribution of the constituent random variables.
- **Phase transitions.** Many high-dimensional random variables have distinct regimes of behavior. For instance, consider a statistical procedure that succeeds with high probability when it has enough data and fails with high probability when it lacks enough data. This shift in performance occurs over a very small increase in the number of data points. These results are nonasymptotic realizations of 0–1 laws from classical probability theory.

The balance of this lecture provides further details about each one of these phenomena, and it describes general principles that explain when we might expect to witness these outcomes.

This course focuses on concentration phenomena, bounds for suprema, and applications of these bounds to empirical process theory. Lacking sufficient time, we will not cover universality, but there is some related material in the CMS/ACM 117 lecture notes from 2019–2022. We will not treat phase transitions in any depth, but the problem sets will explore one example from statistical signal processing.

1.1.2 Application areas

Over the last two decades, HDP has become an increasingly important tool for a wide range of application areas:

- **Numerical algorithms.** Many contemporary algorithms in numerical linear algebra, numerical analysis, and numerical optimization use randomness for efficiency or robustness. For example, the randomized singular value decomposition algorithm is a very effective method for computing truncated SVDs of large matrices that challenge classical direct or iterative algorithms. Monte Carlo methods are a core technique for computing high-dimensional integrals. Stochastic gradient has become one of the workhorse algorithms in modern machine learning.
- **Randomized algorithms.** HDP tools are also central to the development and analysis of other randomized computer algorithms. For example, methods for hashing use (pseudo)random functions to summarize data. Algorithms for processing streaming data are often based on random projections.
- **Statistics and machine learning.** Probability models for data now involve large numbers of variables and large numbers of observations. Core problems include classification, clustering, feature selection, regression, and model identification. Methods for studying empirical processes also depend heavily on HDP.

Roughly, an *empirical process* is the average value of a function at observed (random) data points.

- **Signal processing.** Signal processing engineers often pose random models for signals or for noise contaminating signals. Modern problems often involve very high-dimensional signals or very large numbers of signals. Applications include detection, estimation, prediction, and filtering.

A *signal* is just a function, with the understanding that the function carries information of some type.

- **Information theory.** Random codes play a key role in proving theorems about channel capacity and compression. Furthermore, while practical codes are structured, they mimic the behavior of random codes.
- **Quantum information science.** According to Born's rule, the observation of a quantum system yields a random outcome. As a consequence, probability is at the heart of quantum mechanics and quantum computing. The number

of degrees of freedom in a quantum system grows very quickly (because it is modeling by a tensor product of subsystems), so HDP methods are relevant.

- **Statistical mechanics.** The basic insight behind statistical modeling of large physical systems (e.g., a volume of gas) is that the overall behavior is often predictable, even if components (e.g., individual molecules) are unpredictable. HDP is relevant because of the large number of particles. Other applications include study of percolation and spin glasses.
- **Large random structures.** HDP also arises in the mathematical study of large random structures, including random matrices and random graphs.
- **Asymptotic convex geometry.** HDP has become a central tool in the study of high-dimensional linear spaces and in high-dimensional convex geometry.

The lecturer's personal interest in HDP arises from its application to randomized numerical algorithms. Since we design the algorithms, our probabilistic analysis provides valid insights on the performance of the algorithm.

Similarly, in statistics, (randomized) experimental design is a critical tool for evaluating competing alternatives (e.g., vaccine versus placebo). In this setting, we can have confidence that we make valid inferences because we control the assignment of subjects to test conditions. Indeed, HDP has always had close connections with statistics.

In contrast, many machine learning problems involve observational data, and we have limited information about the mechanism that generates the data. In these settings, strong assumptions like independence must be regarded with suspicion. It is essential that practitioners validate probability models before trusting the outcomes of mathematical analysis based on these models. These concerns are extra-mathematical, but they do influence the choice of applications that we will discuss.

1.2 Nonasymptotic analysis

A key feature of HDP is the nonasymptotic analysis of large probability models. To appreciate why the classical limit theorems are not adequate, let us recall the statement of the weak law of large numbers.

Theorem 1.1 (Weak law of large numbers (WLLN)). Let X be a real random variable with expectation $\mathbb{E}\, X = m$. Consider an iid sequence $(X_i : i \in \mathbb{N})$ of copies of X. Form the running averages:

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \quad \text{for } n \in \mathbb{N}.$$

Then, for each $t > 0$, we have the limit

$$\mathbb{P}\left\{|\bar{X}_n - m| \geq t\right\} \to 0 \quad \text{as } n \to \infty.$$

Recall that *iid* means "independent and identically distributed." The phrase "X_i is a copy of X" means that X_i has the same distribution as X.

For each level t, the probability that the running average $\bar{X}_n$ deviates from the true expectation $m = \mathbb{E}\, X$ by more than t vanishes as the number of terms n increases. Eventually, we can be as confident we desire that the running average approximates the expectation as well as we require.

Example 1.2 (Survey sampling). The students at Hogwarts are selecting a new student body president. Overall, an (unknown) proportion $p \in [0, 1]$ prefer the candidate from Gryffindor over the candidate from Slytherin. Given the importance of this election,

the Wizard Times sends owls to n random students (chosen with replacement) to ask which candidate they prefer. For $i = 1, \dots, n$, define

$$X_i = \begin{cases} 1, & \text{student } i \text{ prefers Gryffindor;} \\ 0, & \text{student } i \text{ prefers Slytherin.} \end{cases}$$

No value judgment is implied by the definition of X_i.

The sample average $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ is the empirical proportion of surveyed students that prefer the candidate from Gryffindor. If the sample is chosen at random, then $\mathbb{E} X_i = p$ and so $\mathbb{E} \bar{X}_n = p$ as well. For $t > 0$, the WLLN ensures that

$$\mathbb{P}\left\{|\bar{X}_n - p| \geq t\right\} \to 0 \quad \text{as } n \to \infty.$$

That is, if enough students are surveyed, then the empirical proportion of votes for Gryffindor is a good approximation of the true preferences of the entire student body.

A practical issue with this application is that the WLLN is an asymptotic claim. It only tells us what happens in the large-sample limit, whereas there are only a finite number of students to survey (at which point the population parameter p is completely determined). A more useful result would quantify the probability that a sample of size n gives an approximation with error level t to the true proportion p. ▪

This example makes it clear why nonasymptotic statements (n fixed) can be essential for applications. Similar challenges arise many other settings in computational mathematics. For example, a randomized algorithm must terminate after a finite number of steps, so nonasymptotic analyses are more informative than limit theorems.

1.3 Concentration

In high-dimensional probability, the most basic phenomenon is that, under fairly weak assumptions, a random variable typically takes values close to its mean.

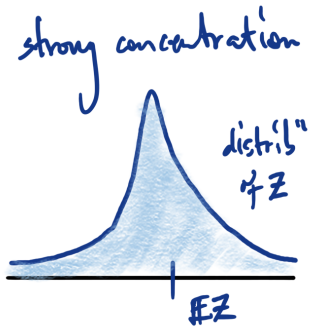
Consider an independent family $(X_1, \dots, X_n)$ of real random variables. For a (measurable) function $f : \mathbb{R}^n \to \mathbb{R}$, we can construct a real random variable

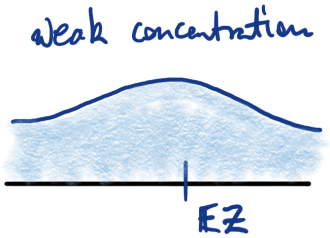
$$Z := f(X_1, \dots, X_n).$$

The random variable Z depends on a large number of independent inputs through the intermediation of the function f, and we can use this structural information to understand the behavior of Z more deeply.

A *concentration inequality* controls the probability that the random variable Z deviates from its median $\mathbb{M} Z$ or its expectation $\mathbb{E} Z$ by a specified amount $t > 0$:

$$\mathbb{P}\{|Z - \mathbb{M} Z| \geq t\} \leq \_\_\_\_\_ \quad \text{or}$$
$$\mathbb{P}\{|Z - \mathbb{E} Z| \geq t\} \leq \_\_\_\_\_.$$





You can think about a concentration inequality as a quantitative version of the WLLN. For contrast, a *tail bound* provides a one-sided deviation inequality (without absolute values); we can combine a lower and an upper tail bound to reach a concentration inequality.

Most elementary probability courses already include some of the fundamental concentration results. First, Chebyshev's inequality states that

$$\mathbb{P}\{|Z - \mathbb{E} Z| \geq t\} \leq \frac{\operatorname{Var}[Z]}{t^2} \quad \text{for } t > 0. \tag{1.1}$$

This result is valid whenever Z has a *finite second moment*. It uses the variance to summarize the fluctuations of the distribution: the smaller the variance, the sharper the concentration. But it only delivers weak tail decay on the order of t^{-2}.

That is, $\mathbb{E}\, Z^2 < +\infty$.

You may also be familiar with the Laplace transform method, which is due to Bernstein. This technique yields a (one-sided) tail bound of the form

This approach is commonly, but inaccurately, referred to the Cramér–Chernoff method or the Hoeffding method.

$$\mathbb{P}\{Z - \mathbb{E}\, Z \geq t\} \leq \inf_{\theta>0} \mathrm{e}^{-\theta t} \cdot \mathbb{E}\, \mathrm{e}^{\theta(Z - \mathbb{E}\, Z)}. \tag{1.2}$$

This result is more limited than (1.1) because it is useful only when the upper tail of Z *decays exponentially*. It uses the moment generating function (mgf) to describe the fluctuations of the distribution, and it can certify the fact that Z has exponential—or better—tail decay.

That is, $\mathbb{P}\{Z \geq t\} \leq C\mathrm{e}^{-ct}$ for all $t > 0$ for certain constants $c, C > 0$.

Chebyshev's inequality (1.1) and the Laplace transform method (1.2) are particularly valuable for independent sums because the variance and the logarithm of the mgf are additive. That is, we specialize to the case where

$$Z = \sum\nolimits_{i=1}^{n} X_i \quad \text{for independent } (X_i).$$

In this situation,

$$\begin{aligned} \operatorname{Var}[Z] &= \sum\nolimits_{i=1}^{n} \operatorname{Var}[X_i], \quad \text{and} \\ \log \mathbb{E}\, \mathrm{e}^{\theta(Z - \mathbb{E}\, Z)} &= \sum\nolimits_{i=1}^{n} \log \mathbb{E}\, \mathrm{e}^{\theta(X_i - \mathbb{E}\, X_i)}. \end{aligned}$$

These facts allow us to exploit information about the individual summands X_i to extract more detailed information about the sum Z. In particular, if each of the summands X_i has controlled variance (or mgf), then so does the sum Z.

For the more general case where $Z = f(X_1, \dots, X_n)$, it is harder to deploy Chebyshev's inequality and the Laplace transform method. To do so, we must deduce bounds on the variance (or mgf) of Z from information about the function f and its random arguments X_i. In the first part of the course, we will develop mathematical techniques that allow us to do so. These results will lead to concentration inequalities for many nonlinear functions.

The basic principle underlying modern concentration theory was enunciated by Michel Talagrand in a 1996 paper [Tal96]:

> "A random variable that depends (in a 'smooth' way) on the influence of many independent variables (but not too much on any of them) is essentially constant."

Our goal will be to understand what this statement means and how to quantify the parentheticals.

1.4 Suprema

Many random variables arising in high-dimensional probability exhibit sharp concentration around the expectation. It is a remarkable fact that we can easily verify concentration properties for many random variables of interest. Nevertheless, concentration inequalities provide no information about the size of the expectation! Obtaining bounds for the expectation typically requires a separate (and more onerous) investigation.

Although we cannot hope to address every possible example, we can specialize our attention to classes of random variables that arise frequently in applications. For an index set T, consider a real-valued random process:

$$\mathcal{X} := (X_t : t \in \mathsf{T}) \in \mathbb{R}^{\mathsf{T}}.$$

In other words, the random process $\mathfrak{X}$ comprises real random variables X_t that are indexed by points in $t \in \mathsf{T}$. The random variables X_t may be correlated, and it is helpful to think about X_t as a smooth function of the index t. From the process, we can construct a real random variable of the form

$$Z := \sup\{X_t : t \in \mathsf{T}\}.$$

For now, let us ignore measurability issues that can arise from this definition.

This type of random variable appears in a huge number of situations. It is relatively easy to obtain concentration inequalities for Z, but studying the behavior of $\mathbb{E}\, Z$ requires a set of deep new ideas.

To indicate how suprema might appear, let us present a simple example involving random matrices.

Example 1.3 (Operator norm of a random matrix). A *random matrix* is a random variable $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ that takes values in a linear space of matrices. Equivalently, $\boldsymbol{X}$ is a rectangular array of real random variables that may or may not be independent. The ℓ_2 operator norm of the random matrix is defined as

$$\|\boldsymbol{X}\|_{\ell_2 \to \ell_2} := \sup\left\{\boldsymbol{u}^* \boldsymbol{X} \boldsymbol{v} : \|\boldsymbol{u}\|_{\ell_2} = 1, \|\boldsymbol{v}\|_{\ell_2} = 1\right\}.$$

A random matrix in captivity:

$$\boldsymbol{X} = \begin{bmatrix} +1 & -1 & -1 & -1 \\ -1 & -1 & -1 & +1 \\ +1 & +1 & -1 & +1 \end{bmatrix}$$

We can regard the norm of the random matrix as the supremum of a random process that is indexed by pairs $t = (\boldsymbol{u}, \boldsymbol{v})$. The index set T is the Cartesian product of two Euclidean unit spheres.

Observe that the family $(\boldsymbol{u}^* \boldsymbol{X} \boldsymbol{v})$ consists of correlated random variables. Indeed, if we change the vector $\boldsymbol{u}$ or $\boldsymbol{v}$ by a small amount, then the bilinear form $\boldsymbol{u}^* \boldsymbol{X} \boldsymbol{v}$ changes by only a small amount. Nevertheless, the bilinear form can vary widely as the index vectors range over the two unit spheres. ▪

Studying the supremum of a random process $\mathfrak{X}$ is challenging. Somehow, we must account for the fact that the process may contain many elements with strong correlation. In other words, if we know the value of one element of the process, then the value of other "nearby" elements is likely to be similar (and in particular not much larger). But more "distant" elements of the process are less correlated, and these may contribute to an increase in the supremum. Thus, the supremum of the process reflects both the rate at which the constituent random variables change and also the size of the index set.

This insight leads to a principle that ultimately goes back to Kolmogorov:

> If the elements of a random process vary in a "smooth" way with the index, then the supremum of the process is controlled by the "complexity" of the index set.

In the second part of the course, we will work to implement the insights behind this statement. The third part of the course contains applications of these ideas to signal processing, statistics, and learning theory.

1.5 Universality

Another fundamental property of probability models in high-dimensions is that the detailed distributions of the constituent random variables have a limited effect on the overall behavior of the model. This idea dramatically generalizes the central limit theorem (CLT).

Let X be a random variable with finite second moment. Consider an iid sequence $(X_i : i \in \mathbb{N})$ of copies of X. The CLT states that appropriately normalized partial sums

of this sequence converge weakly to a normal distribution:

$$n^{-1/2} \sum_{i=1}^{n} X_i \rightsquigarrow \textsc{normal}(\mathbb{E}[X], \operatorname{Var}[X]).$$

In other words, the limit only depends on the distribution of the random variable X through its first and second moment. This is an example of a *universality phenomenon* or an *invariance principle*.

Let us describe part of the argument behind one of the classic proofs of the CLT to see how the result might extend. For a smooth function $f : \mathbb{R} \to \mathbb{R}$, we can expand f in a Taylor series about zero:

$$\left| f(t) - \left[f(0) + f'(0) \cdot t + \tfrac{1}{2} f''(0) \cdot t^2 \right] \right| \le \tfrac{1}{6} \|f'''\|_{\sup} \cdot |t|^3.$$

Now, consider two real random variables X and Y that share the same first and second moment. Applying this formula with both X and Y and taking the expectation, we can reach the bound

$$|\mathbb{E}[f(X) - f(Y)]| \le \tfrac{1}{6} \|f'''\|_{\infty} \cdot \left(\mathbb{E}\,|X|^3 + \mathbb{E}\,|Y|^3 \right).$$

This formula shows that $\mathbb{E}\, f(X) \approx \mathbb{E}\, f(Y)$ for every sufficiently smooth function f. In other words, the typical values of the random variable $f(X)$ only reflect the first two moments of f. With some additional ideas, this argument lifts to multivariate functions. As a particular consequence, this approach leads to a version of the CLT.

We summarize this idea in a general principle, attributed to Lindeberg:

> A random variable that depends in a smooth way on the influence of many independent random variables does not reflect their detailed distributions.

We will not pursue Lindeberg's insight this term, but you may refer to your notes from CMS/ACM 117 for an introduction. Van Handel's notes [van16] also give a brief treatment of these ideas.

1.6 Phase transitions

Last, we describe another phenomenon that occurs in high-dimensional probability models. As you know, water presents in solid form below a temperature of 0° Celsius. Ice melts into water as soon as the temperature becomes positive. It remains liquid until the temperature passes 100° Celsius, when it rapidly changes into a gas.

Similarly, a probability model may exhibit a small number of characteristic behaviors that depend on an underlying model parameter. As the parameter increases, the model stably presents a single behavior. When the parameter passes a threshold, the behavior quickly changes to a new regime. This is called a *phase transition*. It can be viewed as a nonasymptotic counterpart to the 0–1 laws from classical probability.

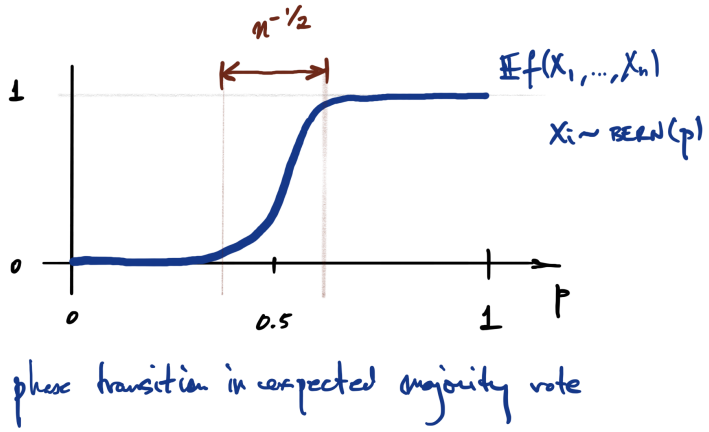
Here is a concrete example arising from survey sampling.

Example 1.4 (Survey sampling). Among the students at Hogwarts, a proportion $p \in [0, 1]$ prefer the Gryffindor candidate to the Slytherin candidate. We interrogate the preferences of n randomly chosen students (with replacement), and we record the iid random variables $X_i \sim \textsc{bernoulli}(p)$. Now, consider whether a strict majority of the sampled students vote in favor of the candidate from Gryffindor:

$$f(X_1, \dots, X_n) = \mathbb{1}\left\{ n^{-1} \textstyle\sum_{i=1}^{n} X_i > \tfrac{1}{2} \right\} = \mathbb{1}\{\text{majority vote for Gryffindor}\}.$$

It is an exercise to compute the expectation of this random variable as a function of p.

Let us present a plot that illustrates the key outcome from this calculation:



In other words, it is extremely unlikely that the majority of the sample prefers Gryffindor when $p \ll \frac{1}{2}$, while it is extremely likely that the majority of the sample prefers Gryffindor when $p \gg \frac{1}{2}$. The key observation is that the transition between the two regimes occurs over the narrow range $p = \frac{1}{2} \pm n^{-1/2}$. The larger the sample size n, the sharper the transition. Indeed, as $n \to \infty$, the smooth curve converges to a step function.

In contrast, one may consider a "junta" function:

$$g(X_1, \ldots, X_n) = \mathbb{1}\{X_1 = 1\}.$$

That is, we record whether the first respondent prefers Gryffindor, and we ignore everyone else's vote. You may confirm that $p \mapsto \mathbb{E}\, g$ is a linear function; it exhibits no sharp change in behavior. ▪

This example generalizes to a wider setting. For Boolean functions, we can formulate a principle that describes when phase transitions take place.

A *Boolean function* is a map $f : \{0,1\}^n \to \{0,1\}$.

> Applied to a family of iid Bernoulli random variables with mean p, a sufficiently symmetric Boolean function exhibits a sharp phase transition as the parameter p increases.

Justifying this claim requires a long excursion into the harmonic analysis of Boolean functions, which would take us too far afield.

Later in this course, we will encounter a beautiful geometric phase transition that occurs in statistical signal processing; this result is an easy consequence of a deep comparison theorem for Gaussian processes. Many other types of high-dimensional probability models exhibit phase transitions, including random graphs, random matrices, spin glasses, and so forth. Unfortunately, we currently lack a unified theory that captures all of these examples.

Notes

This lecture is inspired by Ramon van Handel's notes [van16, Chap. 1].

Lecture bibliography

[Tal96] M. Talagrand. "A new look at independence". In: *Ann. Probab.* 24.1 (1996), pages 1–34.

[van16] R. van Handel. "Probability in High Dimensions". APC 550 Lecture Notes, Princeton Univ. 2016. URL: https://web.math.princeton.edu/~rvan/APC550.pdf.

I. concentration

2. Variance Bounds

Date: 7 January 2021 **Scribe:** Jack Li

Agenda:
1. Concentration, tails, variance
2. Representations and bounds for variance
3. Independent sums
4. Tensorization
5. Efron–Stein–Steele
6. Bounded differences

In this lecture, we begin our discussion of concentration inequalities. As described in Section 2.3, a concentration inequality controls the probability that a random variable deviates from some value (usually its median or expectation) by more than a specified amount.

Most elementary probability courses cover Chebyshev's inequality, which shows how to establish a (weak) concentration inequality for a random variable whose variance is known. In view of this result, it is productive to develop methods for bounding the variance of a random variable.

Recall that the variance of a sum of independent variables is equal to the sum of the variances of the individual terms. We will generalize this property to obtain a bound on the variance of an arbitrary (measurable) function of independent random variables. This important result is called the *tensorization of variance*.

By combining the tensorization result with representations for the variance, we can easily derive the classic Efron–Stein–Steele inequality. By combining tensorization with simple range bounds for the variance, we can derive the bounded differences inequality for the variance. In the next lecture, we will see how these results lead us to variance bounds based on functional inequalities.

2.1 Concentration and tails

Recall that a concentration inequality controls the probability that a real random variable Z deviates from its median $\mathbb{M}\,Z$ or its expectation $\mathbb{E}\,Z$ by more than a specified amount $t > 0$. That is, we seek inequalities of the form

$$\mathbb{P}\{|Z - \mathbb{M}\,Z| \geq t\} \leq \underline{\quad ? \quad} \quad \text{or}$$
$$\mathbb{P}\{|Z - \mathbb{E}\,Z| \geq t\} \leq \underline{\quad ? \quad}.$$

The bounds depend on the structure of the random variable Z, as well as the level t of the deviation from the central tendency. In contrast, a tail bound provides a one-sided deviation inequality (without absolute values); we can use the union bound to combine a lower tail bound and an upper tail bound to reach a concentration inequality.

To motivate our development, we quote the basic principle underlying modern concentration theory:

> "A random variable that depends (in a 'smooth' way) on the influence of many independent variables (but not too much on any of them) is essentially constant."
>
> —Michel Talagrand (1996)

Our goal in the first part of the course is to make sense out of this principle.

To that end, let us consider an independent family $(X_1, \dots, X_n)$ of real random variables. Given a (measurable) function $f : \mathbb{R}^n \to \mathbb{R}$, we can construct the real

random variable

$$Z := f(X_1, \dots, X_n).$$

The random variable Z clearly depends on a large number of independent inputs through the intermediation of the function f. We need to develop techniques for measuring the "smoothness" of the function f, and we must investigate what it means for Z not to depend "too much" on a few inputs. Today's lecture takes the first steps in this program.

2.2 Variance

The simplest and most widely applicable concentration inequalities involve the variance of a random variable.

Definition 2.1 (Variance). Let Z be a real random variable in L_2. The variance of Z is

$$\operatorname{Var}[Z] := \mathbb{E}(Z - \mathbb{E}\, Z)^2 = \mathbb{E}\, Z^2 - (\mathbb{E}\, Z)^2. \tag{2.1}$$

L_2 consists of real random variables that are square-integrable: $\mathbb{E}\, X^2 < +\infty$.

The definition of variance requires us to make a few comments on the notation that we use for expectation.

Notation 2.2 (Expectation). As usual, we abuse notation by writing $\mathbb{E}\, Z$ for the expectation of the random variable Z, which is a constant real number. We abbreviate $\mathbb{E}\, Z^2 := \mathbb{E}[Z^2]$. More generally, nonlinear functions always bind before the expectation.

Given the variance of a random variable, we can bound the tails by means of Chebyshev's inequality.

Proposition 2.3 (Chebyshev). For each real random variable $Z \in \mathsf{L}_2$,

$$\mathbb{P}\{|Z - \mathbb{E}\, Z| \geq t\} \leq \frac{\operatorname{Var}[Z]}{t^2} \quad \text{for all } t > 0. \tag{2.2}$$

Exercise 2.4 (Chebyshev). Prove Proposition 2.3. **Hint:** Use Markov's inequality.

From Chebyshev's inequality, we realize that the variance of a random variable provides information on its concentration about the mean. As a consequence, the variance is a useful summary of how much the random variable deviates from its mean. A useful alternative presentation is

$$\mathbb{P}\{|Z - \mathbb{E}\, Z| \geq t \cdot \operatorname{stdev}[Z]\} \leq 1 \wedge t^{-2}.$$

Recall that the standard deviation, stdev, is the square root of the variance. The wedge $\wedge$ is the infix minimum.

This formulation indicates that the standard deviation is the typical scale on which the random variable fluctuates around its mean. Chebyshev's inequality only produces weak tail decay: the probability of a fluctuation of more than t times the standard deviation is bounded by t^{-2}. We will establish much stronger concentration bounds later in the course.

In the rest of this lecture, we will develop bounds for the variance that can be used in concert with Chebyshev's inequality to obtain concentration inequalities.

2.3 Representations of variance

This section introduces two alternative representations of the variance of a random variable.

The first representation states that the expectation is the point about which a random variable has the minimum expected squared deviation. This result is the starting point for the theory of conditional expectation.

Proposition 2.5 (Variance: Variational formula). Let Z be a real random variable in L_2. Then

$$\operatorname{Var}[Z] = \inf_{a \in \mathbb{R}} \mathbb{E}(Z - a)^2. \tag{2.3}$$

The infimum is attained at $a = \mathbb{E}\, Z$.

Exercise 2.6 (Variance: Variational formula). Establish Proposition 2.5.

The second proposition states that the variance can be written as the expected squared difference between two iid copies of the random variable.

Proposition 2.7 (Variance: Exchangeable pairs). Let $Z, Z' \in \mathsf{L}_2$ be iid real random variables. Then

$$\begin{aligned} \operatorname{Var}[Z] &= \tfrac{1}{2}\, \mathbb{E}(Z - Z')^2 && (2.4)\\ &= \mathbb{E}(Z - Z')_+^2 && (2.5)\\ &= \mathbb{E}(Z - Z')_-^2. && (2.6) \end{aligned}$$

The positive and negative parts of a number are defined as

$$(a)_+ := \max\{a, 0\};$$
$$(a)_- := \max\{-a, 0\}.$$

Note that both are positive!

Exercise 2.8 (Variance: Exchangeable pairs). Prove Proposition 2.7.

As a consequence, for a random variable with small variance, two iid copies tend to be close together in expected mean square. For future reference, we frame a related definition.

Definition 2.9 (Exchangeable pairs). We say that a pair (Z, Z') of random variables is *exchangeable* if (Z, Z') and (Z', Z) have the same distribution.

In particular, when (Z, Z') is exchangeable, both random variables Z and Z' share the same marginal distribution. In Proposition 2.7, since Z and Z' are iid random variables, we can easily confirm that (Z, Z') is an exchangeable pair of random variables. In general, an exchangeable pair need not consist of independent random variables.

Exercise 2.10 (Exchangeable pairs). Find three different examples of an exchangeable pair (Z, Z') where Z is a standard normal random variable.

2.4 Variance bound

Next, we develop a simple bound for the variance of a random variable that takes values in a bounded interval. In this case, the variance is always controlled by the squared length of the interval.

Proposition 2.11 (Variance: Range bound). Let Z be a real random variable in L_2 whose support is contained in the interval $[a, b]$. Then

$$\operatorname{Var}[Z] \le \tfrac{1}{4}(b - a)^2. \tag{2.7}$$

Exercise 2.12 (Variance: Range bound). Establish Proposition 2.11. Show that the bound is saturated by the random variable that places half its mass on each of the two endpoints. **Hint:** Use Proposition 2.5.

The variance bound in Proposition 2.11 does not reflect the "internal" structure of Z. For example, if the random variable places equal mass on the ends of the interval (and nowhere else), then the variance bound is saturated. On the other hand, if the random variable places most of its mass on a small sub-interval, then the variance bound wildly overestimates the true variance.

This discussion suggests that Proposition 2.11 is not particularly useful in isolation. To see how it can lead to more informative results, we will apply the proposition to obtain nontrivial variance bounds for an independent sum of bounded random variables.

2.5 Independent sums

Consider an independent family $(X_1, \dots, X_n)$ of real random variables, each in L_2. Introduce the independent sum

$$Z = \sum\nolimits_{i=1}^{n} X_i \quad \text{with} \quad \mathbb{E}[Z] = \sum\nolimits_{i=1}^{n} \mathbb{E}[X_i]. \tag{2.8}$$

It is well known that the variance of an independent sum is the sum of the variances:

$$\operatorname{Var}[Z] = \sum\nolimits_{i=1}^{n} \operatorname{Var}[X_i]. \tag{2.9}$$

The additivity property isolates the contributions of the individual arguments X_i to the total variance of Z.

Exercise 2.13 (Variance: Independent sum). Confirm (2.9).

Now, let us illustrate the power of the range bound, Proposition 2.11, by applying it to the independent sum (2.8). Assume that each summand X_i has support in an interval $[a_i, b_i]$ for $i = 1, \dots, n$. Abbreviate the length $c_i := |b_i - a_i|$ of the interval, and introduce the vector $\boldsymbol{c} := (c_1, \dots, c_n)$.

For each summand, invoke Proposition 2.11 conditionally. We arrive at the bound

$$\begin{aligned} \operatorname{Var}[Z] &= \sum\nolimits_{i=1}^{n} \operatorname{Var}[X_i] \\ &\le \sum\nolimits_{i=1}^{n} \tfrac{1}{4}(b_i - a_i)^2 = \tfrac{1}{4}\|\boldsymbol{c}\|_2^2. \end{aligned} \tag{2.10}$$

Thus, the standard deviation of the sum satisfies $\operatorname{stdev}[Z] \le \frac{1}{2}\|\boldsymbol{c}\|_2$. We have obtained an elegant bound on the variability of the sum in terms of the ranges of the individual summands.

Exercise 2.14 (Variance of an independent sum: Range bound). Find circumstances where the bound (2.10) holds with equality.

For comparison, let us give a worst-case bound on the range of the independent sum Z:

$$\operatorname{range}[Z] := (\sup Z) - (\inf Z).$$

Using the additive structure of the sum and independence, we may compute that

$$\operatorname{range}[Z] = \sum\nolimits_{i=1}^{n} \operatorname{range}[X_i] \le \sum\nolimits_{i=1}^{n} (b_i - a_i) = \|\boldsymbol{c}\|_1$$

We may now identify circumstances where the standard deviation is much smaller than the range of the independent sum.

Figure 2.1 illustrates two possibilities. When $\|\boldsymbol{c}\|_2 \ll \|\boldsymbol{c}\|_1$, the standard deviation of the sum is much smaller than its range and we obtain sharp concentration around

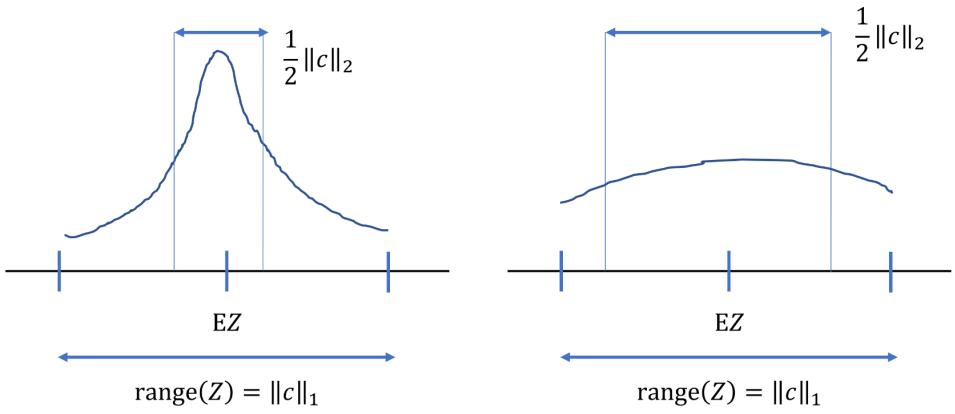


Figure 2.1 (Concentration for an independent sum of bounded random variables). Concentration properties of the sum Z depend on the relation between the bound $\frac{1}{2}\|\boldsymbol{c}\|_2$ for the standard deviation and the bound $\|\boldsymbol{c}\|_1$ for the range. The vector $\boldsymbol{c}$ lists the ranges of the individual summands.

the expectation **[left]**. This situation occurs, for example, when all the individual ranges c_i are similar in magnitude. On the other hand, when $\|\boldsymbol{c}\|_2 \approx \|\boldsymbol{c}\|_1$, the standard deviation of the sum and the range are comparable, and the sum exhibits only weak concentration **[right]**. This situation occurs when a few of the individual ranges c_i are much larger than the others.

This discussion provides a nice illustration of Talagrand's principle. The sum Z is a very smooth function of the independent components X_i. To make sure that the sum concentrates sharply, we need to make sure that it does not depend too much on any of the individual summands. We have quantified how much the sum depends on X_i by means of its range c_i. Our results indicate that we achieve strong concentration for Z when none of the individual summands dominate the sum. Equivalently, none of the c_i is outsized.

2.6 Tensorization of variance

The additivity law (2.9) appears to be a miraculous property of an independent sum. Surprisingly, we can obtain a generalization of the additivity law that holds for a general function of independent random variables. This result is called the *tensorization property* of variance.

As before, consider an independent family $(X_1, \dots, X_n)$ of real random variables. For a (measurable) function $f : \mathbb{R}^n \to \mathbb{R}$, we construct the real random variable

$$Z := f(X_1, \dots, X_n).$$

We always assume that Z is square integrable. Our goal is to control $\mathrm{Var}[Z]$ in terms of the contributions from the individual X_i. To that end, we need to introduce some notation.

Definition 2.15 (Coordinatewise expectation). For $i = 1, \dots, n$, the coordinatewise expectation operator $\mathbb{E}_i$ computes the expectation with respect to X_i, while holding

the remaining random variables $(X_j : j \neq i)$ fixed. That is,

$$\mathbb{E}_i[Z] := \mathbb{E}[Z \mid X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n].$$

This definition implicitly requires that the family (X_i) is independent.

Let us emphasize that $\mathbb{E}_i Z$ is a random variable that depends on $(X_j : j \neq i)$. The coordinatewise expectation operators have some useful algebraic properties that we will exploit heavily.

- **Idempotency.** For each i, we have $\mathbb{E}_i[\mathbb{E}_i Z] = \mathbb{E}_i Z$.
- **Commutativity.** For all i, j, we have $\mathbb{E}_i[\mathbb{E}_j Z] = \mathbb{E}_j[\mathbb{E}_i Z]$.

We encourage the reader to confirm these facts, which depend on the independence assumption.

Definition 2.16 (Coordinatewise variance). The coordinatewise variance $\operatorname{Var}_i[Z]$ of Z with respect to X_i is

$$\operatorname{Var}_i[Z] := \mathbb{E}_i(Z - \mathbb{E}_i Z)^2 = \mathbb{E}_i Z^2 - (\mathbb{E}_i Z)^2.$$

Note that the coordinatewise expectation $\operatorname{Var}_i[Z]$ is a function of $(X_j : j \neq i)$.

We continue to use $\mathbb{E}$ and Var to denote the total expectation and variance functions, which average with respect to all sources of randomness. In this setting, $\mathbb{E} = \mathbb{E}_1 \dots \mathbb{E}_n$.

With these notations, we can state an important theorem.

Theorem 2.17 (Variance tensorizes). With the prevailing notation and assumptions,

$$\operatorname{Var}[Z] \leq \mathbb{E}\left[\sum\nolimits_{i=1}^{n} \operatorname{Var}_i[Z]\right]. \tag{2.11}$$

This result states that we can bound the variance of $Z = f(X_1, \dots, X_n)$ by adding up the variance attributable to each of the individual X_i. The terminology "tensorization" derives from the fact that the probability measure of an independent family $(X_1, \dots, X_n)$ of random variables is the (tensor) product of the marginal distributions. The variance of a function with respect to a product measure is controlled by the sum of the variances with respect to the marginal distributions.

A few more remarks are in order. First, note that the tensorization theorem generalizes our calculation (2.9) for independent sums. Indeed,

$$\operatorname{Var}_i[Z] = \operatorname{Var}_i\left[\sum\nolimits_{j=1}^{n} X_j\right] = \operatorname{Var}[X_i].$$

Thus, for an independent sum, the tensorization inequality for variances holds with equality. This observation also indicates that an independent sum is the *worst case* for tensorization; it is the function where the fluctuation due to each individual random variable is highest.

Proof. The proof uses a Doob martingale to decompose the variance into contributions from the individual random variables. This part of the argument is standard. Introduce a collection $(Y_0, Y_1, \dots, Y_n)$ of random variables:

$$Y_i := \mathbb{E}[Z \mid X_1, \dots, X_i] = \mathbb{E}_n \mathbb{E}_{n-1} \dots \mathbb{E}_{i+1}[Z] \quad \text{for } i = 0, \dots, n.$$

In particular, $Y_0 = \mathbb{E} Z$ and $Y_n = Z$.

As an exercise, confirm that (Y_i) is a (Doob) martingale with respect to the natural filtration $\mathcal{F}_i := \sigma(X_1, \dots, X_i)$:

$$\mathbb{E}[Y_{i+1} \,|\, X_1, \dots, X_i] = Y_i \quad \text{for } i = 0, \dots, n-1.$$

Define the martingale differences $\Delta_i := Y_i - Y_{i-1}$ for $i = 1, \dots, n$. We can express the martingale property as

$$\mathbb{E}[\Delta_{i+1} \,|\, X_1, \dots, X_i] = 0 \quad \text{for } i = 0, \dots, n-1.$$

It is a standard fact that martingale differences are orthogonal:

$$\mathbb{E}[\Delta_i \Delta_j] = 0 \quad \text{for all } i \neq j.$$

You should check this claim as well.

Using the martingale, we can write the variance of Z as a telescoping sum:

$$\begin{aligned}
\operatorname{Var}[Z] &= \mathbb{E}(Z - \mathbb{E}\, Z)^2 \\
&= \mathbb{E}(Y_n - Y_0)^2 = \mathbb{E}\big((Y_n - Y_{n-1}) + (Y_{n-1} - Y_{n-2}) + \cdots + (Y_1 - Y_0)\big)^2 \\
&= \mathbb{E}\left(\sum\nolimits_{i=1}^{n} \Delta_i\right)^2 = \sum\nolimits_{i=1}^{n} \mathbb{E}[\Delta_i^2].
\end{aligned}$$

The last relation follows because the martingale differences are zero mean and orthogonal.

To continue, we will rewrite this expression to isolate the contributions from each of the random variables X_i. Instead of using conditional expectations, it is more transparent to work with the coordinatewise expectations:

$$\begin{aligned}
\operatorname{Var}[Z] &= \sum\nolimits_{i=1}^{n} \mathbb{E}[\Delta_i^2] \\
&= \sum\nolimits_{i=1}^{n} \mathbb{E}\big((\mathbb{E}_n \cdots \mathbb{E}_{i+1}\, Z) - (\mathbb{E}_n \cdots \mathbb{E}_{i+1}\, \mathbb{E}_i\, Z)\big)^2 \\
&\overset{\text{(a)}}{=} \sum\nolimits_{i=1}^{n} \mathbb{E}\big(\mathbb{E}_n \cdots \mathbb{E}_{i+1}(Z - \mathbb{E}_i\, Z)\big)^2 \\
&\overset{\text{(b)}}{\leq} \sum\nolimits_{i=1}^{n} \mathbb{E}\, \mathbb{E}_n \cdots \mathbb{E}_{i+1}(Z - \mathbb{E}_i\, Z)^2 \\
&\overset{\text{(c)}}{=} \sum\nolimits_{i=1}^{n} \mathbb{E}\, \mathbb{E}_i (Z - \mathbb{E}_i\, Z)^2 \\
&= \sum\nolimits_{i=1}^{n} \mathbb{E} \operatorname{Var}_i[Z].
\end{aligned}$$

Step (a) comes from the linearity of conditional expectation, and step (b) comes from Jensen's inequality and the convexity of the square function. Step (c) comes from the representation $\mathbb{E} = \mathbb{E}_n \dots \mathbb{E}_1$ of the total expectation and the fact that $(\mathbb{E}_i : i = 1, \dots, n)$ is a family of commuting idempotent operators. Finally, we recognize the coordinate variance. ∎

2.7 Efron–Stein–Steele inequality

Tensorization is very powerful when combined with the variance representations and bounds that we have already studied. In this section, we combine tensorization with the exchangeable representation of variance to derive the Efron–Stein–Steele inequality.

We maintain the same notation from the previous section. In addition, we introduce an independent copy (X_i') of the original independent sequence (X_i). That is, each X_i'

is an independent copy of X_i, independent from everything else. Define the random variables

$$Z^{(i)} := f(X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n) \quad \text{for } i = 1, \dots, n.$$

Therefore,, $Z^{(i)}$ is an exchangeable counterpart of Z obtained by refreshing the ith coordinate X_i with a new draw X_i' from the same distribution.

We may now state the following corollary of the tensorization theorem.

Corollary 2.18 (Efron–Stein–Steele inequality). With prevailing notation and assumptions,

$$\operatorname{Var}[Z] \le \frac{1}{2} \mathbb{E}\left[\sum\nolimits_{i=1}^{n} (Z - Z^{(i)})^2\right]. \tag{2.12}$$

Proof. Apply the variational representation of the variance (Proposition 2.7) conditionally to see that

$$\operatorname{Var}_i[Z] = \tfrac{1}{2} \mathbb{E}_i (Z - Z^{(i)})^2.$$

Combine with the tensorization of variance (Theorem 2.17). ∎

The Efron–Stein–Steele (ESS) inequality shows that we can control the variance of Z by measuring how much $Z = f(X_1, \dots, X_n)$ changes on average when we refresh the ith coordinate. When Z is insensitive to changes in each coordinate, we can think of it as a smooth function of the inputs X_i. Moreover, we anticipate that the variance of Z is small relative to the range of Z when none of the coordinatewise variances dominates.

The ESS inequality has a distinguished pedigree in statistics and combinatorics. Here are high-level descriptions of some classic applications:

- **Jackknife estimator of variance.** The jackknife estimate of a parameter is obtained by averaging estimates obtained from subsamples that omit one variable. The ESS inequality was developed to analyze the behavior of the jackknife.
- **Bin packing with random weights.** In a bin packing problem, items of different weights must be packed into containers, each with a maximum weight limit, in a way that minimizes the number of containeds used. We can study the variance of the number of bins needed by asking how the number of bins changes by replacing each item with an item drawn from the same distribution.
- **Random graphs.** In a graph coloring problem, we color the vertices of a graph such that no two adjacent vertices (sharing the same edge) have the same color. The smallest number of colors needed to color a graph is called its *chromatic number*. We can study the variance of the chromatic number of a random graph by calculating how the chromatic number changes when we randomly add and remove an edge. Although we can bound the variance, it is generally hard to compute the expected chromatic number.
- **Random traveling salesman problem.** The traveling salesman problem asks us to find the shortest path that visits each city exactly once and returns to the original city, given a list of cities and the distance between each two cities. For random city locations, we can study how the length of the optimal path changes when we relocate one city. Although we can bound the variance, it is generally hard to compute the expected length of a traveling salesman path.
- **Maxima and minima of random processes.** Some examples appear in the problem set.

2.8 Bounded differences

Finally, let us show how we can combine the range bound for variance with the tensorization inequality to obtain a nonlinear analog of our range bound (2.10) for an independent sum.

As usual, let $Z = f(X_1, \dots, X_n)$ be a function of independent random variables. We define the ith discrete "partial derivative" as

$$\begin{aligned}(\mathrm{D}_i f)(X_1, \dots, X_n) &\coloneqq (\operatorname{range}_i f)(X_1, \dots, X_n) && (2.13)\\ &\coloneqq \sup_{x \in \operatorname{supp}(X_i)} f(X_1, \dots, X_{i-1}, x, X_{i+1}, \dots, X_n) && (2.14)\\ &\quad - \inf_{x \operatorname{supp}(X_i)} f(X_1, \dots, X_{i-1}, x, X_{i+1}, \dots, X_n). && (2.15)\end{aligned}$$

Note that $\mathrm{D}_i f$ is a random variable that depends on $(X_j : j \neq i)$.

We have the following corollary of the tensorization of the variance. Compare with (2.10).

Corollary 2.19 (Bounded differences). With prevailing notation and assumptions,

$$\operatorname{Var}[Z] = \operatorname{Var}[f] \leq \frac{1}{4} \mathbb{E}\left[\sum\nolimits_{i=1}^{n} (\mathrm{D}_i f)^2\right]. \tag{2.16}$$

Proof. Apply the range bound for the variance (Proposition 2.11) conditionally to see that

$$\operatorname{Var}_i[Z] \leq \tfrac{1}{4}(\mathrm{D}_i f)^2.$$

Combine with the tensorization of variance (Theorem 2.17). ∎

Exercise 2.20 (Bounded difference: Saturation). Show that the bound in Corollary 2.19 can hold with equality by considering the case of an independent sum.

The results in this section and the previous section show that changes of individual coordinates of $f(X_1, \dots, X_n)$ control the variance of f. In other words, we are bounding the variance using some type of discrete derivative, which reflects the smoothness of f. One may wonder whether it is possible to obtain estimates in terms of the ordinary (calculus) derivative. In the next lecture, we will explore this prospect.

3. Poincaré Inequalities

Date: 12 January 2021 **Scribe:** Nicholas H. Nelsen

Agenda:
1. Poincaré intuition
2. Wirtinger's inequality
3. Spectral interpretation
4. Dynamical interpretation
5. Gaussian Poincaré inequality

In the last lecture, we saw that the variance of a function of independent random variables can be controlled in terms of the "discrete partial derivatives" of the function. In this lecture, we will explore analogs involving the ordinary calculus derivative.

This shift in perspective leads us to study an important class of functional inequalities, known as *Poincaré inequalities*. Such bounds play an important role in the modern study of concentration phenomena and have wide-ranging applicability to the theory of Markov chains and (stochastic) partial differential equations. It is not always possible to obtain a Poincaré inequality for a general probability measure. When it is possible, it often requires some *ad hoc* analysis.

We begin our study with intuitive reasoning and then describe the simplest Poincaré inequality, Wirtinger's inequality for the uniform measure on the torus. Next, we give spectral and dynamical interpretations of this result. Finally, we conclude with a discussion of the powerful Gaussian Poincaré inequality.

In this lecture, we write $\mathsf{L}_2(\mu)$ for the space $\mathsf{L}_2(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \mu)$ of real-valued random variables on the probability space $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \mu)$ with finite second moments, where $\mathscr{B}(\mathbb{R}^n)$ is Borel σ-algebra on $\mathbb{R}^n$.

3.1 Motivation for Poincaré inequalities

Consider an independent family $(X_1, \ldots, X_n)$ of real-valued random variables, and let $f : \mathbb{R}^n \to \mathbb{R}$ be a Borel measurable function. Define the real-valued random variable

$$Z := f(X_1, \ldots, X_n).$$

Recall the *bounded differences inequality*, which may be interpreted as a bound on the variance of Z in terms of the sum of squared first differences or "discrete derivatives" of f:

$$\mathrm{Var}[Z] \leq \frac{1}{4} \mathbb{E}\left[\sum_{i=1}^n |\mathrm{D}_i f|^2 \right],$$

where

$$\mathrm{D}_i f := \sup_{x \in \mathrm{supp}(X_i)} f(X_1, \ldots, X_{i-1}, x, X_{i+1}, \ldots, X_n) - \inf_{x \in \mathrm{supp}(X_i)} f(X_1, \ldots, X_{i-1}, x, X_{i+1}, \ldots, X_n).$$

While aesthetically nice and easy to prove using our previous results on tensorization and variance bounds for L_∞ random variables, the bounded difference inequality has some limitations. First and foremost, the result only holds for functions f bounded almost surely. For example, it does not apply to polynomials in Gaussian random variables. Second, the definition of D_i is in terms of extrema and hence does not take into account finer information about the fluctuations of X_i in each coordinate of Z.

One could imagine extending this idea of first differences to the continuum, bounding the variance of Z by the expected squared Euclidean norm of the gradient of f, sometimes called the "Dirichlet energy." In this lecture, we turn toward the study of such functional inequalities. In particular, *Poincaré inequalities* control the average size of a function by the average size of its derivative.

In what follows, we emphasize functional notation by writing f in place of Z. For example, $\mathrm{Var}[Z] = \mathrm{Var}[f]$, where it is understood that these statements are the same. To emphasize the underlying measure μ, we use $\mathrm{Var}_\mu[Z] = \mathrm{Var}_\mu[f]$.

3.2 Poincaré inequality: Uniform distribution on the torus

Wirtinger's inequality is one of the technically simplest examples of a Poincaré inequality. To avoid complications, we will focus on periodic functions, even though this setting is a little artificial. Problem 3.7 describes a more applicable result.

3.2.1 Functions on the torus

Introduce the one-dimensional torus:

$$\mathbb{T} := \mathbb{R}/\mathbb{Z} \cong [0, 1).$$

Since the torus is isomorphic to the unit circle, it carries a uniform distribution. We can think about $\textsc{uniform}(\mathbb{T})$ as the restriction of the Lebesgue measure to $[0, 1)$.

A function $f : \mathbb{T} \to \mathbb{R}$ can be viewed as a function on $[0, 1)$, extended periodically to the real line. We define the class of continuously differentiable functions on the torus:

$$\mathsf{C}^1(\mathbb{T}) := \{f : \mathbb{T} \to \mathbb{R} : f, f' \text{ continuous}\}.$$

To clarify, a continuous periodic function f must be continuous across the right endpoint: $\lim_{t\to 1} f(t) = f(0)$. The derivative f' of a function on the torus is the ordinary derivative of the periodic extension to the real line, including at the right endpoint $t = 1$ of the fundamental region $[0, 1)$.

3.2.2 Wirtinger's inequality and its consequences

Now, consider a continuously differentiable function $f : \mathbb{T} \to \mathbb{R}$ that has zero mean:

$$\int_0^1 f(x)\,\mathrm{d}x = 0. \tag{3.1}$$

If f is not identically zero, then it must have nontrivial positive and negative parts. That is, the function must oscillate above and below the real axis. For this behavior to occur, the *derivative* f' must be large on average, as compared with f. We formalize this intuition in the following theorem.

Theorem 3.1 (Wirtinger). Let $f \in \mathsf{C}^1(\mathbb{T})$ be a continuously differentiable function on the torus, and assume that f has zero mean (3.1). Then

$$\int_0^1 |f(x)|^2\,\mathrm{d}x \le \frac{1}{(2\pi)^2} \int_0^1 |f'(x)|^2\,\mathrm{d}x.$$

Warning 3.2 (Wirtinger: Hypotheses). Theorem 3.1 requires the assumptions that the function f is periodic and continuously differentiable. We can weaken these assumptions at the cost of increasing the Poincaré constant (see Problem 3.7). ▪

We can reformulate Wirtinger's inequality into probabilistic language, where it is called a *Poincaré inequality* for the uniform distribution on the torus.

Corollary 3.3 (Poincaré: Uniform on torus). Let $\mu := \textsc{uniform}(\mathbb{T})$. Consider a random variable $X \sim \mu$ and a function $f : \mathsf{C}^1(\mathbb{T}) \to \mathbb{R}$. Then

$$\operatorname{Var}[f(X)] \le \frac{1}{(2\pi)^2} \mathbb{E}\left[|f'(X)|^2\right].$$

We need not assume $\mathbb{E} f = 0$ in Corollary 3.3, since replacing $f \mapsto f - \mathbb{E} f$ does not change either side of the display in Corollary 3.3. We also allow the right hand side to be infinite.

We say that $1/(2\pi)^2$ is the *Poincaré constant* of $\textsc{uniform}(\mathbb{T})$ for continuously differentiable functions.

By tensorization, we can extend Corollary 3.3 to multivariate functions on a product of tori. Note that these functions are periodic with respect to each coordinate. For example, $\mathbb{T}^2$ is isomorphic to a donut.

Corollary 3.4 (Multivariate Poincaré: Uniform on torus). Consider a family $(X_1, \dots, X_n)$ of iid $\textsc{uniform}(\mathbb{T})$ random variables. Let $f : \mathbb{T}^n \to \mathbb{R}$ be a continuous function on the n-fold product of the torus, where the first-order partial derivatives of f are continuous. Then the random variable $Z = f(X_1, \dots, X_n)$ satisfies the inequality

$$\operatorname{Var}[f] \le \frac{1}{(2\pi)^2} \mathbb{E}\left[\sum_{i=1}^n |\partial_i f|^2\right] = \frac{1}{(2\pi)^2} \mathbb{E}\left[\|\nabla f\|_2^2\right],$$

where $\|\cdot\|_2$ is the ordinary ℓ_2-norm of a vector.

Problem 3.5 (Multivariate Poincaré: Uniform on torus). Prove Corollary 3.4.

We can be simplified this result further if we further assume that f is L-Lipschitz with respect to the ℓ_2 norm on the product $\mathbb{T}^n$ of tori. That is,

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \le L \cdot \|\boldsymbol{x} - \boldsymbol{y}\|_{\ell_2(\mathbb{T}^n)} \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{T}^n.$$

Let us emphasize that arithmetic with vectors in $\mathbb{T}^n$ is performed modulo $\mathbb{Z}$. Rademacher's theorem states that the Lipschitz function f satisfies

$$\|\nabla f(\boldsymbol{x})\|_2^2 \le L^2 \quad \text{(Lebesgue) almost everywhere on } \mathbb{T}^n.$$

Hence, by Corollary 3.4,

$$\operatorname{Var}[f] \le \frac{L^2}{(2\pi)^2}.$$

Thus we see that (periodic, continuous) Lipschitz functions of uniform random variables have controlled variance, and hence they *always* exhibit concentration.

Warning 3.6 (Poincaré inequalities depend on the distribution). These results may not hold when the random variables $(X_1, \dots, X_n)$ are drawn from a distribution that is not uniform, and they also depend on the class of functions that we consider. ▪

Problem 3.7 (Wirtinger: Minimal assumptions). Consider the Sobolev space $\mathsf{H}^1[0,1]$ of weakly differentiable functions with $f, f' \in \mathsf{L}_2[0,1]$. For this class, prove that

$$\max\left\{\|f - m_f\|_2 : \|f'\|_2 \le 1 \text{ and } f(1/2) = 0\right\} = \frac{1}{\pi^2},$$

where $m_f := \int_0^1 f(x)\,\mathrm{d}x$ is the mean. The maximizers are $f(x) = \pm\sqrt{2}\cos(\pi x)/\pi$. **Hint:** Use calculus of variations. First, develop a prior bound for the maximal value. To establish that the maximizer exists, select a maximizing sequence and apply the Rellich–Kondrakov theorem. Last, use the Lagrange multiplier theorem to find all stationary points.

Deduce that the optimal Poincaré inequality for the uniform random variable $X \sim \textsc{uniform}[0,1]$ reads

$$\operatorname{Var}[f(X)] \le \frac{1}{\pi^2}\,\mathbb{E}[f'(X)^2] \quad \text{for } f \in \mathsf{H}^1[0,1].$$

For an independent family $(X_1, \dots, X_n)$ of copies of X, conclude that

$$\operatorname{Var}[f(X_1,\dots,X_n)] \le \frac{1}{\pi^2}\,\mathbb{E}[\|\nabla f(X_1,\dots,X_n)\|_2^2] \quad \text{for } f \in \mathsf{H}^1([0,1]^n).$$

In particular, when f is L-Lipschitz, we have $\operatorname{Var}[f] \le L^2/\pi^2$.

3.2.3 Proof of Theorem 3.1

We now prove Wirtinger's inequality using Fourier analysis, omitting some technical details about convergence of Fourier series. For example, see [Wei12] for a classical treatment or [Rob20] for a general Hilbert space viewpoint.

Proof. **(Wirtinger's inequality).** Let $\mu := \textsc{uniform}(\mathbb{T})$. Recall that the system

$$\big(\sqrt{2}\sin(2\pi kx), \sqrt{2}\cos(2\pi kx) : k \in \mathbb{N}\big),$$

together with the constant function 1, composes a complete orthonormal system in $\mathsf{L}_2(\mu)$.

Consider a function $f \in \mathsf{C}^1(\mathbb{T})$ with zero mean (3.1). We may expand f in a Fourier series

$$f(x) = \sum\nolimits_{k=1}^{\infty} [a_k \cdot \sqrt{2}\cos(2\pi kx) + b_k \cdot \sqrt{2}\sin(2\pi kx)],$$

where the coefficients $((a_k, b_k) : k \in \mathbb{N})$ may be found using orthonormality of the sines and cosines. Note that the constant term, corresponding to the zero wavenumber $k = 0$, is equal to zero because we have assumed f has zero mean! The convergence of this series is understood in the L_2 sense.

Differentiating the series term-by-term using continuity of the derivative, we obtain

$$f'(x) = \sum\nolimits_{k=1}^{\infty} 2\pi k \cdot [-a_k \cdot \sqrt{2}\sin(2\pi kx) + b_k \cdot \sqrt{2}\cos(2\pi kx)].$$

By two applications of Parseval's identity,

$$\begin{aligned}\int_0^1 |f'(x)|^2\,\mathrm{d}x &= \sum\nolimits_{k=1}^{\infty} (2\pi k)^2 \left[a_k^2 + b_k^2\right] \\ &\ge (2\pi)^2 \sum\nolimits_{k=1}^{\infty} \left[a_k^2 + b_k^2\right] \\ &= (2\pi)^2 \int_0^1 |f(x)|^2\,\mathrm{d}x.\end{aligned}$$

This is the advertised inequality. Observe that equality holds if and only if f is proportional to $\sin(2\pi x)$ or $\cos(2\pi x)$. ∎

3.3 Spectral interpretation

The sharp Wirtinger's inequality of the previous section is fundamentally a functional analytic result that fits into the more general picture of the spectral theory of linear operators [Rob20]. We study a particular linear operator that leads to the same sines and cosines we saw previously, but emphasize that the conceptual approach is much more general. We borrow some notation from the paper [COS20].

As before, we restrict our attention to real-valued functions defined on the torus $\mathbb{T}$. That is, $f : \mathbb{T} \to \mathbb{R}$ formally satisfies *periodic boundary conditions*

In a sense, "periodic boundary conditions" are not really boundary conditions.

$$f(0) = f(1) \quad \text{and} \quad f'(0) = f'(1).$$

Further assume these functions are mean zero, as in (3.1). Define the linear space

$$\dot{\mathsf{L}}_2(\mathbb{T}^1; \mathbb{R}) := \left\{ f : \mathbb{T} \to \mathbb{R} : \int_0^1 |f(x)|^2 \, \mathrm{d}x < \infty, \quad \int_0^1 f(x) \, \mathrm{d}x = 0 \right\}.$$

This linear space is equipped with the $\mathsf{L}_2(\mathbb{T})$ inner product $\langle \cdot, \cdot \rangle$ and the associated norm $\|\cdot\|$ to form a Hilbert space of functions.

Define the linear operator $\boldsymbol{A}$ whose domain $\mathcal{D}(\boldsymbol{A})$ is a dense subset of $\dot{\mathsf{L}}_2(\mathbb{T}; \mathbb{R})$ by

$$\begin{aligned} \boldsymbol{A} : \mathcal{D}(A) &\to \dot{\mathsf{L}}_2(\mathbb{T}; \mathbb{R}) \\ f &\mapsto \boldsymbol{A}f := -\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}. \end{aligned}$$

This operator is simply the negative second derivative, or *Laplacian*, equipped with periodic boundary conditions and restricted to mean zero functions. Under these conditions, the operator $\boldsymbol{A}$ is positive definite, with orthonormal eigenvalue–eigenfunction pairs $\{(\lambda_k, \varphi_k)\}_{k=1}^{\infty}$ given by

Analogous results hold for higher spatial dimensions.

$$\begin{aligned} \varphi_{2j}(x) = \sqrt{2} \cos(2\pi j x) \quad \text{and} \quad \varphi_{2j-1}(x) = \sqrt{2} \sin(2\pi j x); \\ \lambda_{2j} = \lambda_{2j-1} = (2\pi)^2 j^2 > 0. \end{aligned}$$

for $j \in \mathbb{N}$. The system $(\varphi_k : k \in \mathbb{N})$ composes an orthonormal basis for $\dot{\mathsf{L}}_2(\mathbb{T}; \mathbb{R})$. Hence, we may expand any $f \in \dot{\mathsf{L}}_2(\mathbb{T}; \mathbb{R})$ into a series

$$f = \sum_{k=1}^{\infty} f_k \varphi_k, \quad \text{where} \quad f_k = \langle \varphi_k, f \rangle.$$

By choosing the domain of $\boldsymbol{A}$ to be

$$\mathcal{D}(\boldsymbol{A}) := \left\{ f \in \dot{\mathsf{L}}_2(\mathbb{T}; \mathbb{R}) : \sum_{k=1}^{\infty} k^4 f_k^2 < \infty \right\},$$

one obtains a consistent definition.

We are now in a position to make the connection to Wirtinger's inequality. Letting $f \in \mathcal{D}(\boldsymbol{A})$ and integrating by parts,

$$\int_0^1 |f'(x)|^2 \, \mathrm{d}x = -\int_0^1 f(x) f''(x) \, \mathrm{d}x = \langle f, \boldsymbol{A}f \rangle$$

by applying the periodic boundary conditions. Therefore, the Rayleigh quotient satisfies

$$\frac{\langle f, \boldsymbol{A}f \rangle}{\langle f, f \rangle} \geq \min_{g \in \mathcal{D}(\boldsymbol{A})} \frac{\langle g, \boldsymbol{A}g \rangle}{\langle g, g \rangle} = \lambda_{\min}(\boldsymbol{A}) = (2\pi)^2,$$

where we have used the fact that A is positive definite so the minimum of the Rayleigh quotient is the smallest eigenvalue of $\boldsymbol{A}$. This is precisely Wirtinger's inequality:

$$\|f\|^2 = \langle f, f \rangle \leq \frac{1}{(2\pi)^2} \langle f, \boldsymbol{A}f \rangle = \frac{1}{(2\pi)^2} \|f'\|^2.$$

Aside: In spectral graph theory, there are analogous results known as *spectral gap inequalities* that involve the smallest nonzero eigenvalue of the positive semidefinite graph Laplacian, which is often the second eigenvalue $\lambda_2 > 0 = \lambda_1$; hence the terminology "spectral gap." These tools find extensive applications in the study of Markov chains.

3.4 Dynamical interpretation

Poincaré inequalities also find application in a dynamical setting. Consider the *heat* or *diffusion equation* framed on the torus:

$$\begin{cases} \dfrac{\partial u}{\partial t} = \dfrac{\partial^2 u}{\partial x^2}, & \text{on } \mathbb{R} \times \mathbb{T}; \\ u(0, x) = f(x), & \text{for } x \in \mathbb{T}; \\ u : \mathbb{R} \times \mathbb{T} \to \mathbb{R}. \end{cases}$$

This initial value problem can be succinctly written as an abstract equation on function space, using notation of the previous section:

$$\partial_t u = -\boldsymbol{A}u, \; u(0) = f.$$

Note that the periodic boundary conditions are implicit in the requirement that $u(t, \cdot)$ is a function on the torus.

This is a basic linear partial differential equation arising in various fields of science, engineering, and finance. It enjoys a canonical physical interpretation as a model for the diffusion of temperature along a one-dimensional metal ring. For a more probabilistic perspective, one may view the initial data f to be a probability density: $f \geq 0$ a.e. and $\int_0^1 f(x)\,\mathrm{d}x = 1$, and thus the heat equation can be viewed as a model for the evolution of this probability density under a specific stochastic process (i.e., the Brownian motion diffusion process).

If f has Fourier series representation

$$f(x) = a_0 + \sum_{k=1}^{\infty} [a_k \cdot \sqrt{2}\cos(2\pi k x) + b_k \cdot \sqrt{2}\sin(2\pi k x)],$$

then one can show by inspection or separation of variables that

$$u(t, x) = a_0 + \sum_{k=1}^{\infty} \mathrm{e}^{-(2\pi)^2 k^2 t} \cdot [a_k \cdot \sqrt{2}\cos(2\pi k x) + b_k \cdot \sqrt{2}\sin(2\pi k x)].$$

This series representation (see Figure 3.1) implies that high frequencies in the initial condition f are damped *very quickly*, while the lower frequency components of f take longer to decay in time. With $\mu := \textsc{uniform}(\mathbb{T})$, we can invoke $\mathsf{L}_2(\mu)$ orthonormality to obtain the bound

$$\int_0^1 |u(t, x) - a_0|^2\,\mathrm{d}x \leq \mathrm{e}^{-(2\pi)^2 t} \int_0^1 |f(x) - a_0|^2\,\mathrm{d}x.$$

Succinctly,

$$\mathrm{Var}_\mu[u(t, \cdot)] \leq \mathrm{e}^{-(2\pi)^2 t}\,\mathrm{Var}_\mu[u(0, \cdot)].$$

We conclude that the rate of decay of the variance of solutions to the heat equation on the torus is exponential and is controlled by the variance of the initial condition. Note that the Poincaré constant $1/(2\pi)^2$ controls the exponential rate of decay of the variance.

Aside: The simple dynamical interpretation as presented here has much wider scope in the areas of stochastic processes and (potentially stochastic) partial differential equations), especially in the semigroup theory of these topics (see, e.g., [PS08;

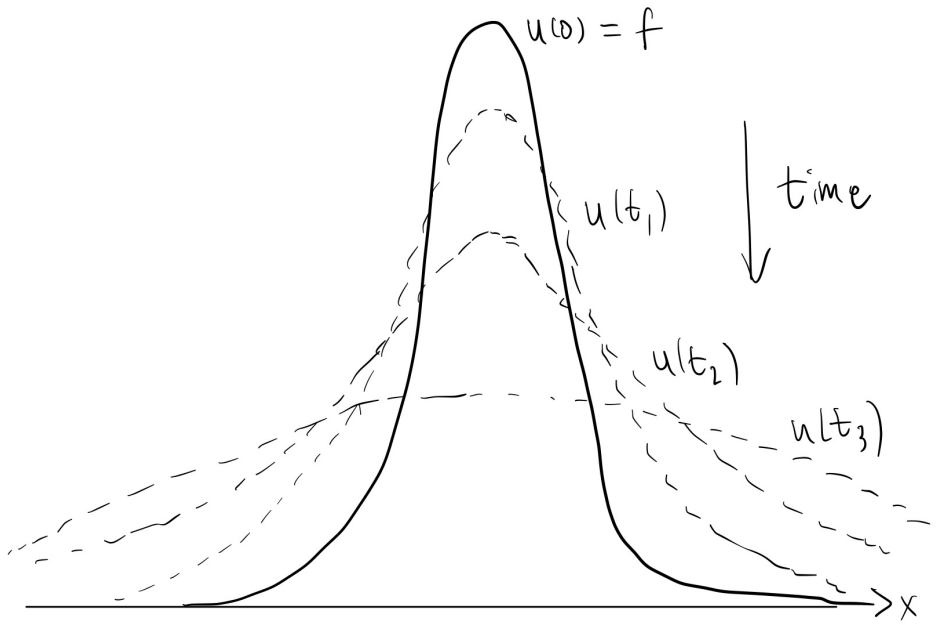


Figure 3.1 Evolution of the solution to the heat equation with initial condition f.

Pazı2]) or the convergence of Markov chains to stationary distributions.

3.5 The Gaussian Poincaré inequality

One may wonder whether it is possible to establish Poincaré inequalities for other probability distributions. Indeed, there are many such examples. We conclude this lecture with a fundamental Poincaré inequality for the Gaussian distribution. We will prove this result using Hilbert space methods, similar in spirit to the arguments behind Wirtinger's inequality.

Theorem 3.8 (Gaussian Poincaré). Let $\gamma := \textsc{normal}(0, 1)$. Draw a random variable $Z \sim \gamma$, and consider a differentiable function $f : \mathbb{R} \to \mathbb{R}$ with $f, f' \in \mathsf{L}_2(\gamma)$. Then

$$\operatorname{Var}[f(Z)] \le \mathbb{E}[|f'(Z)|^2].$$

So, the Poincaré constant of $\textsc{normal}(0, 1)$ is 1.

Proof. Without loss of generality, we may assume that f has mean zero: $\mathbb{E} f(Z) = 0$. Define the (probabilist's) Hermite polynomials $(H_n)_{n=0}^{\infty}$ by

$$H_n(x) := (-1)^n \mathrm{e}^{x^2/2} \frac{\mathrm{d}^n}{\mathrm{d}x^n} \mathrm{e}^{-x^2/2} \quad \text{for } x \in \mathbb{R}.$$

The physicist's Hermite polynomials are a rescaled version of (H_n) obtained by replacing $\mathrm{e}^{-x^2/2} \mapsto e^{-x^2}$.

The system $(H_n : n \in \mathbb{Z}_+)$ compmoses an orthogonal basis for $\mathsf{L}_2(\gamma)$. For all $n, m \in \mathbb{Z}_+$,

$$\mathbb{E}[H_n(Z) H_m(Z)] = \langle H_n,\ H_m \rangle_{\mathsf{L}_2(\gamma)} = \int_{\mathbb{R}} H_n H_m \,\mathrm{d}\gamma = n!\, \delta_{nm}.$$

Here, δ_{nm} is the Kronecker delta.

Hence, the function $f \in \mathsf{L}_2(\gamma)$ admits the expansion

$$f = \sum\nolimits_{k=1}^{\infty} a_k H_k \quad \text{where} \quad a_k = \frac{1}{k!} \langle H_k,\ f \rangle_{\mathsf{L}_2(\gamma)}.$$

Since f has mean zero and $H_0 = 1$, it follows that the coefficient $a_0 = 0$. Using the fact that

$$H_n' = nH_{n-1}$$

and assuming the $(a_k)_{k=1}^{\infty}$ satisfy a sufficient summability condition, we compute

$$\begin{aligned}
\int_{\mathbb{R}} |f'|^2 \,\mathrm{d}\gamma &= \sum\nolimits_{k=1}^{\infty} k^2 a_k^2 \int_{\mathbb{R}} |H_{k-1}|^2 \,\mathrm{d}\gamma \\
&= \sum\nolimits_{k=1}^{\infty} k a_k^2 \cdot k(k-1)! \\
&\geq \sum\nolimits_{k=1}^{\infty} a_k^2 \cdot k! \\
&= \sum\nolimits_{k=1}^{\infty} a_k^2 \int_{\mathbb{R}} |H_k|^2 \,\mathrm{d}\gamma \\
&= \int_{\mathbb{R}} |f|^2 \,\mathrm{d}\gamma.
\end{aligned}$$

This is the required result. ∎

Similarly to Corollary 3.4, we have the following generalization to multivariate functions.

Corollary 3.9 (Multivariate Gaussian Poincaré). Let $\gamma_n \coloneqq \textsc{normal}(\mathbf{0}, \mathbf{I}_n)$. Consider the random vector $\boldsymbol{z} \sim \gamma_n$ and a differentiable function $f \in \mathsf{L}_2(\gamma_n)$. Then

$$\operatorname{Var}[f(\boldsymbol{z})] \leq \mathbb{E}\left[\|\nabla f(\boldsymbol{z})\|_2^2\right].$$

Exercise 3.10 (Multivariate Gaussian Poincaré). Prove Corollary 3.9.

From our previous discussion, it follows that a Lipschitz function of a Gaussian random vector has controlled variance. More precisely, assume that

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L \cdot \|\boldsymbol{x} - \boldsymbol{y}\|_2 \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

Then we have the elegant variance bound

$$\operatorname{Var}[f(\boldsymbol{z})] \leq L^2 \quad \text{where } \boldsymbol{z} \sim \gamma^n.$$

This result quantifies the concentration properties of a Lipschitz function of a standard normal variable.

Aside: The Gaussian Poincaré inequality also enjoys spectral and dynamical interpretations, but we need to replace the Laplacian with another differential operator. Consider the Ornstein–Uhlenbeck linear differential operator $\boldsymbol{A}$, defined by

$$(\boldsymbol{A}u)(x) \coloneqq -\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}(x) + x\left(\frac{\mathrm{d}u}{\mathrm{d}x}(x)\right)$$

for all $u \in \mathscr{D}(\boldsymbol{A})$ and $x \in \mathbb{R}$. Connecting to probability, $\boldsymbol{A}$ is (up to constants) the infinitesimal generator of the *Orstein–Uhlenbeck semigroup*.

On an appropriate Gaussian Hilbert space, the eigenfunctions of the operator $\boldsymbol{A}$ are the Hermite polynomials. The pair $(1, H_1)$ is the minimum eigenvalue and eigenfunction of the operator $\boldsymbol{A}$, restricted to zero mean functions. This is the spectral analog of the fact that the Gaussian Poincaré constant is one.

We may consider the dynamical flow

$$\partial_t u = -\boldsymbol{A}u, \quad u(0) = f$$

This equation models the evolution of a particle diffusing on the real line under the influence of friction.

on an appropriate Gaussian Hilbert state space. With $\gamma := \text{NORMAL}(0, 1)$, the solution variance satisfies

$$\text{Var}_\gamma[u(t, \cdot)] \leq \mathrm{e}^{-t} \text{Var}_\gamma[u(0, \cdot)].$$

This formula gives a dynamical interpretation of the Gaussian Poincaré inequality.

Lecture bibliography

[COS20] Y. Chen, H. Owhadi, and A. M. Stuart. "Consistency of Empirical Bayes And Kernel Flow For Hierarchical Parameter Estimation". In: *arXiv preprint arXiv:2005.11375* (2020).

[PS08] G. Pavliotis and A. Stuart. *Multiscale methods: averaging and homogenization*. Springer Science & Business Media, 2008.

[Paz12] A. Pazy. *Semigroups of linear operators and applications to partial differential equations*. Springer Science & Business Media, 2012.

[Rob20] J. C. Robinson. *An Introduction to Functional Analysis*. Cambridge University Press, 2020.

[Wei12] H. F. Weinberger. *A first course in partial differential equations: with complex variables and transform methods*. Courier Corporation, 2012.

4. Exponential Concentration

Date: 14 January 2021 **Scribe:** Ziyun Zhang

Agenda:
1. Laplace transform method
2. Hoeffding
3. Chernoff
4. Bernstein

Last time, we established some Poincaré inequalities, and we used them to bound the variance of a nonlinear function of independent random variables. In this lecture, we begin our study of exponential concentration inequalities. Today, we will consider the case of an independent sum, and next time we will start to develop results on the concentration of nonlinear functions.

Let Z be a real random variable in L_2. Chebyshev's inequality shows that

$$\mathbb{P}\{|Z - \mathbb{E}\,Z| \geq t \cdot \operatorname{stdev}[Z]\} \leq 1 \wedge t^{-2} \quad \text{for all } t > 0.$$

In other words, the typical scale for fluctuations around the mean is the standard deviation, $\operatorname{stdev}[Z] := \sqrt{\operatorname{Var}[Z]}$. By bounding the variance, we can obtain results on the concentration of Z. Unfortunately, Chebyshev's inequality only yields weak tail decay of t^{-2}. Can we do better?

To understand the prospects, consider an independent family $(X_1, \dots, X_n)$ of real random variables in L_2, and define the independent sum $Z = \sum_{i=1}^n X_i$. By additivity of variance,

$$\operatorname{Var}[Z] = \sum\nolimits_{i=1}^n \operatorname{Var}[X_i].$$

This result allows us to decompose the variance of the sum into terms that reflect the variability of individual summands. When the individual variances are small in comparison with the total, we anticipate that the sum concentrates on a scale that is small relative to the possible range of the random variable. By tensorization, similar results hold for nonlinear random variables.

The central limit theorem states that a (standardized) sum of iid random variables in L_2 converges weakly to a standard normal distribution. The tails of the standard normal distribution decay at a rate of $\exp(-t^2/2)$. This rate is dramatically better than what we achieve from Chebyshev's inequality. One may wonder whether it is possible to obtain concentration inequalities for a sum that reflect that rapid tail decay of the normal distribution. In fact, this feat is possible if we are willing to make stronger assumptions on the summands.

In this lecture, we will develop the Laplace transform method, an analog of Chebyshev's inequality that can certify exponential tail decay for a random variable. We will show how to apply the Laplace transform method to a sum of independent, *bounded* random variables. Depending on the assumptions that we pose on the summands, we obtain several excellent tail bounds for the sum. In the next lecture, we will extend this approach to nonlinear functions.

4.1 Laplace transform method

The basic approach to developing exponential concentration inequalities is called the Laplace transform method.

4.1.1 Moments and cumulants

To develop these ideas, we introduce two functions that package up information about the tail decay of a random variable.

Definition 4.1 (Moment generating function; cumulant generating function). Let X be a real random variable. The *moment generating function (mgf)* of X is

$$m_X(\theta) := \mathbb{E}\, \mathrm{e}^{\theta X} \quad \text{for} \quad \theta \in \mathbb{R}.$$

The *cumulant generating function function (cgf)* of X is

$$\xi_X(\theta) := \log \mathbb{E}\, \mathrm{e}^{\theta X} = \log m_X(\theta) \quad \text{for} \quad \theta \in \mathbb{R}.$$

The mgf and cgf are well defined for all $\theta \in \mathbb{R}$, but they can take the value $+\infty$.

In practice, the mgf and cgf are only interesting for random variables where one of the tails decays exponentially. For example,

$$\mathbb{P}\{X \geq t\} \leq C \cdot \mathrm{e}^{-ct} \quad \text{for all } t > 0.$$

In this case, the mgf and cgf finite for a strictly positive value of the parameter θ.

Example 4.2 (Bernoulli distribution: Mgf and cgf). Let $X \sim \textsc{bernoulli}(p)$ be a Bernoulli random variable with expectation $p \in [0, 1]$. The mgf and cgf satisfy

$$m_X(\theta) = 1 + (\mathrm{e}^{\theta} - 1)p \quad \text{and} \quad \xi_X(\theta) \leq (\mathrm{e}^{\theta} - 1)p \quad \text{for all } \theta \in \mathbb{R}.$$

These results follow from an easy calculation. ∎

Example 4.3 (Normal distribution: Mgf and cgf). Let $X \sim \textsc{normal}(0, \sigma^2)$ be a normal random variable with mean zero and variance σ^2. The mgf and cgf take the form

$$m_X(\theta) = \mathrm{e}^{\theta^2\sigma^2/2} \quad \text{and} \quad \xi_X(\theta) = \tfrac{1}{2}\theta^2\sigma^2 \quad \text{for all } \theta \in \mathbb{R}.$$

This result requires a nontrivial calculation. ∎

Aside: You may wonder why the mgf and cgf are referred to as generating functions. By a formal Taylor expansion,

$$m_X(\theta) = \sum\nolimits_{p=0}^{\infty} \frac{\theta^p}{p!} \cdot \mathbb{E}\, X^p.$$

As a consequence, the derivatives of the mgf at zero report the polynomial moments:

$$\frac{\mathrm{d}^p}{(\mathrm{d}\theta)^p}\bigg|_{\theta=0} m_X(\theta) = \mathbb{E}\, X^p \quad \text{for each } p \in \mathbb{N}.$$

In other words, the mgf is the exponential generating function of the sequence $(\mathbb{E}\, X^p : p \in \mathbb{N})$ of polynomial moments.

Similarly, we can expand the cgf is the exponential generating function of a sequence of numbers, called *cumulants*:

$$\xi_X(\theta) = \sum\nolimits_{p=1}^{\infty} \frac{\theta^p}{p!} \cdot \kappa_p(X).$$

> The numbers κ_p are called the *cumulants* or *semi-invariants* of X. In particular, we have $\kappa_0(X) = 0$ and $\kappa_1(X) = \mathbb{E}\, X$, and $\kappa_2(X) = \mathrm{Var}[X]$. The first p cumulants can be written as a polynomial in the first p polynomial moments and vice versa. Although less familiar, the cumulants have very elegant mathematical properties.

4.1.2 Cgfs and tails

We can obtain a bound for the tail decay of a real random variable in terms of its cgf.

Proposition 4.4 (Laplace transform method). Let X be a real random variable. Then

$$\mathbb{P}\{X \geq t\} \leq \inf_{\theta>0} \mathrm{e}^{-\theta t + \xi_X(\theta)} = \exp\big(-\sup_{\theta>0}(\theta t - \xi_X(\theta))\big),$$
$$\mathbb{P}\{X \leq t\} \leq \inf_{\theta<0} \mathrm{e}^{-\theta t + \xi_X(\theta)} = \exp\big(-\sup_{\theta<0}(\theta t - \xi_X(\theta))\big).$$

Proof. Fix a parameter $\theta > 0$. Since $t \mapsto \mathrm{e}^{\theta t}$ is strictly increasing,

$$\mathbb{P}\{X \geq t\} = \mathbb{P}\left\{\mathrm{e}^{\theta X} \geq \mathrm{e}^{\theta t}\right\}.$$

Since $\mathrm{e}^{\theta t}$ and $\mathrm{e}^{\theta X}$ are positive, we can apply Markov's inequality to obtain

$$\mathbb{P}\left\{\mathrm{e}^{\theta X} \geq \mathrm{e}^{\theta t}\right\} \leq \mathrm{e}^{-\theta t}\, \mathbb{E}\, \mathrm{e}^{\theta X}.$$

Optimize over $\theta > 0$ to arrive at the upper tail. ∎

Exercise 4.5 (Lower tails). Establish the lower tail bound in Proposition 4.4.

Example 4.6 (Normal distribution: Tails). We can invoke the Laplace transform method to obtain an elegant tail bound for a normal random variable $X \sim \textsc{normal}(0, \sigma^2)$. Indeed, using Example 4.3,

$$\mathbb{P}\{X \geq t\sigma\} \leq \exp\big(-\sup_{\theta>0}(\theta t\sigma - \tfrac{1}{2}\theta^2\sigma^2)\big) = \mathrm{e}^{-t^2/2} \quad \text{for } t > 0.$$

The infimum is attained at $\theta = t/\sigma$. We have obtained a very good bound for the upper tail, showing normal-type decay on the scale of the standard deviation σ. A similar bound applies to the lower tail of the normal distribution. ▪

4.1.3 Additivity of the cgf

The Laplace transform method is an excellent tool for studying independent sums because the cgf of an independent sum is *additive*.

Proposition 4.7 (The cgf is additive). Consider an independent family $(X_1, \dots, X_n)$ of real random variables, and define $Z = \sum_{i=1}^n X_i$. Then

$$\xi_Z(\theta) = \sum\nolimits_{i=1}^n \xi_{X_i}(\theta) \quad \text{for all } \theta \in \mathbb{R}.$$

Proof. Observe that

$$m_Z(\theta) = \mathbb{E}\, \mathrm{e}^{\theta \sum_i X_i} = \mathbb{E}\left[\prod\nolimits_i \mathrm{e}^{\theta X_i}\right] = \prod\nolimits_i \mathbb{E}\, \mathrm{e}^{\theta X_i} = \prod\nolimits_i m_{X_i}(\theta).$$

We have used the fact that the expectation of a product of independent random variables equals the product of the expectations. Finally, take the logarithm and recognize the cgfs. ∎

Example 4.8 (Binomial distribution: Tails). For a simple illustration of Proposition 4.7, let us consider a binomial random variable $Z \sim \textsc{bin}(n, p)$ with mean $\mathbb{E}\, Z = np$. Since we can write Z as a sum of n iid copies of the random variable $X \sim \textsc{bernoulli}(p)$,

$$\xi_Z(\theta) = n\xi_X(\theta) \le (\mathrm{e}^\theta - 1) \cdot np \quad \text{for all } \theta \in \mathbb{R}.$$

We have used Example 4.2. Invoking the Laplace transform method,

$$\begin{aligned}\mathbb{P}\{Z - np \ge t \cdot np\} &\le \exp\left(-\sup_{\theta>0}(tnp - (\mathrm{e}^\theta - 1) \cdot np)\right) \\ &= \left(\frac{\mathrm{e}^t}{(1+t)^{1+t}}\right)^{np} \quad \text{for } t \ge 0.\end{aligned}$$

This is called the Chernoff bound for the binomial tail. It gives decay on the scale np of the mean at the gamma rate t^{-t}, which is slightly faster than exponential. A related result holds for the lower tail. ∎

More generally, by combining the Laplace transform method and the additivity of cgfs, we can control the tails of an independent sum using *local* information about the summands. To pursue this approach, it suffices to produce upper bounds for the cgfs of the summands. By careful engineering, this procedure leads to simple and appealing concentration inequalities. In this lecture, we study three main examples:

1. **Hoeffding's inequality.** Designed for bounded summands.
2. **Chernoff's inequality.** Designed for bounded *positive* summands.
3. **Bernstein's inequality.** Designed for bounded summands with *small variance*.

Although we focus on bounded summands, many of these results are valid in more general scenarios.

4.1.4 Cramér's theorem

The Laplace transform method is more than a clever trick. It actually produces *sharp* asymptotic bounds for iid sums. Let us discuss this point briefly.

Consider an iid family $(X_1, \dots, X_n)$ of copies of a real random variable X whose mgf is finite for some $\theta > 0$. Combining Propositions 4.4 and 4.7, we arrive at the bound

$$\mathbb{P}\left\{\frac{1}{n}\sum\nolimits_{i=1}^n X_i \ge t\right\} \le \exp\big(-n \cdot \sup_{\theta>0}(\theta t - \xi_X(\theta))\big).$$

When $t > \mathbb{E}\, X$, you may verify that the supremum occurs for strictly positive θ, so we can relax the range of the supremum to $\theta \in \mathbb{R}$. Let us introduce the *rate function* of the random variable X:

$$\Lambda_X(t) := \sup_{\theta\in\mathbb{R}} (\theta t - \xi_X(\theta)).$$

The rate function is called the *Fenchel–Legendre conjugate* of the cgf; it is always a lower-semicontinuous convex function. With this notation, we can rewrite the last display as

$$\frac{1}{n}\log\mathbb{P}\left\{\frac{1}{n}\sum\nolimits_{i=1}^n X_i \ge t\right\} \le -\Lambda_X(t) \quad \text{for } t > \mathbb{E}\, X.$$

Cramér's theorem asserts that the last bound is asymptotically sharp (under minimal conditions on X):

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\left\{\frac{1}{n}\sum\nolimits_{i=1}^n X_i \ge t\right\} = -\Lambda_X(t) \quad \text{for } t > \mathbb{E}\, X.$$

In other words, the rate function describes the *exact* asymptotic behavior of an iid sum.

4.2 Concentration for bounded summands

In this section, we develop our first packaged concentration inequality. This result, attributed to Hoeffding, shows that an independent sum of bounded random variables exhibits normal concentration.

4.2.1 Hoeffding's cgf bound

The key ingredient in the proof of Hoeffding's inequality is a cgf bound for a bounded random variable. We will state and prove a sharp version of this result using an integral representation for the cgf and an exponential tilting of the probability distribution.

Lemma 4.9 (Hoeffding cgf bound). Assume that $a \le X \le b$. Then

$$\xi_{X - \mathbb{E}X}(\theta) \le \tfrac{1}{8}\theta^2 (b-a)^2.$$

Proof. Without loss of generality, we may assume that $\mathbb{E}X = 0$. Fix the parameter $\theta \in \mathbb{R}$. Write the cgf as an integral:

$$\xi_X(\theta) = \log \int_a^b \mathrm{e}^{\theta x} \,\mathrm{d}\mu_X,$$

where μ_X is the distribution of X, supported on $[a, b]$. Observe that $\xi_X(0) = \xi_X'(0) = 0$ because $\mathbb{E}X = 0$. We will control the cgf by bounding its second derivative and then integrating the resulting inequality.

Let us compute the first derivative of ξ_X with respect to θ.

$$\xi_X'(\theta) = \frac{\int_a^b x\mathrm{e}^{\theta x} \,\mathrm{d}\mu_X}{\int_a^b \mathrm{e}^{\theta x} \,\mathrm{d}\mu_X}.$$

We have used bounded convergence to pass the derivative through the integral. Now, define the *tilted probability distribution*:

$$\mathrm{d}\nu_\theta(x) := \frac{\mathrm{e}^{\theta x} \,\mathrm{d}\mu_X(x)}{\int_a^b \mathrm{e}^{\theta y} \,\mathrm{d}\mu_X(y)}.$$

Observe that $\int_a^b \mathrm{d}\nu_\theta = 1$, so ν_θ is another probability measure supported on $[a, b]$. With this new notation, we have

$$\xi_X'(\theta) = \int_a^b x \,\mathrm{d}\nu_\theta$$

In other words, $\xi_X'(\theta)$ is the mean of the tilted distribution ν_θ.

Let us compute the second derivative of $\xi_X(\theta)$. We have

$$\xi_X''(\theta) = \frac{\int_a^b x^2\mathrm{e}^{\theta x} \,\mathrm{d}\mu_X}{\int_a^b \mathrm{e}^{\theta x} \,\mathrm{d}\mu_X} - \left(\frac{\int_a^b x\mathrm{e}^{\theta x} \,\mathrm{d}\mu_x}{\int_a^b \mathrm{e}^{\theta x} \,\mathrm{d}\mu_x}\right)^2 = \int_a^b x^2 \,\mathrm{d}\nu_\theta - \left(\int_a^b x \,\mathrm{d}\nu_\theta\right)^2.$$

In the second line, we have recognized integrals with respect to the tilted probability measure ν_θ. As a consequence, $\xi_X''(\theta)$ is the variance of the distribution ν_θ, supported on $[a, b]$. Using the interpretation as a variance, we derive that

$$0 \le \xi_X''(\theta) \le \tfrac{1}{4}(b-a)^2,$$

using the range bound for the variance (Proposition 2.11) from Lecture 2.

Finally, we use this result to bound the cgf itself. Invoking the fundamental theorem of calculus twice,

$$\begin{aligned}\xi_X(\theta) &= \int_0^\theta \int_0^s \xi_X''(t)\,\mathrm{d}t\,\mathrm{d}s \\ &\le \int_0^\theta \int_0^s \tfrac{1}{4}(b-a)^2\,\mathrm{d}t\,\mathrm{d}s = \tfrac{1}{8}\theta^2(b-a)^2.\end{aligned}$$

We rely on the facts that $\xi_X(0) = \xi_X'(0) = 0$. ∎

Exercise 4.10 (Hoeffding cgf). Try to develop a shorter proof of Lemma 4.9 with the constant 1 in place of the sharp constant $1/8$.

4.2.2 Hoeffding's inequality

With this cgf bound at hand, we can easily produce a concentration inequality for an independent sum of bounded random variables.

Theorem 4.11 (Hoeffding). Consider an independent family $(X_1, \dots, X_n)$ of real random variables that satisfy $a_i \le X_i \le b_i$ for $i = 1, \dots, n$. Construct the sum $Z = \sum_{i=1}^n X_i$, and introduce the variance proxy

$$v := \frac{1}{4}\sum_{i=1}^n (b_i - a_i)^2.$$

Then

$$\mathbb{P}\left\{|Z - \mathbb{E}\,Z| \ge t\sqrt{v}\right\} \le 2\mathrm{e}^{-t^2/2} \quad \text{for all } t \ge 0.$$

Proof. By linearity of expectation,

$$Z - \mathbb{E}\,Z = \sum_{i=1}^n (X_i - \mathbb{E}\,X_i).$$

By the additivity of cgfs (Proposition 4.7) and the Hoeffding cgf bound (Lemma 4.9), we have

$$\xi_{Z-\mathbb{E}\,Z}(\theta) = \sum_{i=1}^n \xi_{X_i - \mathbb{E}\,X_i}(\theta) \le \frac{1}{8}\theta^2 \sum_{i=1}^n (b_i - a_i)^2 = \frac{\theta^2}{2} v.$$

By the Laplace transform method (Proposition 4.4),

$$\begin{aligned}\mathbb{P}\left\{Z - \mathbb{E}\,Z \ge t\sqrt{v}\right\} &\le \inf_{\theta>0} \exp(-\theta t\sqrt{v} + \xi_{Z-\mathbb{E}\,Z}(\theta)) \\ &\le \inf_{\theta>0} \exp(-\theta t\sqrt{v} + \theta^2 v/2) = \mathrm{e}^{-t^2/2}.\end{aligned}$$

The infimum is achieved when $\theta = t/\sqrt{v}$. The lower tail bound follows from a similar argument. Combine the two results using a union bound. ∎

To understand the Hoeffding inequality, recall that the bounded difference inequality for the variance yields the comparison $\mathrm{Var}[Z] \le v$. Thus, we can interpret the variance proxy v as an upper bound that stands in for the variance. The Hoeffding inequality shows that the independent sum Z has tail decay at least as fast as a normal random variable with variance v.

For comparison, the central limit theorem (CLT) suggests that the distribution of Z should be close to a normal variable with variance $\mathrm{Var}[Z]$, assuming that the

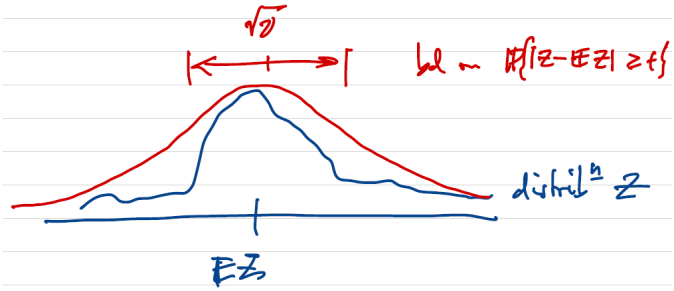


Figure 4.1 (Hoeffding). The Hoeffding bound shows that the tails of an independent sum decay as fast as the tails of a normal random variable with variance v. The number v is called the variance proxy, and it reflects the sum of the squared ranges of the summands.

summands have comparable size. Nevertheless, the CLT does not give good control on the tails, so the concentration inequality gives a different perspective on the behavior of an independent sum.

4.3 Concentration for bounded, positive summands

Next, we turn to a concentration inequality, due to Chernoff, that is suitable for random variables that are bounded and positive. A new feature of this result is that the scale for concentration depends on the size of the expectation, rather than the variance.

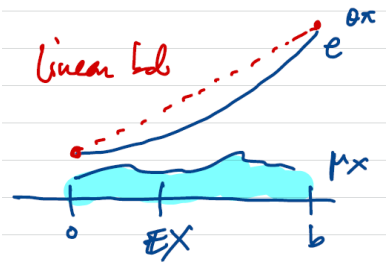
4.3.1 Chernoff's cgf bound

The main task in proving Chernoff's inequality is to develop an appropriate cgf bound. This argument will be based on a graphical estimate for the exponential function.

Lemma 4.12 (Chernoff cgf bound). Assume that $0 \le X \le b$. Then

$$\xi_X(\theta) \le \frac{\mathrm{e}^{\theta b} - 1}{b} \cdot (\mathbb{E}\, X) \quad \text{for all } \theta \in \mathbb{R}.$$

Proof. The idea of the proof is best conveyed graphically. Owing to the convexity of the function $x \mapsto \mathrm{e}^{\theta x}$, we can bound it above on the interval $[0, b]$ by a straight line connecting the endpoints:



Converting this diagram into algebraic terms, we see that

$$m_X(\theta) = \mathbb{E}\left[\mathrm{e}^{\theta X}\right] \le \mathbb{E}\left[1 + \frac{\mathrm{e}^{\theta b} - 1}{b} \cdot X\right] = 1 + \left(\frac{\mathrm{e}^{\theta b} - 1}{b}\right) \cdot \mathbb{E}\, X.$$

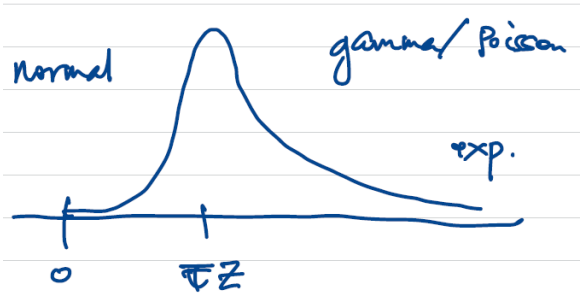


Figure 4.2 (Chernoff). The Chernoff inequalities show that an independent sum of bounded, positive random variables has tails that decay at a rate similar to a gamma random variable. The left tail has a normal decay, while the right tail has gamma decay.

The bracket is the point–slope form of the linear bound. Using the fact that $\log(1+a) \leq a$ for $a > -1$, we conclude that

$$\xi_X(\theta) = \log m_X(\theta) \leq \left(\frac{\mathrm{e}^{\theta b} - 1}{b}\right) \cdot (\mathbb{E}\, X).$$

This is the Chernoff cgf bound. ∎

4.3.2 Chernoff's inequalities

With this cgf bound at hand, we can obtain a concentration inequality for an independent sum of bounded, positive random variables. This result is often used to count the number of independent events that occur by applying it to a sum of indicator random variables. Compare this result with Example 4.8.

Theorem 4.13 (Chernoff inequalities). Consider an independent family $(X_1, \dots, X_n)$ of positive random variables that satisfy $0 \leq X_i \leq b$ for each $i = 1, \dots, n$. Construct the sum $Z = \sum_{i=1}^{n} X_i$. Then

Note that all summands are bounded above by the same number b.

$$\mathbb{P}\{Z \geq (1+t)(\mathbb{E}\, Z)\} \leq \left(\frac{\mathrm{e}^{t}}{(1+t)^{1+t}}\right)^{(\mathbb{E}\, Z)/b} \quad \text{for} \quad t > 0,$$

$$\mathbb{P}\{Z \leq (1-t)(\mathbb{E}\, Z)\} \leq \left(\frac{\mathrm{e}^{-t}}{(1-t)^{1-t}}\right)^{(\mathbb{E}\, Z)/b} \quad \text{for} \quad t \in [0, 1].$$

Exercise 4.14 (Chernoff inequalities). Prove Theorem 4.13.

The Chernoff inequality yields *gamma concentration*. That is, the tails of a sum of independent bounded random variables look like the tails of a gamma random variable (Figure 4.2). The left tail behaves like a normal distribution $\mathrm{e}^{-t^2/2}$, while the right tails decays like t^{-t}, which is slightly faster than exponential.

The asymmetry stems from the positivity of the summands X_i, which ensures that the sum always increases as we add more terms. As a consequence, it is very hard for the sum to be close to zero, which pushes the mass upward toward the mean. In contrast, the sum can become quite large if a few summands take unusually large values, which is why the upper tail bound is weaker.

4.4 Concentration for bounded summands with small variance

Last, we develop a concentration inequality due to Bernstein that is suitable for bounded summands that have small variance.

4.4.1 Bernstein's cgf bound

As usual, our duty is to obtain a cgf bound for a single random variable. This argument is based on a Taylor expansion of the exponential.

Lemma 4.15 (Bernstein cgf bound). Suppose that X is a centered, bounded random variable: $\mathbb{E}\, X = 0$ and $|X| \le b$ almost surely. Then

$$\xi_X(\theta) \le \frac{(\theta^2/2)\operatorname{Var}[X]}{1 - b|\theta|/3} \quad \text{for all } \theta \in \mathbb{R}.$$

Proof. By Taylor expansion, we have

$$\begin{aligned} m_X(\theta) = \mathbb{E}[\mathrm{e}^{\theta X}] &= 1 + \theta \cdot (\mathbb{E}\, X) + \sum\nolimits_{p \ge 2} \frac{\theta^p}{p!} \mathbb{E}\, X^p \\ &\le 1 + \sum\nolimits_{p \ge 2} \frac{|\theta|^p}{p!} (\mathbb{E}\, X^2) \cdot b^{p-2} \\ &\le 1 + \left(\frac{\theta^2}{2} \operatorname{Var}[X] \right) \cdot \sum\nolimits_{p=0}^{\infty} \frac{|\theta|^p}{3^p} b^p \\ &= 1 + \frac{(\theta^2/2)\operatorname{Var}[X]}{1 - b|\theta|/3}. \end{aligned}$$

Taking the logarithm and using the fact that $\log(1 + a) \le a$ for $a > -1$, we obtain the Bernstein cgf bound. ∎

4.4.2 Bernstein's inequality

We may now state Bernstein's inequality for an independent sum. This result is probably the single most useful concentration inequality for an independent sum. It is widely applicable, and it provides accurate bounds for moderate deviations. There are many extensions to this result that weaken the hypotheses.

Theorem 4.16 (Bernstein inequality). Consider an independent family $(X_1, \dots, X_n)$ of bounded, real random variables: $|X_i - \mathbb{E}\, X_i| \le b$ for each $i = 1, \dots, n$. Construct the sum $Z = \sum_{i=1}^n X_i$. Then

$$\mathbb{P}\{|Z - \mathbb{E}\, Z| \ge t\} \le 2 \exp\left(\frac{-t^2/2}{\operatorname{Var}[Z] + bt/3} \right) \quad \text{for } t > 0.$$

Exercise 4.17 (Bernstein inequality). Prove Theorem 4.16. **Hint:** Choose the parameter $\theta = t/(\operatorname{Var}[Z] + bt/3)$.

To understand Bernstein's inequality, we consider two parameter regimes. For moderate t, we obtain normal concentration. For large t, we obtain exponential concentration. See Figure 4.3. For small or medium t, the tail bound looks like $\exp(-t^2/(2\operatorname{Var}[Z]))$, so it yields normal concentration determined by the actual variance of the sum. (In contrast, the variance proxy in Hoeffding's inequality depends on the ranges of the random variables, and it can be far larger than the true variance.) For large t, the tail bound looks like $\exp(-2t/(3b))$, which gives exponential decay

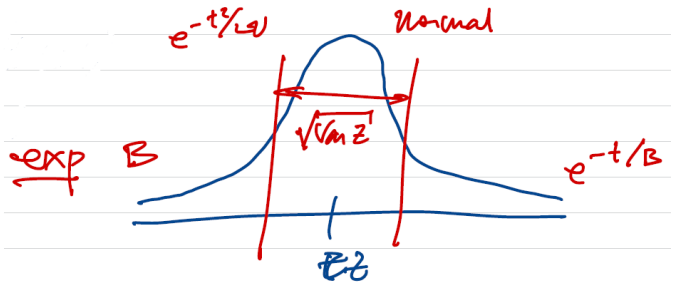


Figure 4.3 (Bernstein). Bernstein's inequality shows that an independent sum of bounded random variables has normal tail decay on the scale of the variance for moderate deviations. For large deviations, the sum exhibits exponential tail decay on the scale of the upper bound for the summands.

on the scale of the upper bound b. These large deviations are driven by the occasional situations where a few summands take unusually large values. These phenomena can be observed empirically.

5. Entropy and Concentration

Date: 19 January 2021 **Scribe:** Sarina Liu

Agenda:
1. Concentration entropy
2. Herbst argument
3. Entropy tensorizes
4. Entropy bounds
5. Bounded differences

Last time, we introduced exponential tail bounds for independent sums based on the Laplace transform method. Today, we will begin to develop exponential concentration inequalities for nonlinear functions of independent random variables.

Consider the random variable $Z = f(X_1, ..., X_n)$, where the family (X_i) is independent. Recall that the variance provides a way of understanding the fluctuations around the mean. Using the Chebyshev inequality, we can get weak concentration bounds from the variance. However, we want a stronger exponential concentration for Z so that we have greater control over the tail decay.

To move from independent sums, where the variance is additive, to nonlinear functions, we find inspiration in the fact that variance tensorizes:

$$\operatorname{Var}[Z] \le \sum_{i=1}^{n} \mathbb{E} \operatorname{Var}_i[Z].$$

That is, for a nonlinear function, the variance is controlled by the variance of individual coordinates of f.

To obtain exponential concentration results for nonlinear functions, we will study notions of entropy, and we will show that entropy also tensorizes. Using a classic argument, due the Herbst, we can derive concentration inequalities from entropy bounds. This approach allows us to obtain exponential concentration for a nonlinear function of independent random variables. In the simplest case, the variance proxy is the same quantity that appeared in the bounded differences inequality for the variance.

5.1 Entropy for random variables

Entropy is a measure of the unpredictability or dispersion of a random variable. In this section, we introduce some measures of entropy that are designed to study concentration properties.

5.1.1 Entropy and relative entropy

We begin with some entropy-like functions defined for positive numbers.

Definition 5.1 (Entropy; relative entropy). The (negative) *entropy* is the convex function

$$h(t) := t \log t \quad \text{for } t \in \mathbb{R}_+.$$

We use the convention $0 \log 0 = 0$.

For positive numbers a and t, the entropy of a relative to t is given by

$$\mathrm{D}(a \,\|\, t) := a(\log a - \log t) - (a - t) \quad \text{for } a, t \in \mathbb{R}_+.$$

We use the conventions $\mathrm{D}(0\|0) = 0$ and $\mathrm{D}(a\|0) = +\infty$ for $a > 0$.

The function D is called the *relative entropy* or the *generalized information divergence*.

Let us take a moment to understand the basic properties of relative entropy. First, note that the relative entropy is the Bregman divergence associated with the (univariate)

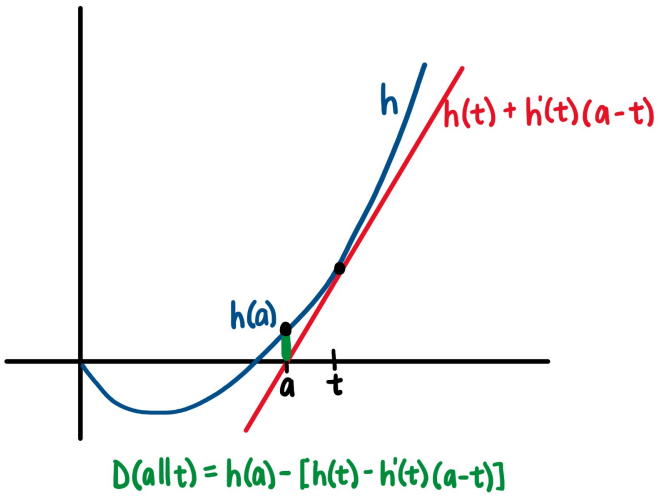


Figure 5.1 Relative entropy is the Bregman divergence associated with the convex function h.

entropy function h:

$$\mathrm{D}(a \,\|\, t) = h(a) - h(t) - h'(t)(a - t) \quad \text{for } a \geq 0 \text{ and } t > 0.$$

In other words, we subtract from $h(a)$ the first-order Taylor expansion of h around the point t. As illustrated in Figure 5.1, the tangent line of a convex function is always a lower bound, so we deduce that the divergence is positive:

$$\mathrm{D}(a \,\|\, t) \geq 0 \quad \text{for all } a, t \geq 0.$$

Equality holds if and only if $a = t$. Thus, we can regard the relative entropy as a measure of the "distance" from the point a to the point t. Note, however, that the relative entropy is not symmetric, nor does it satisfy the triangle inequality.

5.1.2 Concentration entropy

Next, we introduce a notion of concentration entropy that describes the fluctuation of a random variable about its expected value.

Definition 5.2 (Concentration entropy). Let Y be a positive random variable. Define the *concentration entropy* of Y to be the number

$$\begin{aligned} \mathrm{ent}(Y) &:= \mathbb{E}[Y \log Y] - (\mathbb{E}\, Y) \log(\mathbb{E}\, Y) \\ &= \mathbb{E}[Y(\log Y - \log(\mathbb{E}\, Y)) - (Y - \mathbb{E}\, Y)] = \mathbb{E}[\mathrm{D}(Y \,\|\, \mathbb{E}\, Y)]. \end{aligned}$$

In other words, the concentration entropy measures the average divergence of a realization of Y from its expected value $\mathbb{E}\, Y$. Since the relative entropy is positive,

$$\mathrm{ent}(Y) \geq 0.$$

The last relation also follows from Jensen's inequality and the fact that the entropy is convex.

In fact, the expectation $\mathbb{E}\, Y$ is the number from which the random variable Y has the least average divergence.

Exercise 5.3 (Infimal representation of entropy). Prove the variational formula $\operatorname{ent}(Y) = \inf_{t>0} \mathbb{E}[\mathrm{D}(Y \,\|\, t)]$. **Hint:** Differentiate with respect to t.

Exercise 5.4 (Entropy of an exponential). Most commonly, we apply the concentration entropy to a random variable of the form $Y = \mathrm{e}^Z$ where Z is a real random variable. Calculate $\operatorname{ent}(Y)$ in terms of Z. Draw a connection with mgfs and cgfs.

Exercise 5.5 (Scale invariance). For a positive random variable Y, show that $\operatorname{ent}(aY) = \operatorname{ent}(Y)$ for all $a \in \mathbb{R}_{++}$. Deduce that $\operatorname{ent}(\mathrm{e}^Z) = \operatorname{ent}(\mathrm{e}^{Z - \mathbb{E} Z})$.

5.2 The Herbst argument

To understand why the concentration entropy is a valuable tool for understanding concentration, we need to draw a connection with other concepts from the theory of concentration. The next result shows that bounds for the concentration entropy lead to bounds for the cgf. This observation is attributed to Herbst (unpublished work!), and it is commonly called the *Herbst argument*.

Proposition 5.6 (Herbst). Consider an independent family $(X_1, \dots, X_n)$ of real random variables, and form $Z = f(X_1, \dots, X_n)$. Suppose that we have an entropy bound of the form

$$\operatorname{ent}(\mathrm{e}^{\theta Z}) \le \tfrac{1}{2}\theta^2 v \cdot m_Z(\theta) \quad \text{for } \theta \in \mathbb{R}.$$

As usual m_Z is the mgf. Then the cgf of the centered random variable $Z - \mathbb{E} Z$ admits the bound

$$\xi_{Z - \mathbb{E} Z}(\theta) = \log \mathbb{E}\, \mathrm{e}^{\theta(Z - \mathbb{E} Z)} \le \tfrac{1}{2}\theta^2 v.$$

Proof. Without loss of generality, we may assume that $\mathbb{E} Z = 0$. (Why?) Thus, the cgf satisfies $\xi_Z(0) = 0$ and $\xi'(Z) = \mathbb{E} Z = 0$. For future reference, observe that

$$\lim_{\theta \to 0} \theta^{-1} \xi_Z(\theta) = 0$$

because of L'Hôpital's rule.

Now, let us calculate a derivative:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\theta}\left[\frac{1}{\theta}\xi_Z(\theta)\right] &= \frac{1}{\theta}\xi_Z'(\theta) - \frac{1}{\theta^2}\xi_Z(\theta) \\
&= \frac{1}{\theta}\frac{\mathbb{E}[Z\mathrm{e}^{\theta Z}]}{\mathbb{E}\,\mathrm{e}^{\theta Z}} - \frac{1}{\theta^2}\log \mathbb{E}\,\mathrm{e}^{\theta Z} \\
&= \frac{1}{\theta^2}\frac{\mathbb{E}[\mathrm{e}^{\theta Z}\log \mathrm{e}^{\theta Z}] - (\mathbb{E}\,\mathrm{e}^{\theta Z})\log \mathbb{E}\,\mathrm{e}^{\theta Z}}{\mathbb{E}\,\mathrm{e}^{\theta Z}} \\
&= \frac{1}{\theta^2}\frac{\operatorname{ent}(\mathrm{e}^{\theta Z})}{m_Z(\theta)}.
\end{aligned}$$

This formula shows how a bound on the entropy relative to the mgf is connected to the rate of change of the cgf.

We will now combine the last two displays with the Fundamental Theorem of Calculus to get bounds for the cgf in terms of the entropy. For $\theta > 0$,

$$\begin{aligned}
\frac{1}{\theta}\xi_Z(\theta) &= \int_0^\theta \frac{\mathrm{d}s}{s^2} \cdot \frac{\operatorname{ent}(\mathrm{e}^{sZ})}{m_Z(s)} \\
&\le \int_0^\theta \frac{\mathrm{d}s}{s^2} \cdot \frac{1}{2}s^2 v = \frac{1}{2}\theta v.
\end{aligned}$$

The second line follows because of our assumption on the entropy.

We deduce that $\xi_Z(\theta) \le \frac{1}{2}\theta^2 v$ when $\theta > 0$. A similar argument is valid when $\theta < 0$. This point completes the proof. ∎

Exercise 5.7 (Negative values of θ). Complete the proof of Proposition 5.6 by deriving the cgf bound for $\theta < 0$.

We now know that a certain type of bound for $\operatorname{ent}(\mathrm{e}^{\theta Z})$ induces a bound on the cgf $\xi_{Z-\mathbb{E}Z}$. Therefore, if we have a quadratic bound on concentration entropy, we obtain a quadratic bound on the cgf. A routine application of the Laplace transform method yields normal concentration for Z.

Exercise 5.8 (Entropy and normal concentration). Deduce that the bound

$$\frac{\operatorname{ent}(\mathrm{e}^{\theta Z})}{m_Z(\theta)} \le \tfrac{1}{2}\theta^2 v \quad \text{for all } \theta \in \mathbb{R}$$

implies the concentration inequality

$$\mathbb{P}\{|Z - \mathbb{E}Z| \ge t\} \le 2\,\mathrm{e}^{-t^2/(2v)} \quad \text{for all } t \ge 0.$$

Problem 5.9 (Entropy: Range bound). For a random variable Z with mean zero and $a \le Z \le b$, prove that

$$\frac{\operatorname{ent}(\mathrm{e}^{\theta Z})}{m_Z(\theta)} \le \tfrac{1}{8}\theta^2(b-a)^2.$$

Hint: Write the left-hand side in terms of the cgf $\xi_Z(\theta)$ and its derivative $\xi_Z'(\theta)$. Bound this expression using the fundamental theorem of calculus and Hoeffding's cgf inequality $\xi_Z''(\theta) \le \frac{1}{4}(b-a)^2$.

5.3 Entropy tensorizes

The concentration entropy is a powerful tool for studying concentration in large part because it has a tensorization property, analogous with the tensorization property of the variance. To state this result, we need a definition of coordinatewise entropy.

Definition 5.10 (Coordinatewise entropy). Consider an independent family $(X_1, \dots, X_n)$ of random variables, and form $Y = f(X_1, \dots, X_n)$. The *coordinatewise entropy* describes the entropy production due to changes in the ith coordinate:

$$\operatorname{ent}_i(Y) := \mathbb{E}_i[Y(\log Y - \log \mathbb{E}_i Y)].$$

Recall that $\mathbb{E}_i$ is the expectation with respect to the ith random variable X_i only and holding X_j for $j \ne i$ constant. Therefore, ent_i is a function of $(X_j : j \ne i)$.

Now, we will show that the total entropy of a function of independent random variables is controlled by the sum of the conditional entropies.

Theorem 5.11 (Entropy tensorizes). With the prevailing notation and assumptions,

$$\operatorname{ent}(Y) \le \mathbb{E}\left[\sum\nolimits_{i=1}^{n} \operatorname{ent}_i(Y)\right].$$

In other words, the total fluctuation of Y around its mean is controlled by the sum of the fluctuations due to changes in the individual coordinates. To prove this result, we need the following fundamental fact about relative entropy.

Proposition 5.12 (Relative entropy is convex). The function $(a, t) \mapsto \mathrm{D}(a \,\|\, t)$ is convex on $\mathbb{R}^2_{++}$.

Proof. Let h be the univariate entropy function. Observe that the divergence can be written as

$$\mathrm{D}(a \,\|\, t) = t \cdot h(a/t) - (a - t).$$

In convex analysis, the bivariate function $(a, t) \mapsto t\, h(a/t)$ is called the *perspective transformation* of the convex function h. It is a standard fact that the perspective of a convex function is convex. The second term is linear, so it does not affect the convexity properties. ∎

As a consequence of Proposition 5.12 and Jensen's inequality, we have the relation

$$\mathrm{D}(\mathbb{E}\, Y \,\|\, \mathbb{E}\, Z) \le \mathbb{E}\, \mathrm{D}(Y \,\|\, Z) \quad \text{for all positive random variables } Y, Z.$$

The same kind of averaging inequality also holds conditionally.

Proof of Theorem 5.8. The proof is similar in spirit to the proof that the variance tensorizes. We can decompose the concentration entropy using a Doob martingale. This expression allows us to isolate the contribution of each individual random variable X_i to the entropy.

Much as before, define

$$Y_i = \mathbb{E}_1 \dots \mathbb{E}_i\, Y \quad \text{for each } i = 0, \dots, n.$$

By convention, $Y_0 = Y$ and $Y_n = \mathbb{E}\, Y$. We have the telescoping sum

$$\begin{aligned}
\operatorname{ent}(Y) &= \mathbb{E}[Y(\log Y - \log \mathbb{E}\, Y)] \\
&= \mathbb{E}\left[Y\left(\sum\nolimits_{i=1}^{n} \log Y_{i-1} - \log Y_i\right)\right] \\
&= \mathbb{E}\left[\sum\nolimits_{i=1}^{n} Y_{i-1}(\log Y_{i-1} - \log \mathbb{E}_i\, Y_{i-1})\right] \\
&= \mathbb{E}\left[\sum\nolimits_{i=1}^{n} \mathrm{D}(Y_{i-1} \,\|\, \mathbb{E}_i\, Y_{i-1})\right].
\end{aligned}$$

Recall that $Y_{i-1} = \mathbb{E}_1 \cdots \mathbb{E}_{i-1}\, Y$. Therefore, we can apply Jensen's inequality conditionally to draw out the expectation $\mathbb{E}_1 \cdots \mathbb{E}_{i-1}$. Indeed,

$$\begin{aligned}
\operatorname{ent}(Y) &\le \mathbb{E}\left[\sum\nolimits_{i=1}^{n} \mathbb{E}_1 \cdots \mathbb{E}_{i-1}\, \mathrm{D}(Y \,\|\, \mathbb{E}_i\, Y)\right] \\
&= \mathbb{E}\left[\sum\nolimits_{i=1}^{n} \mathbb{E}_i\, \mathrm{D}(Y \,\|\, \mathbb{E}_i\, Y)\right] \\
&= \mathbb{E}\left[\sum\nolimits_{i=1}^{n} \operatorname{ent}_i(Y)\right].
\end{aligned}$$

We have liberally applied the fact that the coordinatewise expectations are commuting, idempotent operators. ∎

5.4 Entropy bounds

In the last section, we observed that we can control the entropy of a multivariate function in terms of the coordinatewise entropies. Therefore, we can exploit bounds on the entropy of a univariate function to obtain bounds for the total entropy.

5.4.1 Entropy and exchangeability

First, we will show that the entropy can be bounded by a more symmetrical function involving a pair of iid random variables. As a particular example, we can obtain an elegant bound for the entropy of an exponential.

Proposition 5.13 (Exchangeable bound for entropy). Let Y, Y' be iid positive random variables.

$$\operatorname{ent}(Y) \le \frac{1}{2}\,\mathbb{E}[(Y - Y')(\log Y - \log Y')].$$

In particular, for iid real random variables Z, Z', we have

$$\operatorname{ent}(\mathrm{e}^Z) \le \frac{1}{2}\,\mathbb{E}[(\mathrm{e}^Z - \mathrm{e}^{Z'})(Z - Z')].$$

Proof. For the first inequality, by convexity of the negative logarithm,

$$\begin{aligned}\operatorname{ent}(Y) &= \mathbb{E}_Y[Y(\log Y - \log \mathbb{E}_{Y'} Y')] \\ &\le \mathbb{E}_{Y,Y'}[Y(\log Y - \log Y')].\end{aligned}$$

We write $\mathbb{E}_X$ for the expectation with respect to the randomness in a variable X. The second relation follows from Jensen's inequality and independence. Since (Y, Y') has the same distribution as (Y', Y), it also follows that

$$\operatorname{ent}(Y) = \operatorname{ent}(Y') \le \mathbb{E}_{Y,Y'}[Y'(\log Y' - \log Y)].$$

Thus, by averaging,

$$\operatorname{ent}(Y) = \frac{1}{2}\,\mathbb{E}[(Y - Y')(\log Y - \log Y')].$$

Substituting in the exponential $Y = \mathrm{e}^Z$ and $Y' = \mathrm{e}^{Z'}$, we discover that

$$\operatorname{ent}(\mathrm{e}^Z) \le \frac{1}{2}\,\mathbb{E}[(\mathrm{e}^Z - \mathrm{e}^{Z'})(Z - Z')].$$

This is the required result. ∎

5.4.2 Discrete MLS

We are now prepared to establish a univariate entropy bound, which is called a *discrete modified logarithmic Sobolev inequality* (MLS). The terminology will be discussed in more detail in subsequent lectures.

Theorem 5.14 (Discrete MLS). Let Z, Z' be iid real random variables. Then

$$\operatorname{ent}(\mathrm{e}^Z) \le \frac{1}{2}\,\mathbb{E}\left[\,\mathbb{E}'[(Z - Z')^2] \cdot \mathrm{e}^Z\right].$$

We write $\mathbb{E}'$ for the expectation with respect to the randomness in Z'.

Proof. The argument depends on a simple numerical inequality. For real a, b,

$$\begin{aligned}(a-b)(\mathrm{e}^a - \mathrm{e}^b) &= (a-b)^2 \int_0^1 \mathrm{e}^{(1-\tau)a + \tau b}\,\mathrm{d}\tau \\ &\le (a-b)^2 \int_0^1 \left[(1-\tau)\mathrm{e}^a + \tau \mathrm{e}^b\right]\mathrm{d}\tau = \tfrac{1}{2}(a-b)^2(\mathrm{e}^a + \mathrm{e}^b).\end{aligned}$$

The first relation follows from the Fundamental Theorem of Calculus, and the inequality holds because the exponential function is convex.

Applying this inequality to the formula in Proposition 5.13,

$$\begin{aligned}\operatorname{ent}(\mathrm{e}^Z) &\le \frac{1}{2}\,\mathbb{E}[(Z - Z')(\mathrm{e}^Z - \mathrm{e}^{Z'})] \\ &\le \frac{1}{4}\,\mathbb{E}[(Z - Z')^2 \cdot (\mathrm{e}^Z + \mathrm{e}^{Z'})] = \frac{1}{2}\,\mathbb{E}\left[\,\mathbb{E}'[(Z - Z')^2] \cdot \mathrm{e}^Z\right].\end{aligned}$$

In the last step, we have used the fact that Z, Z' are iid. ∎

Using tensorization, we obtain a multivariate extension of the last result. To that end, we recall the notation that arose in the Efron–Stein–Steele inequality. Consider an independent family $(X_1, \dots, X_n)$ of random variables, and form the function $Z = f(X_1, \dots, X_n)$. Let (X_i') be an independent draw of (X_i), and define the function

$$Z^{(i)} := f(X_1, \dots, X_i', \dots, X_n) \quad \text{for } i = 1, \dots, n.$$

That is, $Z^{(i)}$ is obtained from Z by refreshing the ith coordinate X_i with an independent copy X_i'. Conditional on the values $(X_j : j \neq i)$, the pair $(Z, Z^{(i)})$ is iid.

Corollary 5.15 (Multivariate discrete MLS). With the prevailing notation,

$$\operatorname{ent}(\mathrm{e}^Z) \le \mathbb{E}[V\mathrm{e}^Z], \quad \text{where} \quad V = \frac{1}{2}\sum\nolimits_{i=1}^n \mathbb{E}'(Z - Z^{(i)})^2.$$

Here, $\mathbb{E}'$ computes the expectation with respect to the randomness in (X_i') only.

Proof. The tensorization of entropy (Theorem 5.11) states that

$$\operatorname{ent}(\mathrm{e}^Z) = \mathbb{E}\left[\sum\nolimits_{i=1}^n \operatorname{ent}_i(\mathrm{e}^Z)\right].$$

Applying the discrete MLS (Theorem 5.14) coordinatewise,

$$\operatorname{ent}_i(\mathrm{e}^Z) \le \frac{1}{2}\,\mathbb{E}_i\left[\,\mathbb{E}_i'[(Z - Z^{(i)})^2] \cdot \mathrm{e}^Z\right].$$

Combine the two displays to complete the proof. ∎

The random variable V is the same quantity that appears in the Efron–Stein–Steele inequality as a bound for the variance. It may be interpreted as an estimate of the squared energy of the random variable Z at the random point $(X_1, \dots, X_n)$.

5.5 From entropy bounds to concentration

Finally, we derive concentration inequalities from the entropy bounds in the last section.

5.5.1 Uniform bounds on the variance

The simplest setting for the discrete MLS (Corollary 5.15) occurs when the random variance proxy V admits a uniform bound.

Corollary 5.16 (Discrete MLS: Uniform bound). With the prevailing notation, for $\theta \in \mathbb{R}$,

$$\operatorname{ent}(\mathrm{e}^{\theta Z}) \le \tfrac{1}{2}\theta^2 v \cdot m_Z(\theta) \quad \text{where} \quad v = \left\|\sum\nolimits_{i=1}^n \mathbb{E}'(Z - Z^{(i)})^2\right\|_\infty.$$

Recall that $\|\cdot\|_\infty$ is the essential supremum of a random variable.

This result is an immediate consequence of Corollary 5.15. Combining this result with the Herbst argument (Proposition 5.6), we see that

$$\mathbb{P}\left\{|Z - \mathbb{E}\, Z| \geq t\sqrt{v}\right\} \leq \mathrm{e}^{-t^2/2} \quad \text{for all } t \geq 0.$$

This is our first nonlinear concentration inequality.

5.5.2 Bounded differences inequality

To appreciate the content of Corollary 5.16, let us make a further estimate for the uniform variance proxy v. Recall our definition of the discrete "derivative" of a function:

$$(\mathrm{D}_i f) := \sup_{x \in \operatorname{supp}(X_i)} f(X_1, \ldots, x, \ldots, X_n) - \inf_{x \in \operatorname{supp}(X_i)} f(X_1, \ldots, x, \ldots, X_n).$$

Since $Z - Z^{(i)}$ only changes in the ith coordinate, it is obvious that

$$\mathbb{E}'(Z - Z^{(i)})^2 \leq (\mathrm{D}_i f)^2.$$

Therefore, the variance proxy in Corollary 5.16 is bounded as

$$v \leq \left\| \sum_{i=1}^n (\mathrm{D}_i f)^2 \right\|_\infty.$$

These considerations lead to a bounded difference inequality.

Theorem 5.17 (Bounded differences). Consider an independent family $(X_1, \ldots, X_n)$ of random variables, and consider a function $f : \mathbb{R}^n \to \mathbb{R}$. Define the variance proxy

$$v := \left\| \sum_{i=1}^n (\mathrm{D}_i f)^2 \right\|_\infty.$$

Then

$$\mathbb{P}\left\{|Z - \mathbb{E}\, Z| \geq t\sqrt{v}\right\} \leq 2\mathrm{e}^{-t^2/2} \quad \text{for all } t \geq 0.$$

We have obtained normal concentration for a nonlinear function in terms of the discrete derivatives. This result is an analog of Hoeffding's inequality, proven in the previous chapter. The constants here are worse than the result for independent sums, but they can be improved with further care.

In the next lecture, we will see that the discrete MLS is just one example from a larger class of modified logarithmic Sobolev inequalities. These results allow us to prove more refined nonlinear concentration inequalities for particular distributions.

Exercise 5.18 (Bounded differences: Sharp constant). Deduce that Theorem 5.17 holds with v replaced by $v/4$. **Hint:** Use Problem 5.9.

6. Log-Sobolev Inequalities

Date: 21 January 2021 **Scribe:** Anushri Dixit

In the last lecture, we used entropy as a tool for proving normal concentration for nonlinear functions of independent random variables. Among other things, this approach yields a bounded difference inequality where the variance proxy is the sum of squared "discrete derivatives" of the random variable. In this lecture, we will develop some nonlinear concentration inequalities involving the squared (calculus) derivative of the random variable. The key input is a type of functional inequality called a modified log-Sobolev inequality (MLSI).

Agenda:
1. MLSIs and concentration
2. Rademacher MLSI
3. Gaussian MLSI
4. Convex MLSI

6.1 Recap: Entropy and concentration

Let us begin with a review of the arguments from the last lecture. First, recall that the relative entropy (or generalized information divergence) measures of distance between two positive numbers:

$$\mathrm{D}(a \,\|\, t) := a(\log a - \log t) - (a - t) \geq 0 \quad \text{for } a, t \geq 0.$$

Aside: The KL divergence often arises in information theory and statistical physics.

The relative entropy is positive and convex.

Next, we define the concentration entropy, which measures the average divergence between a realization of a random variable and its mean.

Definition 6.1 (Concentration entropy). The entropy of a positive random variable Y is given by

$$\mathrm{ent}(Y) := \mathbb{E}[\mathrm{D}(Y \,\|\, \mathbb{E}\, Y)] = \mathbb{E}[Y(\log Y - \log \mathbb{E}\, Y)].$$

The concentration entropy is always positive, and it equals zero if and only if the random variable Y is constant: $Y = \mathbb{E}\, Y$ almost surely. The expectation plays a distinguished role in the definition; indeed,

$$\mathrm{ent}(Y) = \inf_{t>0} \mathbb{E}[\mathrm{D}(Y \,\|\, t)]. \tag{6.1}$$

Like the variance, the concentration entropy tensorizes. That is, we can control the entropy of a random variable in terms of the entropy produced by each coordinate. Consider a positive function $Y = f(X_1, \dots, X_n)$ of independent random variables (X_i). The coordinatewise entropy is

$$\mathrm{ent}_i(Y) := \mathbb{E}_i[\mathrm{D}(Y \,\|\, \mathbb{E}_i\, Y)].$$

Notation: $\mathbb{E}_i$ averages over only X_i, holding $(X_j : j \neq i)$ fixed. The symbol $\mathbb{E}$ refers to the total expectation, while $\mathbb{E}'$ denotes the expectation with respect to an independent copy of the underlying random variables.

We can now state the tensorization theorem.

Theorem 6.2 (Entropy tensorizes). With the prevailing notation,

$$\operatorname{ent}(Y) \le \mathbb{E}\left[\sum_{i=1}^{n} \operatorname{ent}_i(Y)\right].$$

This results follows from a Doob martingale decomposition and the convexity of the information divergence.

Tensorization allows us to lift bounds for univariate entropy to obtain entropy bounds for nonlinear functions of independent random variables.

Example 6.3 (Discrete MLSI). Let us consider the entropy of the exponential of a real random variable Z. As discussed in the last lecture, we have the univariate discrete MLSI:

$$\operatorname{ent}(\mathrm{e}^Z) \le \frac{1}{2}\,\mathbb{E}\left[(Z - Z')^2 \cdot \mathrm{e}^Z\right],$$

where Z' is an independent copy of Z.

We can tensorize this univariate bound to obtain a discrete MLSI for a multivariate function $Z = f(X_1, \dots, X_n)$ of independent random variables:

$$\operatorname{ent}(\mathrm{e}^Z) \le \frac{1}{2}\,\mathbb{E}\left[\left(\sum_{i=1}^{n} (Z - Z^{(i)})^2\right) \cdot \mathrm{e}^Z\right],$$

where $Z^{(i)} = f(X_1, \dots, X_i', \dots, X_n)$ and (X_i') is an independent copy of (X_i). This inequality can be rewritten as

$$\operatorname{ent}(\mathrm{e}^Z) \le \frac{1}{2}\,\mathbb{E}[V \mathrm{e}^Z] \quad \text{where} \quad V = \mathbb{E}'\left[\sum_{i=1}^{n} (Z - Z^{(i)})^2\right].$$

The random variable V is a random proxy for the variance. Up to scaling, recall that V also appears on the right-hand side of the Efron–Stein–Steele inequality.

We can make use of this entropy inequality to get a parameterized bound that depends quadratically on a parameter s and is controlled by the maximum size of the variance proxy V. In other words,

The norm $\|\cdot\|_\infty$ returns the essential supremum of a random variable.

$$\operatorname{ent}(\mathrm{e}^{sZ}) \le \frac{s^2}{2}\|V\|_\infty\,\mathbb{E}[\mathrm{e}^{sZ}], \quad \text{for } s \in \mathbb{R}.$$

Now, the Herbst argument yields bounds on the cgf. For $\theta > 0$,

$$\begin{aligned}\frac{1}{\theta}\xi_{Z-\mathbb{E}Z}(\theta) &= \int_0^\theta \frac{\mathrm{d}s}{s^2}\frac{\operatorname{ent}(\mathrm{e}^{sZ})}{\mathbb{E}\,\mathrm{e}^{sZ}} \\ &\le \int_0^\theta \frac{\mathrm{d}s}{2}\|V\|_\infty = \frac{\theta}{2}\|V\|_\infty.\end{aligned}$$

A similar bound holds for $\theta < 0$. Finally, the Laplace transform method tells us that $Z - \mathbb{E}Z$ has normal concentration with variance proxy $\|V\|_\infty$. See Problem Set 1, Exercise 3.

To summarize, we used the discrete MLSI to bound the entropy by the mgf. Using the Herbst argument and the Laplace transform method, we arrive at a normal concentration inequality. ∎

6.2 Modified log-Sobolev inequalities and concentration

In the previous example, we developed a bound on the entropy using a uniform estimate for the random variance proxy. This approach yields an entropic formulation of the fact that the nonlinear random variable has normal concentration.

Nevertheless, this approach has several shortcomings. The discrete variance proxy V is not always easy to work with, and we might like results formulated in terms of ordinary (calculus) derivatives. In addition, the bound only applies to the case where the variance proxy admits a uniform bound, which excludes a number of important examples (such as quadratic forms). In this lecture, we will develop nonlinear concentration results involving derivatives. You will engage the second issue on Problem Set 2.

Consider a real-valued function $f(X)$ of a real random variable X. In this discussion, it is convenient to use functional notation for random variables. Informally, suppose we can prove an inequality in one dimension of the form

The symbol f' denotes the derivative of f, not an independent copy.

$$\operatorname{ent}(\mathrm{e}^f) \le \frac{1}{2}\,\mathbb{E}\left[|f'|^2\,\mathrm{e}^f\right].$$

This bound is also called a (modified) log-Sobolev inequality (MLSI). It is a "continuous" analog of the discrete MLSI from Example 6.3. Note that the MLSI can be rewritten as

$$\operatorname{ent}(\mathrm{e}^f) \le 2\,\mathbb{E}\left[\left((\mathrm{e}^{f/2})'\right)^2\right].$$

In other words, the right-hand side is related to the average energy of $\mathrm{e}^{f/2}$.

By tensorization, we can extend this inequality to a random variable of the form $f(X_1, \dots, X_n)$ where the X_i are iid copies of X. This step yields a multivariate MLSI:

$$\operatorname{ent}(\mathrm{e}^f) \le \frac{1}{2}\,\mathbb{E}\left[\|\nabla f\|^2\,\mathrm{e}^f\right].$$

Applying the Herbst argument and the Laplace transform method, we obtain a normal concentration inequality:

$$\mathbb{P}\left\{|f - \mathbb{E}\,f| \ge t\right\} \le 2\,\mathrm{e}^{-t^2/(2v)} \quad \text{where} \quad v = \big\|\|\nabla f\|_2^2\big\|_\infty.$$

In other words, the variance proxy v is a uniform bound on the squared Euclidean norm of the gradient, also known as the Dirichlet energy.

Recall that a Poincaré inequality controls the variance in terms of the expectation of the Dirichlet energy:

$$\operatorname{Var}[f] \le \mathbb{E}\left[\|\nabla f\|_2^2\right].$$

We can see the univariate and multivariate MLSIs are a kind of entropic counterpart to the Poincaré inequality. Whereas we only derived variance bounds from the Poincaré inequality, the MLSI yields normal concentration.

As with Poincaré inequalities, we need to develop *ad hoc* arguments to establish an MLSI for a particular random variable X. An advantage of the discrete MLSI (Example 6.3) is that it holds for any distribution. Later, we will encounter a convex MLSI that holds for all bounded distributions—but only for convex functions.

Problem 6.4 (MLSI implies Poincaré). In fact, the MLSI is a stronger result that the Poincaré inequality. Show that an MLSI implies a Poincaré inequality. **Hint:** Apply the MLSI to the function $\log(1 + \eta f)$ and take $\eta \to 0$.

Problem 6.5 (Poincaré implies exponential concentation). In fact, the Poincaré inequality can be used to obtain exponential concentration. To prove this fact, apply the Poincaré inequality to $\mathrm{e}^{\theta f}$ to obtain a recursive formula for the mgf.

> **Aside:** (Modified) log-Sobolev inequalities have a very long history. They play an important role in probability, analysis and mathematical physics. As with other Sobolev inequalities, they quantify the idea that we can trade differentiability properties for integrability properties. That is, we control a moment of a function in terms of a moment of its derivative. Log-Sobolev inequalities are particularly important because they extend to infinite-dimensional settings, whereas many other types of Sobolev inequalities depend on the underlying dimension of the space.

6.3 Rademacher MLSI

First, let us develop an MLSI for the simplest nontrivial probability distribution, the Rademacher distribution. We will derive this result as a simple consequence of the discrete MLSI, although we will not achieve sharp constants.

Recall that a Rademacher random variable $\varepsilon \sim \textsc{uniform}\{\pm 1\}$. Let us apply the univariate discrete MLSI to the random variable $f(\varepsilon)$ where $f : \{\pm 1\} \to \mathbb{R}$. After a short calculation, we obtain

$$\operatorname{ent}(\mathrm{e}^{f(\varepsilon)}) \le \frac{1}{4}\,\mathbb{E}\left[(f(+1) - f(-1))^2\,\mathrm{e}^{f(\varepsilon)}\right].$$

We can interpret the first term in the expectation on the right-hand side as the square of the discrete derivative of f.

Now, consider an independent family $(\varepsilon_1, \dots, \varepsilon_n)$ of Rademacher random variables and a real random variable $Z = f(\varepsilon_1, \dots, \varepsilon_n)$. As in the last paragraph, we calculate the expectation due to the randomness in the ith coordinate as

$$\mathbb{E}'(Z - Z^{(i)})^2 = \frac{1}{2}\left[f(\varepsilon_1, \dots, +1, \dots, \varepsilon_n) - f(\varepsilon_1, \dots, -1, \dots, \varepsilon_n)\right]^2 = \frac{1}{2}(\mathrm{D}_i f)^2.$$

We have used our standard notation D_i for the "discrete derivative" in the ith coordinate. Hence, the random variance proxy of Z is

$$V = \frac{1}{2}\sum\nolimits_{i=1}^{n} (\mathrm{D}_i f)^2.$$

Keep in mind that the variance proxy V is a function of $\varepsilon_1, \dots, \varepsilon_n$. We can interpret V as the squared Euclidean norm of the discrete gradient of f. Using the multivariate discrete MLSI from Example 6.3, we immediately obtain the following result.

Theorem 6.6 (Rademacher MLSI). With the prevailing notation

$$\operatorname{ent}(\mathrm{e}^{f}) \le \frac{1}{4}\,\mathbb{E}\left[\left(\sum\nolimits_{i=1}^{n} (\mathrm{D}_i f)^2\right)\mathrm{e}^{f}\right].$$

The theorem gives us a bound on the entropy of a nonlinear function of Rademacher random variables using a random variance proxy. The entropy bound reflects the energy, the sum of squared (discrete) derivatives.

Using the Herbst argument and the Laplace transform method, we obtain normal concentration with variance proxy

$$v = \|V\|_\infty = \max\nolimits_{\varepsilon_i} \sum\nolimits_{i=1}^{n} \left(\mathrm{D}_i f(\varepsilon_1, \dots, \varepsilon_n)\right)^2.$$

In the next section, we will use this result to obtain bounds on the entropy of a standard normal random variable.

Problem 6.7 (Rademacher MLSI: Optimal constants). For a univariate function $f : \{\pm 1\} \to \mathbb{R}$, prove that

$$\operatorname{ent}(\mathrm{e}^f) \le \frac{1}{8}\,\mathbb{E}[(f(+1) - f(-1))^2\,\mathrm{e}^f]$$

Hint: Reduce to the case where $f(-1) = 0$ and $f(+1) = s$. Check the case $s = 0$. Consider the derivatives of both sides with respect to s.

6.4 Gaussian MLSI

In this section, we will derive MLSIs for the standard normal distribution. We begin with the univariate case, and we use tensorization to obtain the multivariate generalization.

Theorem 6.8 (Gaussian MLSI: Univariate case). Let $\gamma \sim \textsc{normal}(0, 1)$. For a nice function $f : \mathbb{R} \to \mathbb{R}$,

$$\operatorname{ent}(\mathrm{e}^{f(\gamma)}) \le \mathbb{E}\left[(f'(\gamma))^2\,\mathrm{e}^{f(\gamma)}\right]. \tag{6.2}$$

In fact, the inequality holds for all f where the both sides are defined and finite.

Proof sketch. The key idea is to use the central limit theorem (CLT) to condense the univariate MLSI for a standard normal random variable from the multivariate MLSI for Rademacher random variables. To justify the calculations, assume that f and its derivatives f', f'' are bounded.

Consider the sequence of random variables

$$Z_n \coloneqq \frac{1}{\sqrt{n}} \sum\nolimits_{i=1}^{n} \varepsilon_i.$$

As $n \to \infty$, the sequence (Z_n) converges weakly to a standard normal distribution by the CLT.

To apply the Rademacher MLSI (Theorem 6.6), we need to compute discrete derivatives. For each index $i = 1, \dots, n$,

$$\begin{aligned}
\mathrm{D}_i f &= f\left(\frac{1}{\sqrt{n}}\left(+1 + \sum\nolimits_{j\neq i} \varepsilon_j\right)\right) - f\left(\frac{1}{\sqrt{n}}\left(-1 + \sum\nolimits_{j\neq i} \varepsilon_j\right)\right) \\
&= \frac{2}{\sqrt{n}}\, f'\left(\frac{1}{\sqrt{n}} \sum\nolimits_{i=1}^{n} \varepsilon_i\right) + \mathcal{O}(n^{-1}) \\
&= \frac{2}{\sqrt{n}}\, f'(Z_n) + O(n^{-1}).
\end{aligned}$$

We have employed a first-order Taylor expansion of f around the standardized sum Z_n, using the assumption that f'' is bounded. Summing the squares of the discrete derivatives,

$$\sum\nolimits_{i=1}^{n} (\mathrm{D}_i f)^2 = 4 f'(Z_n) + O(n^{-1/2}).$$

From Theorem 6.6, we obtain the bound

$$\begin{aligned}
\operatorname{ent}(\mathrm{e}^f) &\le \frac{1}{4}\,\mathbb{E}\left[\left(\sum\nolimits_{i=1}^{n} (\mathrm{D}_i f)^2\right)\mathrm{e}^f\right] \\
&\approx \mathbb{E}\left[f'(Z_n)^2\,\mathrm{e}^{f(Z_n)}\right] \to \mathbb{E}\left[f'(\gamma)^2\,\mathrm{e}^{f(\gamma)}\right].
\end{aligned}$$

The last limit follows from the CLT since we are considering a bounded, continuous function of Z_n. ∎

Now we can tensorize Theorem 6.8 to obtain a multivariate Gaussian MLSI.

Corollary 6.9 (Gaussian MLSI). Consider a standard normal random vector $\boldsymbol{\gamma} \sim \text{NORMAL}(\mathbf{0}, \mathbf{I}_n)$ and a (nice) function $f : \mathbb{R}^n \to \mathbb{R}$. Then

$$\operatorname{ent}(\mathrm{e}^{f(\boldsymbol{\gamma})}) \le \mathbb{E}\left[\|\nabla f(\boldsymbol{\gamma})\|_2^2 \, \mathrm{e}^{f(\boldsymbol{\gamma})}\right]. \tag{6.3}$$

Exercise 6.10 (Gaussian MLSI). Derive Corollary 6.9.

Corollary 6.9 leads to particularly simple results when the function f is L-Lipschitz with respect to the Euclidean norm. In this case, $\|\nabla f\|_2^2 \le L^2$ almost everywhere, and we recognize that

$$\operatorname{ent}(\mathrm{e}^{sf}) \le s^2 L^2 \, \mathbb{E}[\mathrm{e}^{sf}] \quad \text{for } s \in \mathbb{R}.$$

Invoking the Herbst argument, we learn that f has normal concentration with variance proxy $2L^2$. The constant two is spurious, and it can be removed if we use the optimal Gaussian MLSI.

Exercise 6.11 (Gaussian MLSI: Optimal constants). Using Problem 6.7, deduce that the Gaussian MLSI holds with constant $1/2$.

> **Aside:** The Gaussian MLSI is intimately related to the convergence of the Ornstein–Uhlenbeck (OU) process, which models the velocity of a massive Brownian particle under the influence of friction.
>
> The equilibrium distribution of the OU process is a standard normal distribution. The generator of the OU process is the second-order differential operator $\boldsymbol{A}f = f'' - xf'$, defined for (smooth) functions on the real line. The associated parabolic PDE takes the form
>
> $$\partial_t u_t = \boldsymbol{A} u_t, \quad \text{where} \quad u_0 = f.$$
>
> The initial condition f is a function on the real line, and u_0 is the solution at time t. The solutions converge to the constant function $u_\infty(x) = \mathbb{E}[f(\gamma)]$, where γ is a standard normal random variable.
>
> The (sharp) Gaussian MLSI implies that the concentration entropy of $u_t(\gamma)$ decays at an exponential rate:
>
> $$\operatorname{ent}(u_t(\gamma)) \le \mathrm{e}^{-t} \operatorname{ent}(u_0(\gamma)) \quad \text{for } t \ge 0.$$
>
> The exponential decay in variance, promised by the Gaussian Poincaré inequality, is a fact about the spectrum of the OU differential operator. In contrast, the Gaussian MLSI requires deeper properties of the operator.

6.5 Convex MLSI

We derived MLSIs for the multivariate Gaussian distributions using the univariate Gaussian MLSI, which in turn we calculated from the Rademacher MLSI. All of these results required independent arguments. One may wonder whether there is a general scheme for deriving MLSIs for a wider class of distributions.

In fact, there is a univariate MLSI that holds for any bounded distribution, but only applies to the case of a convex function. We will establish this result and its multivariate extension in this section.

Theorem 6.12 (Convex MLSI: Univariate case). Consider a bounded, real random variable X taking values in $[a, b]$. Let $f : [a, b] \to \mathbb{R}$ be a *convex* function. Then

$$\operatorname{ent}(\mathrm{e}^f) \le \frac{1}{2}(b-a)^2 \,\mathbb{E}\left[(f')^2 \,\mathrm{e}^f\right]. \tag{6.4}$$

Proof. Using the variational formula for entropy (6.1),

$$\operatorname{ent}(\mathrm{e}^f) = \inf_{t>0} \mathbb{E}\left[\mathrm{e}^{f(X)}(\log \mathrm{e}^{f(X)} - \log(t)) - (\mathrm{e}^{f(X)} - t)\right].$$

Fix a point $y \in \arg\min f$, and select $t = \mathrm{e}^{f(y)}$. Then

$$\begin{aligned}
\operatorname{ent}(\mathrm{e}^f) &\le \mathbb{E}\left[\mathrm{e}^{f(X)}(f(X) - f(y)) - (\mathrm{e}^{f(X)} - \mathrm{e}^{f(y)})\right] \\
&= \mathbb{E}\left[\mathrm{e}^{f(X)}\left(\mathrm{e}^{f(y)-f(X)} - (f(y) - f(X)) - 1\right)\right] \\
&= \mathbb{E}\left[\frac{1}{2}\mathrm{e}^{f(X)}\big(f(y) - f(X)\big)^2\right].
\end{aligned}$$

We have used the fact that $f(y) - f(X) \le 0$ to instantiate the numerical inequality $\mathrm{e}^t - t - 1 \le \frac{1}{2}t^2$, valid for $t \le 0$. Since f is convex, it is supported by its tangent at the random point X. That is,

$$f(y) - f(X) \ge f'(X)(y - X).$$

Negating this inequality, we find that $0 \le f(X) - f(y) \le f'(X)(X - y)$. Taking the square,

$$(f(X) - f(y))^2 \le f'(X)^2(X - y)^2 \le f'(X)(b - a)^2.$$

Combine the displays to complete the proof. ∎

As usual, we may tensorize the univariate bound to obtain a multivariate convex MLSI for an unusual class of functions.

Corollary 6.13 (Convex MLSI). Consider an independent family $(X_1, \dots, X_n)$ of real random variables taking values in the interval $[a, b]$. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a *separately convex* function. Then

$$\operatorname{ent}(\mathrm{e}^f) \le \frac{1}{2}(b-a)^2 \,\mathbb{E}\left[\|\nabla f\|^2 \,\mathrm{e}^f\right]. \tag{6.5}$$

A separately convex function is convex when restricted to each coordinate. This requirement is weaker than convexity. See Problem Set 1.

Exercise 6.14 (Convex MLSI). Derive Corollary 6.13.

As a consequence of Corollary 6.13, a *convex* and L-Lipschitz function of independent, bounded random variables exhibits a normal upper tail. It is important to realize that the corresponding lower tail inequality may not be valid. Indeed, we cannot follow the same argument for the *concave* function $-f$.

The inequality (6.5) is analogous to the Hoeffding bound that we obtained in Lecture 4. However, the uniform bound $\|\nabla f\|^2 \le L^2$ seems as if it may be wasteful. A natural question that arises is whether we can obtain a Bernstein-type inequality that better accounts for the typical size of the gradient. We investigate this question in Problem Set 2.

7. Moment Inequalities

Date: 26 January 2021 **Scribe:** Eitan Levin

Agenda:
1. Symmetrization
2. Khintchine's inequality
3. Rosenthal inequalities

We have seen that Chebyshev's inequality controls the tail of a random variable in terms of its variance. The variance captures the typical (as opposed to worst-case) fluctuation of the random variable, and it can often be effectively bounded (e.g., using Poincaré inequalities). On the other hand, the resulting t^{-2} tail bound is weaker than we might hope. Indeed, we anticipate that an independent sum has subgaussian tails, owing to the central limit theorem. We obtained this type of result using the Laplace transform method and estimates for the cgf.

Unfortunately, the cgf involves an exponential moment. Bounding the cgf is equivalent to bounding polynomial moments of all orders (see Exercise 3 in Problem set 1). Thus, the assumptions needed to control the cgf are stronger than the assumptions needed to control the variance. Our estimates for the cgf of an independent sum depend on the worst-case behavior of the summands. For example, the Hoeffding, Chernoff, and Bernstein inequalities all assume that the random variables are bounded almost surely. Likewise, the nonlinear tail bounds we obtained using modified log-Sobolev inequalities involve the uniform norm of (some notion of) the gradient.

It is natural to wonder whether we can obtain polynomial tail decay for an independent sum Z without assuming that the summands have moments of all orders. In this lecture, we obtain bounds on the polynomial moments $\mathbb{E}\,|Z|^p$ of the sum. By Markov's inequality, these estimates give us tail bounds of the form

$$\mathbb{P}\{|Z| \geq t\} \leq \frac{\mathbb{E}\,|Z|^p}{t^p}, \quad p > 0,\ t \geq 0.$$

Given the function $p \mapsto \mathbb{E}\,|Z|^p$ for all $p > 0$, we can optimize the tail bound over p. The result is at least as good as the best tail bound one obtains using the Laplace transform method.

To bound these moments, we again look to the central limit theorem for inspiration.

The reason that an independent sum of zero-mean random variables has rapidly decaying tails is that many of the summands cancel each other out. To make this cancellation explicit, we develop a method called *symmetrization* to inject random signs into the sum. This modification has a minimal effect on the size of the moments.

Afterward, we prove *Khintchine's inequality*, which shows how the random signs can be used to produce moment bounds. Finally, we combine symmetrization and Khintchine's inequality to derive polynomial moment bounds analogous to the Chernoff and Bernstein cgf bounds. The new moment inequalities involve the expected maximum of the summands, rather than a uniform bound on the summands, so they give more precise information than the exponential inequalities.

7.1 Symmetrization

This discussion is based on [van16, Sec. 7.1]

Consider a sum $\sum_{i=1}^n (X_i - \mathbb{E}\,X_i)$ of iid centered random variables. Naïvely, we might expect this sum to have order $O(n)$ since it contains n terms each of order $O(1)$. However, the central limit theorem implies that this sum is likely to have order $O(\sqrt{n})$. The reason is that the summands are independent and have zero mean, so they are likely to have opposite signs and cancel each other out. The random signs,

$\operatorname{sgn}(X_i - \mathbb{E}\, X_i)$, are responsible for the subgaussian tail decay, while the magnitudes $|X_i - \mathbb{E}\, X_i|$ only determine the width of the gaussian. This discussion suggests that, in order to effectively bound moments of independent sums, we need to make the random signs of the summands explicit and find a way to exploit them. The process of making these signs explicit is called symmetrization:

Theorem 7.1 (Symmetrization). Consider an independent family $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ of random variables taking values in a normed linear space, and let $\Phi\colon \mathbb{R}_+ \to \mathbb{R}_+$ be increasing and convex. Then

$$\mathbb{E}_{\boldsymbol{x}}\, \Phi\left(\left\|\sum\nolimits_{i=1}^{n} (\boldsymbol{x}_i - \mathbb{E}\, \boldsymbol{x}_i)\right\|\right) \le \mathbb{E}_{\boldsymbol{x},\varepsilon}\, \Phi\left(2\left\|\sum\nolimits_{i=1}^{n} \varepsilon_i (\boldsymbol{x}_i - \mathbb{E}\, \boldsymbol{x}_i)\right\|\right), \tag{7.1}$$

where (ε_i) comprises iid Rademacher variables, independent from $(\boldsymbol{x}_i)$.

Recall that a Rademacher random variable $\varepsilon \sim \textsc{uniform}\{\pm 1\}$.

Later in this lecture, we will employ this result when the summands $\boldsymbol{x}_i$ are real-valued and the convex function $\Phi(t) = t^p$. For an example where symmetrization is used with matrix-valued random variables, see Problem Set 2.

Proof. For each index i, let $\boldsymbol{x}_i'$ be an independent copy of $\boldsymbol{x}_i$. The function $\boldsymbol{x} \mapsto \Phi(\|\boldsymbol{x}\|)$ is convex, so Jensen's inequality gives

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}\, \Phi\left(\left\|\sum\nolimits_i (\boldsymbol{x}_i - \mathbb{E}\, \boldsymbol{x}_i)\right\|\right) &= \mathbb{E}_{\boldsymbol{x}}\, \Phi\left(\left\|\mathbb{E}_{\boldsymbol{x}'} \sum\nolimits_i (\boldsymbol{x}_i - \boldsymbol{x}_i')\right\|\right) \\ &\le \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'}\, \Phi\left(\left\|\sum\nolimits_i (\boldsymbol{x}_i - \boldsymbol{x}_i')\right\|\right).\end{aligned}$$

Because $\boldsymbol{x}_i - \boldsymbol{x}_i' \sim \boldsymbol{x}_i' - \boldsymbol{x}_i$ and the family $(\boldsymbol{x}_i)$ is independent, we have

$$\sum\nolimits_i (\boldsymbol{x}_i - \boldsymbol{x}_i') \sim \sum\nolimits_i \zeta_i (\boldsymbol{x}_i - \boldsymbol{x}_i') \quad \text{for all } \textit{fixed} \text{ signs } \zeta_i \in \{\pm 1\}.$$

Therefore, we can choose *random* signs ε_i and average to obtain

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'} \Phi\left(\left\|\sum\nolimits_i (\boldsymbol{x}_i - \boldsymbol{x}_i')\right\|\right) &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}',\boldsymbol{\epsilon}}\, \Phi\left(\left\|\sum\nolimits_i \varepsilon_i (\boldsymbol{x}_i - \boldsymbol{x}_i')\right\|\right) \\ &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}',\boldsymbol{\epsilon}}\, \Phi\left(\left\|\sum\nolimits_i \varepsilon_i (\boldsymbol{x}_i - \mathbb{E}\, \boldsymbol{x}_i) - \sum\nolimits_i \varepsilon_i (\boldsymbol{x}_i' - \mathbb{E}\, \boldsymbol{x}_i')\right\|\right) \\ &\le \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}',\boldsymbol{\epsilon}}\, \Phi\left(\left\|\sum\nolimits_i \varepsilon_i (\boldsymbol{x}_i - \mathbb{E}\, \boldsymbol{x}_i)\right\| + \left\|\sum\nolimits_i \varepsilon_i (\boldsymbol{x}_i' - \mathbb{E}\, \boldsymbol{x}_i')\right\|\right) \\ &\le \mathbb{E}_{\boldsymbol{x},\boldsymbol{x}',\boldsymbol{\epsilon}}\left[\frac{1}{2}\Phi\left(2\left\|\sum\nolimits_i \varepsilon_i (\boldsymbol{x}_i - \mathbb{E}\, \boldsymbol{x}_i)\right\|\right) + \frac{1}{2}\Phi\left(2\left\|\sum\nolimits_i \varepsilon_i (\boldsymbol{x}_i' - \mathbb{E}\, \boldsymbol{x}_i')\right\|\right)\right] \\ &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{\epsilon}}\, \Phi\left(2\left\|\sum\nolimits_i \varepsilon_i (\boldsymbol{x}_i - \mathbb{E}\, \boldsymbol{x}_i)\right\|\right).\end{aligned}$$

In sequence, we have used the triangle inequality, convexity of Φ, and identical distribution. This is stated inequality. ∎

Exercise 7.2 (Symmetrization: Lower bound). The upper bound in Theorem 7.1 has a matching lower bound, namely

$$\mathbb{E}_{\boldsymbol{x},\varepsilon}\, \Phi\left(\frac{1}{2}\left\|\sum\nolimits_i \varepsilon_i (\boldsymbol{x}_i - \mathbb{E}\, \boldsymbol{x}_i)\right\|\right) \le \mathbb{E}_{\boldsymbol{x}}\, \Phi\left(\left\|\sum\nolimits_i (\boldsymbol{x}_i - \mathbb{E}\, \boldsymbol{x}_i)\right\|\right).$$

This shows that, by symmetrizing, we change the pth moment $\left(\mathbb{E}_{\boldsymbol{x}} \left\|\sum_i (\boldsymbol{x}_i - \mathbb{E}\, \boldsymbol{x}_i)\right\|^p\right)^{1/p}$ by at most a factor of 4.

Exercise 7.3 (Symmetrization: Upper bound without centering). By modifying the proof of Theorem 7.1, show that we can obtain an upper bound involving *uncentered* summands:

$$\mathbb{E}_{\boldsymbol{x}} \Phi\left(\left\|\sum_i (\boldsymbol{x}_i - \mathbb{E}\,\boldsymbol{x}_i)\right\|\right) \le \mathbb{E}_{\boldsymbol{x},\varepsilon} \Phi\left(2\left\|\sum_i \varepsilon_i \boldsymbol{x}_i\right\|\right).$$

Is there a matching lower bound?

7.2 Khintchine's inequality

At first sight, it is not clear how to exploit the random signs to bound the right-hand side in (7.1). Indeed, bounding the expectation $\mathbb{E}_{\boldsymbol{x}}$ for fixed signs ε_i is as hard as bounding the left-hand side in (7.1). The key is to *condition* on the values of $(\boldsymbol{x}_i)$ and compute the expectation only with respect to the random signs (ε_i). This approach allows us to exploit special bounds for the moments of Rademacher series.

In this section, we will prove Khintchine's inequality, which bounds the polynomial moments of a real-valued Rademacher series $\sum_i \varepsilon_i a_i$.

Theorem 7.4 (Khintchine). Consider the Rademacher series $Z = \sum_{i=1}^n \varepsilon_i a_i$ where (ε_i) are iid Rademacher variables and $\boldsymbol{a} = (a_1, \dots, a_n) \in \mathbb{R}^n$. For $p = 1$ and $p \ge 2$, we have

$$\left(\mathbb{E}\,|Z|^{2p}\right)^{1/2p} \le \sqrt{2p-1}\,\|\boldsymbol{a}\|_2.$$

This bound implies that Z is subgaussian with variance proxy proportional to $\|\boldsymbol{a}\|_2^2$ using Exercise 3(b) on Problem Set 1.

The proof relies on a useful numerical inequality, which we have seen before in the exchangeable bound for entropy. Let us isolate the general form of this result.

Exercise 7.5 (Mean value inequality). Suppose that $\varphi : \mathbb{R} \to \mathbb{R}$ is continuously differentiable and φ' is convex. For all $a, b \in \mathbb{R}$, we have

$$(a-b)(\varphi(a) - \varphi(b)) \le \tfrac{1}{2}(a-b)^2(\varphi'(a) + \varphi'(b)).$$

Establish a similar result under the alternative assumption that φ is convex, so that φ' is increasing.

Proof of Khintchine's inequality. For $p = 1$ we have an equality (check it!), so we may assume $p \ge 2$. To simplify matters slightly, we will also assume that $p \in \mathbb{N}$. The case of non-integer p is left as an exercise.

Consider the function $\varphi(t) = t^{2p-1}$, and note that φ' is convex on all of $\mathbb{R}$. Let (ε_i') be an independent copy of (ε_i), and define the exchangeable counterparts $Z^{(i)} = \varepsilon_i' a_i + \sum_{j \ne i} \varepsilon_j a_j$. Imitating the argument behind the discrete MLSI, we see that

The idea behind this proof is to mimic the familiar computation of the $2p$th moment of a standard normal random variable. That argument uses (repeated) integration by parts. The approach here is a kind of discrete analog of integration by parts.

$$\begin{aligned}
\mathbb{E}\,Z^{2p} &= \mathbb{E}[Z \cdot Z^{2p-1}] \\
&= \frac{1}{2}\sum_{i=1}^n \mathbb{E}\left[(Z - Z^{(i)})(\varphi(Z) - \varphi(Z^{(i)}))\right] \\
&\le \frac{1}{4}\sum_{i=1}^n \mathbb{E}[(Z - Z^{(i)})^2(\varphi'(Z) + \varphi'(Z^{(i)}))] \\
&= \frac{1}{2}\sum_{i=1}^n \mathbb{E}[(Z - Z^{(i)})^2 \varphi'(Z)] \\
&= \frac{1}{2}\,\mathbb{E}\left[\left(\sum_{i=1}^n \mathbb{E}_{\varepsilon_i'}(Z - Z^{(i)})^2\right)\varphi'(Z)\right].
\end{aligned}$$

The second line follows from a simple computation based on the fact that $(Z, Z^{(i)})$ is exchangeable. The inequality follows from Exercise 7.5. Afterward, we use exchangeability again to simplify the expression.

To complete the argument, we make use of the structure of the Rademacher series. Observe that

$$Z - Z^{(i)} = (\varepsilon_i - \varepsilon_i')a_i \quad \text{and} \quad \mathbb{E}_{\varepsilon_i'}(Z - Z^{(i)})^2 = 2a_i^2.$$

Since $\varphi'(Z) = (2p-1)Z^{2(p-1)}$, we have

$$\begin{aligned}\mathbb{E}\, Z^{2p} &\le \frac{1}{2}\,\mathbb{E}\left[\left(\sum\nolimits_{i=1}^{n} \mathbb{E}_{\varepsilon_i'}(Z - Z^{(i)})^2\right)\varphi'(Z)\right] \\ &= (2p-1)\|\boldsymbol{a}\|_2^2\left(\mathbb{E}\, Z^{2(p-1)}\right) \\ &\le (2p-1)\|\boldsymbol{a}\|_2^2\left(\mathbb{E}\, Z^{2p}\right)^{1-1/p}.\end{aligned}$$

The last inequality is Lyapunov's. It is straightforward to solve for $\mathbb{E}\, Z^{2p}$, which gives the stated result. ∎

The proof of Theorem 7.4 is an example of the *moment comparison method*, in which we bound a moment by a power of itself. We will use this method again in this lecture and in Problem Set 2.

Exercise 7.6 (Khintchine: Remaining values of p). Extend the proof of Theorem 7.4 to non-integer $p \ge 1$. Develop a bound for the case $p \in (1, 2)$ using a variant of the mean value inequality (Exercise 7.5).

Exercise 7.7 (Khintchine: Best constants for $p \ge 1$). The best constant in Theorem 7.4 is

$$\mathbb{E}\, Z^{2p} \le (2p-1)!!\,\|\boldsymbol{a}\|_2^{2p} \quad \text{for all } p \ge 1.$$

The sharp constants were obtained by Haagerup [Haa81].

Extend the proof given above to obtain the best constant for *integer* values of $p \ge 2$. We recognize that $(2p-1)!! = \mathbb{E}[\gamma^{2p}]$, the $2p$th moment of a standard normal random variable γ. Asymptotically, $(2p-1)!! \sim [(2p-1)/\mathrm{e}]^p$, so we have only lost a small constant factor.

Problem 7.8 (Khintchine: Lower bound). For $p \ge 1/2$, show that there is a constant c_p for which

$$c_p\,\|\boldsymbol{a}\|_2 \le \left(\mathbb{E}\left|\sum\nolimits_{i=1}^{n} \varepsilon_i a_i\right|^{2p}\right)^{1/2p}.$$

The sharp result for $p = 1/2$ is due to Szarek; see Latała and Oleszkiewicz [LO94].

For $p \ge 1$, the constant $c_p = 1$. The optimal constant $c_{1/2} = 1/\sqrt{2}$. **Hint:** Prove the result for $p = 1/2$ by applying Hölder's inequality to the case $p = 1$.

Problem 7.9 (Khintchine–Kahane). There is an extension of Khintchine's inequality, due to Kahane, that holds in any normed linear space. Consider a family $(\boldsymbol{a}_i)$ of fixed vectors and a Rademacher sequence (ε_i). For $p > 0$, prove that

$$\left(\mathbb{E}\left\|\sum\nolimits_{i=1}^{n} \varepsilon_i \boldsymbol{a}_i\right\|^p\right)^{1/p} \le \text{const} \cdot \sqrt{p} \cdot \mathbb{E}\left\|\sum\nolimits_{i=1}^{n} \varepsilon_i \boldsymbol{a}_i\right\|.$$

Hint: Check that $\boldsymbol{\epsilon} \mapsto \|\sum_i \varepsilon_i \boldsymbol{a}_i\|$ is convex and Lipschitz, and show that the Lipschitz constant is comparable with the expectation on the right-hand side. (Not easy!) Use the convex MLSI to obtain a subgaussian upper tail bound and integrate.

7.3 Polynomial moment bounds

We are now in a position to prove moment bounds by combining symmetrization with Khintchine's inequality.

7.3.1 A Hoeffding-type moment bound

We begin with the polynomial moment analog of Hoeffding's normal concentration inequality.

Theorem 7.10 (Moment inequality: Hoeffding form). Consider an independent sum $Z = \sum_{i=1}^n (X_i - \mathbb{E}\, X_i)$ of real random variables that satisfy $a_i \le X_i \le b_i$ for each i. For $p = 1$ and $p \ge 2$, we have

$$\left(\mathbb{E}\,|Z|^{2p}\right)^{1/2p} \le 2\sqrt{2p-1}\sqrt{v} \lesssim \sqrt{pv}, \quad \text{where} \quad v = \sum\nolimits_{i=1}^n (b_i - a_i)^2$$

Up to constants, v is the variance proxy that appears in Hoeffding's concentration inequality. The use of symmetrization and Khintchine's inequality in the proof yields inferior constants. The argument is just intended as a demonstration of a more general approach.

Proof. We symmetrize the sum (Theorem 7.1) and then apply Khintchine's inequality (Theorem 7.4).

$$\begin{aligned}
\left(\mathbb{E}\,|Z|^{2p}\right)^{1/2p} &\le 2\left(\mathbb{E}_{\boldsymbol{X},\boldsymbol{\epsilon}} \left|\sum\nolimits_i \varepsilon_i (X_i - \mathbb{E}\, X_i)\right|^{2p}\right)^{1/2p} \\
&\le 2\sqrt{2p-1}\left(\mathbb{E}_{\boldsymbol{X}} \left|\sum\nolimits_i (X_i - \mathbb{E}\, X_i)^2\right|^{p}\right)^{1/2p} \\
&\le 2\sqrt{2p-1}\sqrt{\sum\nolimits_i (b_i - a_i)^2}.
\end{aligned}$$

In the last line, we apply the obvious fact that $(X_i - \mathbb{E}\, X_i)^2 \le (b_i - a_i)^2$. ∎

As with the Hoeffding variance and cgf bounds, this estimate does not use any information about the X_i beyond their extreme values. It tends to be accuate only when the distributions of the X_i concentrates its mass near the endpoints of the interval $[a_i, b_i]$. Recall from Exercise 3 in Problem Set 1 that a bound on the pth moment of the form $\lesssim \sqrt{pv}$ is equivalent to subgaussian tails with variance proxy v (up to a constant). Thus, the above moment bound is equivalent with the bound from Hoeffding's normal concentration inequality.

7.3.2 A Chernoff-type moment bound

Next, we present some new moment bounds. These results will provide more refined information than the inequalities based on the Laplace transform method. These moment estimates are usually called *Rosenthal inequalities*, but the versions here are more closely associated with Nagaev, Pinelis, and Utev.

First, we prove the polynomial moment analog of Chernoff's bound for an independent sum of positive random variables. In contrast with Chernoff's inequality, we will not assume that the summands are uniformly bounded.

Theorem 7.11 (Moment bound: Chernoff form). Consider a family $(X_1, \dots, X_n)$ of

independent *positive* random variables. Then for $p = 1$ or $p \geq 2$, we have

$$\left(\mathbb{E}\left|\sum_i X_i\right|^{2p}\right)^{1/2p} \leq \left[\left(\sum_i \mathbb{E}\, X_i\right)^{1/2} + 2\sqrt{2p-1}\left(\mathbb{E}\max_i X_i^{2p}\right)^{1/4p}\right]^2$$
$$\lesssim \left(\sum_i \mathbb{E}\, X_i\right) + p\left(\mathbb{E}\max_i X_i^{2p}\right)^{1/4p}.$$

A similar bound holds for $p \in (1, 2)$ with slightly different constants.

Theorem 7.11 implies that the sum has a subexponential tail, with decay on the scale of the typical size of the largest summand. This quantity is never larger than a uniform bound on all of the summands, and it can account for situations where the summands have significantly different sizes.

Proof. We center, symmetrize, and apply Khintchine's inequality to obtain a moment comparison. Here are the details. Write $Z = \sum_{i=1}^n X_i$.

$$\begin{aligned}
\left(\mathbb{E}\, Z^{2p}\right)^{1/2p} &\leq (\mathbb{E}\, Z) + \left[\mathbb{E}\left|\sum_i (X_i - \mathbb{E}\, X_i)\right|^{2p}\right]^{1/2p} \\
&\leq (\mathbb{E}\, Z) + 2\left[\mathbb{E}_{\boldsymbol{X},\boldsymbol{\epsilon}}\left|\sum_i \varepsilon_i X_i\right|^{2p}\right]^{1/2p} \\
&\leq (\mathbb{E}\, Z) + 2\sqrt{2p-1}\left[\mathbb{E}\left(\sum_i X_i^2\right)^p\right]^{1/2p} \\
&\leq (\mathbb{E}\, Z) + 2\sqrt{2p-1}\left[\,\mathbb{E}\left((\max_i X_i)\cdot Z\right)^p\,\right]^{1/2p} \\
&\leq (\mathbb{E}\, Z) + 2\sqrt{2p-1}\left(\mathbb{E}\max_i X_i^{2p}\right)^{1/4p}\left(\mathbb{E}\, Z^{2p}\right)^{1/4p}.
\end{aligned}$$

The first bound follows from the triangle inequality for the L_{2p} norm. The second and third lines require symmetrization (Exercise 7.3) and Khintchine's inequality (Theorem 7.4). We bound the sum using Hölder's inequality, exploiting the fact that the summands are positive. Finally, we apply the Cauchy–Schwarz inequality.

Observe that we have arrived at a moment comparison of the form

$$E^2 \leq a + bE \quad \text{where} \quad E = \left(\mathbb{E}\, Z^{2p}\right)^{1/4p}.$$

As an exercise, you may confirm that the solution to this quadratic inequality satisfies $E \leq \sqrt{a} + b$. Therefore,

$$\left(\mathbb{E}\, Z^{2p}\right)^{1/4p} \leq (\mathbb{E}\, Z)^{1/2} + \left(\mathbb{E}\max_i X_i^{2p}\right)^{1/4p}.$$

Square this expression to complete the proof. ∎

Exercise 7.12 (Quadratic inequalities). Suppose that $E^2 \leq a + bE$ for positive a, b. Deduce that $E \leq \sqrt{a} + b$.

7.3.3 A Bernstein-type moment bound

Finally, we consider an independent sums of centered random variables, and we prove a polynomial moment analog of Bernstein's concentration inequality.

Theorem 7.13 (Moment inequality: Bernstein form). Consider a family $(X_1, \dots, X_n)$ of independent *centered*, real random variables. For $p \geq 2$, we have

$$\left(\mathbb{E}\left|\sum_i X_i\right|^{2p}\right)^{1/2p} \lesssim \sqrt{p \operatorname{Var}\left(\sum_i X_i\right)} + p\left(\mathbb{E}\max_i X_i^{2p}\right)^{1/2p}.$$

This bound implies that the independent sum has subgaussian tails on the scale of the variance, and subexponential tails on the scale of the typical size of the largest summand. This result is qualitatively similar to Bernstein's inequality, but we no longer need to assume that the summands are uniformly bounded.

Proof. For simplicity, we will assume that $p = 2$ or $p \geq 4$ and obtain precise estimates for the constants. The argument follows our standard pattern: symmetrize and invoke Khintchine's inequality. Indeed, since the summands are centered,

$$\begin{aligned}\left(\mathbb{E}\left|\sum_i X_i\right|^{2p}\right)^{1/2p} &\leq 2\left(\mathbb{E}_{\boldsymbol{X},\boldsymbol{\epsilon}}\left|\sum_i \epsilon_i X_i\right|^{2p}\right)^{1/2p} \\ &\leq 2\sqrt{2p-1}\left[\mathbb{E}\left(\sum_i X_i^2\right)^p\right]^{1/2p} \\ &\leq 2\sqrt{2p-1}\left[\left(\sum_i \mathbb{E}\, X_i^2\right)^{1/2} + 2\sqrt{p-1}\left(\mathbb{E}\max_i X_i^{2p}\right)^{1/2p}\right].\end{aligned}$$

The last inequality is Theorem 7.11, applied to control the pth moment of a sum $\sum_i X_i^2$ of independent, positive random variables. ∎

Polynomial moment inequalities for nonlinear functions also hold, but they are significantly harder to prove. Symmetrization and Khintchine's inequality apply in a wide variety of circumstances. On Problem Set 2, we will see that they allow us to prove polynomial moment inequalities for an independent sum of random matrices. Later, we will use related methods to study the supremum of an empirical process.

Lecture bibliography

[Haa81] U. Haagerup. "The best constants in the Khintchine inequality". eng. In: *Studia Mathematica* 70.3 (1981), pages 231–283. URL: http://eudml.org/doc/218383.

[LO94] R. Latała and K. Oleszkiewicz. "On the best constant in the Khinchin–Kahane inequality". In: *Studia Math.* 109.1 (1994), pages 101–104.

[van16] R. van Handel. "Probability in High Dimensions". APC 550 Lecture Notes, Princeton Univ. 2016. URL: https://web.math.princeton.edu/~rvan/APC550.pdf.

8. Matrix Concentration

Date: 28 January 2021 **Scribe:** Jing Yu

Agenda:
1. Independent Sum Model
2. Matrix Laplace Transform Method
3. Matrix cumulant generating functions are subadditive
4. Matrix Bernstein

The theory of matrix concentration is halfway in between the study of concentration inequalities and the study of suprema, which is the next main topic in this course. These results show that a random matrix concentrates around its expected value, with deviations measured in the spectral norm. As a consequence, we can obtain bounds for the expected norm of the random matrix. The most direct formulation of the spectral norm involves a supremum, which can also be studied with completely different tools.

We will start with an independent sum model for a random matrix. Next, we introduce a matrix version of the Laplace transform method. We will then give a partially proof of a fundamental result which states that matrix cumulant generating functions are subadditive. As an example of the general methodology, we will establish the matrix Benstein inequality, which has become a major tool for numerical analysis, data science, and statistics in the last decade.

8.1 Introduction to matrix concentration

We will be interested in the maximum eigenvalue $\lambda_{\max}(\boldsymbol{A})$ of an Hermitian matrix $\boldsymbol{A} \in \mathbb{H}_d$. Let us recall some basic properties of the maximum eigenvalue map:

- $\lambda_{\max}(\boldsymbol{A})$ is a real number.
- $\lambda_{\max}(\boldsymbol{A}) = \sup_{\|\boldsymbol{u}\|_2=1} \boldsymbol{u}^* \boldsymbol{A} \boldsymbol{u}$, according to the Rayleigh theorem.
- $\boldsymbol{A} \mapsto \lambda_{\max}(\boldsymbol{A})$ is convex and 1-Lipschitz function with respect to $\|\cdot\|_{\mathrm{F}}$.

Consider a random Hermitian matrix $\boldsymbol{X} \in \mathbb{H}_d$. In earlier lectures, we have established some concentration inequalities that lead to bounds on the quantity

$$\mathbb{P}\left\{|\lambda_{\max}(\boldsymbol{X}) - \mathbb{E}\,\lambda_{\max}(\boldsymbol{X})| \geq t\right\}.$$

For example, since $\lambda_{\max}(\cdot)$ is a convex and 1-Lipschitz function, we can invoke the convex Poincaré inequality (Lecture 3) or the convex MLSI (Lecture 6) to study the upper tail behavior of $\lambda_{\max}(\boldsymbol{X})$. These results apply, for example, when the entries of $\boldsymbol{X}$ are independent, bounded random variables.

Note, however, that our scalar concentration inequalities provide no information about $\mathbb{E}\,\lambda_{\max}(\boldsymbol{X})$, the point at which the random variable $\lambda_{\max}(\boldsymbol{X})$ concentrates. In this lecture, we will develop some tools that yield upper bounds for $\mathbb{E}\,\lambda_{\max}(\boldsymbol{X})$.

Matrix concentration inequalities describe the concentration of a random Hermitian matrix around its expectation, as a matrix, with deviations measured using the maximum eigenvalue. That is, they inform us about the quantity

$$\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X}) \geq t\right\}.$$

Note the contrast with the usual concentration quantity $\mathbb{P}\left\{\lambda_{\max}(\boldsymbol{X}) - \mathbb{E}\,\lambda_{\max}(\boldsymbol{X}) \geq t\right\}$ that we have been studying. The latter tells us how far the random maximum eigenvalue $\lambda_{\max}(\boldsymbol{X})$ falls from its expectation. The former treats the random matrix as the random

variable and asks how well this random variable concentrates around its expectation, as measured by the maximum eigenvalue function.

Using integration by parts, tail bounds for $\lambda_{\max}(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X})$ also lead to inequalities for the expectation:

$$\mathbb{E}\,\lambda_{\max}(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X}) \le \text{(bound)}.$$

Invoking Weyl's inequality for the maximum eigenvalue, we can obtain estimates of the form

$$\mathbb{E}\,\lambda_{\max}(\boldsymbol{X}) = \lambda_{\max}(\mathbb{E}\,\boldsymbol{X}) \pm \text{(bound)}.$$

This approach leads to nontrivial results on the expectation of the maximum eigenvalue that may be very hard to achieve by other means.

Bounds for $\lambda_{\max}(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X})$ have many implications, including

- Control on the eigenvalues $\boldsymbol{X}$ as compared with those of $\mathbb{E}\,\boldsymbol{X}$.
- Control of the eigenvectors of $\boldsymbol{X}$ as compared with those of $\mathbb{E}\,\boldsymbol{X}$, at least for isolated eigenvalues.
- Control on all linear functionals $\boldsymbol{X}$ as compared with those of $\mathbb{E}\,\boldsymbol{X}$.

Altogether, matrix concentration inequalities provide a powerful tool for studying properties of random matrices.

8.2 The independent sum model

In our study of matrix concentration, we need to introduce a model for a random matrix that is flexible enough to handle many applications while remaining analytically tractable. Taking inspiration from classical probability, we will consider a random matrix that can be expressed as a sum of independent random matrices.

Consider a statistically independent family $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n) \in \mathbb{H}_d$ of random, Hermitian matrices with common dimension d. Construct the random matrix

$$\boldsymbol{Y} = \sum\nolimits_{i=1}^{n} \boldsymbol{X}_i.$$

Our goal is to exploit properties of the summands $\boldsymbol{X}_i$ to obtain probabilistic bounds for

$$\lambda_{\max}(\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y}) \quad \text{and} \quad \lambda_{\min}(\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y}).$$

We will focus on tail bounds, but related methods lead directly to expectation bounds.

Although it may not be obvious, the independent sum model is widely applicable. As a first example, let us consider the plug-in estimator for the sample covariance matrix of a random vector. Suppose that $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$ are iid copies of a *centered* random vector $\boldsymbol{x} \in \mathbb{R}^d$. The standard sample covariance estimator is the random matrix

$$\boldsymbol{Y} = \frac{1}{n} \sum\nolimits_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^*.$$

It is easy to see that $\mathbb{E}\,\boldsymbol{Y} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^*]$, so the sample covariance is an unbiased estimator for the true covariance. The model posits that the samples $\boldsymbol{x}_i$ are independent, so we can view the sample covariance as an independent sum of random Hermitian matrices.

There are many other examples that fit in this framework. For instance, the Laplacian of a random graph can be written as the sum of the Laplacian matrices associated with each edge of the graph; when the edges are independent, we can express the Laplacian using the random sum model.

There are many other examples involving rectangular matrices. A random matrix with independent entries can be rewritten as an independent sum of sparse random

matrices. Likewise, a random matrix with independent columns can be written as a sum of independent column-sparse random matrices. We will see that results for random rectangular matrices follow as a formal consequence of our results for random Hermitian matrices.

8.3 The Laplace transform method for random matrices

To develop matrix concentration inequalities for independent sums, we will imitate the scalar Laplace transform method that has already proven so useful. These ideas were formulated in a 2002 quantum information paper [AW02] by Ahlswede and Winter.

8.3.1 Standard matrix functions

First, let us recall the definition of a standard matrix function. Every Hermitian matrix $\boldsymbol{A} \in \mathbb{H}_d$ has a unique spectral resolution:

$$\boldsymbol{A} = \sum\nolimits_{\lambda \in \operatorname{spec}(A)} \lambda \boldsymbol{P}_\lambda,$$

where $\operatorname{spec}(\boldsymbol{A}) \subset \mathbb{C}$ contains the (distinct) eigenvalues of $\boldsymbol{A}$, and $\boldsymbol{P}_\lambda$ is the orthogonal projector onto the eigenspace associates with eigenvalue λ.

Using the spectral resolution, we can lift a scalar function $f : \mathbb{R} \to \mathbb{R}$ to a matrix function $f : \mathbb{H}_d \to \mathbb{H}_d$. Indeed, we define

$$f(\boldsymbol{A}) \coloneqq \sum\nolimits_{\lambda \in \operatorname{spec}(A)} f(\lambda) \boldsymbol{P}_\lambda \in \mathbb{H}_d.$$

In other words, we simply apply the scalar function to each eigenvalue of the Hermitian matrix without changing the eigenspaces.

Equivalently, if we have an eigenvalue factorization $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^*$, then we may define $f(\boldsymbol{A}) = \boldsymbol{U} f(\boldsymbol{\Lambda}) \boldsymbol{U}^*$. In this definition, we apply the function f to each diagonal element of the diagonal matrix $\boldsymbol{\Lambda}$.

Whenever we apply a familiar scalar function to an Hermitian matrix, we are referring to the associated standard matrix function. Common examples include the exponential function and integer powers. For a positive-definite matrix, we can form the matrix logarithm or compute noninteger powers.

8.3.2 Generating functions for random matrices

For scalar random variables, we constructed generating functions that pack up information about the moments of the random variables. We can make similar definitions in the matrix setting.

Definition 8.1 (Matrix moment generating function and cumulant generating function). Let $\boldsymbol{X} \in \mathbb{H}_d$ be a random Hermitian matrix. The *matrix moment generating function* (mgf) and *matrix cumulant generating function* (cgf) are respectively defined as

$$\begin{aligned} \boldsymbol{M}_{\boldsymbol{X}}(\theta) &\coloneqq \mathbb{E} \exp(\theta \boldsymbol{X}); \\ \boldsymbol{\Xi}_{\boldsymbol{X}}(\theta) &\coloneqq \log \mathbb{E} \exp(\theta \boldsymbol{X}). \end{aligned}$$

The parameter $\theta \in \mathbb{R}$.

In this definition, the matrix exponential and logarithm are standard matrix functions, as discussed in the previous subsection. As in the scalar setting, we can

make a formal Taylor series expansion of the mgf

$$\boldsymbol{M}_{\boldsymbol{X}}(\theta) = \sum\nolimits_{p=0}^{\infty} \frac{\theta^p}{p!} \mathbb{E}\, \boldsymbol{X}^p.$$

Similarly, the matrix cgf has the expansion

$$\boldsymbol{\Xi}_{\boldsymbol{X}}(\theta) = \theta(\mathbb{E}\, \boldsymbol{X}) + \frac{\theta^2}{2}(\mathbb{E}\, \boldsymbol{X}^2 - (\mathbb{E}\, \boldsymbol{X})^2) + \cdots.$$

The second-order term in the cgf is a matrix analog of the variance.

8.3.3 Matrix Laplace transform method

We can obtain tail bounds for the maximum eigenvalue of a random Hermitian matrix in terms of the matrix mgf. This simple result is called the *matrix Laplace transform method*. The idea is due to Ahlswede & Winter [AW02], and the easy proof here is due to Oliveira [Oli10].

Proposition 8.2 (Matrix Laplace transform method). Let $\boldsymbol{Y} \in \mathbb{H}_d$ be a random Hermitian matrix. Then

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) > t\} \le \inf_{\theta>0} \mathrm{e}^{-\theta t} \cdot \mathbb{E}\, \mathrm{tr}\, \exp(\theta \boldsymbol{Y});$$
$$\mathbb{P}\{\lambda_{\min}(\boldsymbol{Y}) < t\} \le \inf_{\theta<0} \mathrm{e}^{-\theta t} \cdot \mathbb{E}\, \mathrm{tr}\, \exp(\theta \boldsymbol{Y}).$$

Proof. We consider the tail bound for the maximum eigenvalue. Fix $\theta > 0$. By Markov's inequality,

$$\begin{aligned}
\mathbb{P}\{\lambda_{\max}(\boldsymbol{Y}) > t\} &= \mathbb{P}\left\{\mathrm{e}^{\theta \lambda_{\max}(\boldsymbol{Y})} \ge \mathrm{e}^{\theta t}\right\} \\
&\le \mathrm{e}^{-\theta t}\, \mathbb{E}\, \mathrm{e}^{\theta \lambda_{\max}(\boldsymbol{Y})} \\
&= \mathrm{e}^{-\theta t}\, \mathbb{E}\, \lambda_{\max}(\mathrm{e}^{\theta \boldsymbol{Y}}) \\
&\le \mathrm{e}^{-\theta t} \cdot \mathbb{E}\, \mathrm{tr}\, \mathrm{e}^{\theta \boldsymbol{Y}} \\
&\le \inf_{\theta>0} \mathrm{e}^{-\theta t} \cdot \mathbb{E}\, \mathrm{tr}\, \exp(\theta \boldsymbol{Y}).
\end{aligned}$$

The second equality is due to spectral mapping theorem for a function on Hermitian matrices. The second-to-last inequality holds because the eigenvalues of $\mathrm{e}^{\theta \boldsymbol{Y}}$ are positive, so the maximum eigenvalue is bounded above by the trace. Finally, we optimize over $\theta > 0$. ∎

Exercise 8.3 (Minimum eigenvalue). Establish the tail bound for the minimum eigenvalue that is stated in Proposition 8.2. **Hint:** $\lambda_{\min}(\boldsymbol{A}) = -\lambda_{\max}(-\boldsymbol{A})$.

Exercise 8.4 (Expectation bounds). Let $\boldsymbol{Y} \in \mathbb{H}_d$ be a random Hermitian matrix. Show that

$$\mathbb{E}\, \lambda_{\max}(\boldsymbol{Y}) \le \inf_{\theta>0} \theta^{-1} \log \mathrm{tr}\, \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{Y}};$$
$$\mathbb{E}\, \lambda_{\min}(\boldsymbol{Y}) \le \inf_{\theta>0} \theta^{-1} \log \mathrm{tr}\, \mathbb{E}\, \mathrm{e}^{-\theta \boldsymbol{Y}}.$$

> **Aside:** The matrix Laplace transform method is a relatively recent technical innovation. Nevertheless, there is an older line of research that pursues the idea that we can establish moment inequalities for random matrices that parallel similar results for random scalars. Françoise Lust-Piquard [LP86] obtained the first matrix moment inequality, a remarkable result known as the the noncommutative Khintchine inequality. (See Problem Set 2 for a modern proof.) Later, Gilles Pisier and coauthors recognized that many inequalities for scalar martingales extend to matrix martingales. These works provide the intellectual foundations for the field of matrix concentration, and they precede the matrix Laplace transform method.

8.4 Subadditivity of matrix cgfs

To activate the matrix Laplace transform method, we need to obtain bounds for the trace of the matrix mgf. As in the scalar setting, we will focus on the case of an independent sum of random matrices.

8.4.1 The failure of the matrix mgf

Drawing inspiration from the scalar setting, we would like to develop an analog of the fact that the scalar mgf of an independent sum is multiplicative. That is, for an independent family (X_i) of real random variables,

$$\mathbb{E}\, \mathrm{e}^{\sum_i X_i} \;\; \mathbb{E} \prod_i \mathrm{e}^{X_i} = \prod_i \mathbb{E}\, \mathrm{e}^{X_i}. \tag{8.1}$$

Let us see what happens when we try to generalize this formula to Hermitian matrices.

The first relation in (8.1) depends on the fact that the exponential of a real sum is the product of exponentials. In contrast, for Hermitian $\boldsymbol{A}, \boldsymbol{B}$,

$$\mathrm{e}^{\boldsymbol{A}+\boldsymbol{B}} = \mathrm{e}^{\boldsymbol{A}}\mathrm{e}^{\boldsymbol{B}} \quad \text{if and only if } \boldsymbol{A}, \boldsymbol{B} \text{ commute.}$$

Commutativity is a severe restriction that we cannot abide. In our application, we are interested in the trace of the exponential, and one may wonder whether the situation is better when we take the trace. In general,

$$\operatorname{tr} \mathrm{e}^{\boldsymbol{A}+\boldsymbol{B}} \neq \operatorname{tr}[\mathrm{e}^{\boldsymbol{A}}\mathrm{e}^{\boldsymbol{B}}].$$

Nevertheless, the Golden–Thompson inequality does provide a substitute:

$$\operatorname{tr} \mathrm{e}^{\boldsymbol{A}+\boldsymbol{B}} \leq \operatorname{tr}[\mathrm{e}^{\boldsymbol{A}}\mathrm{e}^{\boldsymbol{B}}].$$

This bound is always valid. Unfortunately, the Golden–Thompson inequality does not extend to three matrices:

$$\operatorname{tr} \mathrm{e}^{\boldsymbol{A}+\boldsymbol{B}+\boldsymbol{C}} \not\leq |\operatorname{tr}[\mathrm{e}^{\boldsymbol{A}}\mathrm{e}^{\boldsymbol{B}}\mathrm{e}^{\boldsymbol{C}}]|.$$

To manage this issue, Ahlswede & Winter [AW02] and Oliveira [Oli10] apply the Golden–Thompson inequality iteratively. This is an workable approach, but it drains some of the effectiveness from the Laplace transform method.

Problem 8.5 (Triple Golden–Thompson). Find an example to confirm that the Golden–Thompson inequality is not valid for three matrices. **Hint:** Use the Pauli spin matrices.

8.4.2 Subadditivity of matrix cgfs

What went wrong with the computations involving the matrix mgf? One simple observation is that the product of Hermitian matrices is typically not Hermitian, so the eigenvalues of the product can behave in complicated ways. It would be better to works with sums of Hermitian matrices, which remain Hermitian.

To that end, let us focus on the cgf. In the scalar setting, for an independent family (X_i) of real random variables, we have

$$\log \mathbb{E}\, \mathrm{e}^{\sum_i X_i} = \sum_i \log \mathbb{E}\, \mathrm{e}^{X_i}.$$

This formula is a more promising candidate for generalization to the matrix setting. Unfortunately, if we replace the random scalars by random Hermitian matrices, then the cgf identity is no longer valid. Nevertheless, there is an excellent substitute, due to your instructor [Tro15a].

Theorem 8.6 (Subadditivity of matrix cgfs). Consider a family $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$ of independent, random Hermitian matrices in $\mathbb{H}_d$. Then

$$\mathbb{E}\, \operatorname{tr} \exp\left(\sum\nolimits_{i=1}^n \boldsymbol{X}_i\right) \le \operatorname{tr} \exp\left(\sum\nolimits_{i=1}^n \log \mathbb{E}\, \mathrm{e}^{\boldsymbol{X}_i}\right).$$

Equivalently, for all $\theta \in \mathbb{R}$,

$$\operatorname{tr} \exp\left(\Xi_{\sum_{i=1}^n \boldsymbol{X}_i}(\theta)\right) \le \operatorname{tr} \exp\left(\sum\nolimits_{i=1}^n \Xi_{\boldsymbol{X}_i}(\theta)\right).$$

To prove Theorem 8.6, we rely on a deep fact from matrix analysis [Lie73].

Fact 8.7 (Concavity of trace-exp-log). Let $\boldsymbol{H} \in \mathbb{H}_d$ be a fixed Hermitian matrix. The function $\boldsymbol{A} \mapsto \operatorname{tr} \exp(\boldsymbol{H} + \log \boldsymbol{A})$ is concave on positive-definite matrices. ▪

Fact 8.7 is an easy consequence of the convexity of the quantum relative entropy function, another deep result from matrix analysis. See [Tro15a] for a complete proof of Fact 8.7 along these lines. We will use Fact 8.7 by way of a simple corollary.

Corollary 8.8 (Tropp 2010). Let $\boldsymbol{H} \in \mathbb{H}_d$ be fixed, and let $\boldsymbol{X} \in \mathbb{H}_d$ be a random matrix. Then

$$\mathbb{E}\, \operatorname{tr} \exp(\boldsymbol{H} + \boldsymbol{X}) \le \operatorname{tr} \exp(\boldsymbol{H} + \log \mathbb{E}\, \mathrm{e}^{\boldsymbol{X}}).$$

Proof. Combine Fact 8.7 and Jensen's inequality. ∎

With this corollary at hand, we can easily establish Theorem 8.6 by iteration.

Exercise 8.9 (Proof of Theorem 8.6). Use Corollary 8.8 to establish Theorem 8.6.

8.4.3 The master theorem for independent sums

Combining the matrix Laplace transform method with the subadditivity of cgfs, we arrive at an abstract matrix concentration inequality [Tro15a].

Theorem 8.10 (Master theorem). Consider a family $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$ of independent, random Hermitian matrices in $\mathbb{H}_d$. Then

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_{i=1}^n \boldsymbol{X}_i\right) \ge t\right\} \le \inf_{\theta > 0} \mathrm{e}^{-\theta t} \cdot \operatorname{tr} \exp\left(\sum\nolimits_{i=1}^n \Xi_{\boldsymbol{X}_i}(\theta)\right)$$

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum\nolimits_{i=1}^n \boldsymbol{X}_i\right) \le t\right\} \le \inf_{\theta < 0} \mathrm{e}^{-\theta t} \cdot \operatorname{tr} \exp\left(\sum\nolimits_{i=1}^n \Xi_{\boldsymbol{X}_i}(\theta)\right).$$

Proof. Invoke the matrix Laplace transform method (Proposition 8.2). Use the subadditivity of matrix cgfs (Theorem 8.6) to bound the trace of the matrix mgf. ∎

Exercise 8.11 (Expectation bounds). Develop an analog of Theorem 8.10 that directly yields bounds for $\mathbb{E}\,\lambda_{\max}(\boldsymbol{Y})$ and $\mathbb{E}\,\lambda_{\min}(\boldsymbol{Y})$, where $\boldsymbol{Y}$ is an independent sum of Hermitian random matrices.

8.5 The matrix Bernstein inequality

As an example of the master theorem at work, we will prove the matrix Bernstein inequality, which is the single most useful matrix concentration inequality.

Theorem 8.12 (Matrix Bernstein). Consider a family $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$ of independent, random Hermitian matrices in $\mathbb{H}_d$. Assume that $\mathbb{E}\,\boldsymbol{X}_i = \mathbf{0}$ and $\|\boldsymbol{X}_i\| \le B$ for each index i. Define the variance proxy

$$v = \lambda_{\max}\left(\sum\nolimits_{i=1}^{n} \mathbb{E}\,\boldsymbol{X}_i^2\right).$$

Then

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum\nolimits_{i=1}^{n} \boldsymbol{X}_i\right) \ge t\right\} \le d \cdot \exp\left(\frac{-t^2/2}{v + Bt/3}\right).$$

If $d = 1$, so that (X_i) is a family of real random variables, Theorem 8.12 reduces to the scalar Bernstein inequality—up to an including the constants. In this case, the variance proxy is simply the variance of the independent sum. The tail bound reflects normal concentration on the scale of the variance proxy v and exponential concentration on a scale determined by the upper bound B on the summands.

For $d > 1$, we are in the matrix setting. In this case, we need to generalize the ordinary scalar variance. Observe that the variance proxy satisfies

$$v = \lambda_{\max}\left(\mathbb{E}\left(\sum\nolimits_{i=1}^{n} \boldsymbol{X}_i\right)^2\right).$$

This is the "magnitude" of the expected square of the (centered) independent sum, which is a natural generalization for variance in the matrix setting. The only other difference from the scalar setting is the appearance of the dimensional factor d, which arises from a simple bound on the trace. This factor is necessary in general, but it has a very small impact because the tail probability decays exponentially.

8.5.1 Proof of the matrix Bernstein inequality

The key step in the proof is a bound for the matrix cgf of a centered random matrix subject to a uniform spectral norm bound.

Lemma 8.13 (Matrix Bernstein cgf). Consider a random Hermitian matrix $\boldsymbol{X} \in \mathbb{H}_d$ that satisfies $\mathbb{E}\,\boldsymbol{X} = \mathbf{0}$ and $\|\boldsymbol{X}\| \le B$. For $\theta < 3/B$,

$$\log \mathbb{E}\, \mathrm{e}^{\theta \boldsymbol{X}} \preccurlyeq \frac{\theta^2/2}{1 - B|\theta|/3} \cdot \left(\mathbb{E}\,\boldsymbol{X}^2\right).$$

Recall that $\|\cdot\|$ is the spectral norm of a matrix; $\preccurlyeq$ is the psd partial order.

Problem 8.14 (Matrix Bernstein cgf). Prove the matrix Bernstein cgf bound (Lemma 8.13). The proof parallels the scalar argument. You will also need the fact that the matrix logarithm preserves psd inequalities.

Problem 8.15 (Proof of matrix Bernstein). Prove the matrix Bernstein inequality (Theorem 8.12) using the master tail bound and Lemma 8.13. You will also need the fact that the trace exponential is monotone with respect to the psd order.

Problem 8.16 (Matrix Bernstein: Expectation bound). Under the assumptions of Theorem 8.12, show that

$$\mathbb{E}\,\lambda_{\max}\left(\sum\nolimits_{i=1}^{n}\boldsymbol{X}_i\right) \le \sqrt{2v\log d} + \frac{1}{3}B\log d.$$

8.6 Rectangular matrix Bernstein

Finally, let us turn to the problem of producing bounds for the spectral norm of an independent sum of rectangular random matrices. This may seem like a daunting challenge, but it actually follows as a formal consequence of the results we have already developed. As an illustration, we will establish the rectangular version of the matrix Bernstein inequality.

Corollary 8.17 (Matrix Bernstein: Rectangular case). Consider a family $(\boldsymbol{Z}_1,\dots,\boldsymbol{Z}_n)$ of independent, random matrices in $\mathbb{C}^{d_1\times d_2}$. Assume that $\mathbb{E}\,\boldsymbol{Z}_i = 0$ and $\|\boldsymbol{Z}_i\| \le B$ for each index i. Define the variance proxy

$$v = \left\|\sum\nolimits_{i=1}^{n}\mathbb{E}\left(\boldsymbol{Z}_i\boldsymbol{Z}_i^*\right)\right\| \vee \left\|\sum\nolimits_{i=1}^{n}\mathbb{E}\left(\boldsymbol{Z}_i^*\boldsymbol{Z}_i\right)\right\|.$$

Then

$$\mathbb{P}\left\{\left\|\sum\nolimits_{i=1}^{n}\boldsymbol{Z}_i\right\| \ge t\right\} \le (d_1+d_2)\exp\left(\frac{-t^2/2}{v+Bt/3}\right).$$

Furthermore,

$$\mathbb{E}\left\|\sum\nolimits_{i=1}^{n}\boldsymbol{Z}_i\right\| \le \sqrt{2v\log(d_1+d_2)} + \frac{1}{3}B\log(d_1+d_2).$$

Corollary 8.17 is entirely similar to Theorem 8.12, except that the variance proxy now reflects the fact that a rectangular matrix has two distinct squares, the "column square" and the "row square." The dimensional factors are simply the total of the two dimensions of the matrix. In the scalar setting, $d_1 + d_2 = 2$, which is the same constant that arises from applying the union bound to merge an upper and lower tail inequality.

8.6.1 The Hermitian dilation

Let us outline the device that is used to derive Corollary 8.17 from Theorem 8.12.

Definition 8.18 (Hermitian dilation). For $\boldsymbol{C} \in \mathbb{C}^{d_1\times d_2}$, define the *Hermitian dilation*

$$\mathscr{H}(\boldsymbol{C}) := \begin{bmatrix}\mathbf{0} & \boldsymbol{C}\\ \boldsymbol{C}^* & \mathbf{0}\end{bmatrix} \in \mathbb{H}_{d_1+d_2}.$$

We frame an exercise that collects some of the basic facts about this construction.

Exercise 8.19 (Hermitian dilation). Verify that $\boldsymbol{C} \mapsto \mathscr{H}(\boldsymbol{C})$ is a real-linear map. Check that

$$\mathscr{H}(\boldsymbol{C})^2 = \begin{bmatrix}\boldsymbol{C}\boldsymbol{C}^* & \mathbf{0}\\ \mathbf{0} & \boldsymbol{C}^*\boldsymbol{C}\end{bmatrix}.$$

Show that $\lambda_{\max}(\mathscr{H}(\boldsymbol{C})) = \|\boldsymbol{C}\|$.

Problem 8.20 (Rectangular matrix Bernstein). We can obtain Corollary 8.17 by applying Theorem 8.12 to the Hermitian dilation of the sum $\sum_{i=1}^{n}\boldsymbol{Z}_i$. Do it.

Lecture bibliography

[AW02] R. Ahlswede and A. Winter. "Strong converse for identification via quantum channels". In: *IEEE Transactions on Information Theory* 48.3 (2002), pages 569–579.

[Lie73] E. H. Lieb. "Convex trace functions and the Wigner–Yanase–Dyson conjecture". In: *Les rencontres physiciens-mathématiciens de Strasbourg-RCP25* 19 (1973), pages 0–35.

[LP86] F. Lust-Piquard. "Inégalités de Khintchine dans C_p $(1 < p < \infty)$". In: *CR Acad. Sci. Paris* 303 (1986), pages 289–292.

[Oli10] R. Oliveira. "Sums of random Hermitian matrices and an inequality by Rudelson". In: *Electronic Communications in Probability* 15 (2010), pages 203–212.

[Tro15a] J. A. Tropp. "An introduction to matrix concentration inequalities". In: *Found. Trends Mach. Learn.* 8.1–2 (2015), pages 1–230.

II.

suprema

9. Packing and Covering

Date: 2 February 2020 **Scribe:** Riley Murray

This lecture develops deterministic tools for use in the second half of the course. Specifically, we develop concepts of *covering numbers* and *packing numbers* in metric spaces. Section 9.1 begins by reviewing supremum problems encountered earlier in the course as well as basic concepts of metric spaces. Following that we introduce covering and packing problems in metric spaces (Sections 9.3–9.5). Finally, Sections 9.6 and 9.7 present two technical approaches for bounding the covering and packing numbers of a set.

Agenda:
1. Preliminaries
2. Covering numbers
3. Packing numbers
4. Duality
5. Volumetric bounds
6. The empirical method

9.1 Supremum problems

Supremum problems involve the analysis of a random variable Y of the form

$$Y = \sup_{t \in \mathsf{T}} X_t \quad \text{where} \quad (X_t : t \in \mathsf{T}) \text{ is a random process.}$$

Aside: Uncountable suprema can lead to measurability issues. We will return to this point later.

The index set T here is abstract; it could denote time, or something else entirely. For example, the operator norm of a random matrix $\boldsymbol{A}$ can be expressed as the supremum of the random process $(\|\boldsymbol{Au}\|_2 : \|\boldsymbol{u}\|_2 = 1)$ indexed by unit vectors.

We encountered two supremum problems on Problem Set 1. In Problem 1(d), we considered a point set $\mathsf{T} \subseteq \mathbb{R}^n$ and a vector $\boldsymbol{\epsilon} = (\varepsilon_1, \dots, \varepsilon_n)$ of independent Rademachers. We constructed the *Rademacher process*

$$X_{\boldsymbol{t}} = \langle \boldsymbol{\epsilon}, \boldsymbol{t} \rangle = \sum_{i=1}^n \varepsilon_i t_i \quad \text{for } \boldsymbol{t} \in \mathsf{T}.$$

We developed an upper bound for $\mathrm{Var}[\sup_{\boldsymbol{t} \in \mathsf{T}} X_{\boldsymbol{t}}]$ by an appeal to the Efron–Stein–Steele inequality.

Problem 1(d) contains an example of the type of random processes that we will encounter. There is structure described by a deterministic set T, a source $\boldsymbol{\epsilon}$ of randomness, and a map by which $\boldsymbol{\epsilon}$ interacts with T. We will make a deeper study of supremum problems with this tripartite structure.

In Exercise 4(a), we used convexity of the cumulant generating function to show that $\mathbb{E}\, Z \le \theta^{-1} \xi_Z(\theta)$ holds for any random variable Z and all $\theta > 0$. As a particular example, we considered a random vector $(X_i : i = 1, \dots, n)$ of centered (not necessarily independent!) subgaussian random variables with variance proxies bounded by v. We derived the bound $\mathbb{E} \max_i X_i \le \sqrt{2v \log n}$.

Exercise 4(a) reflects the types of conclusions we might like to reach. We intend to develop bounds on *the expected size of the supremum*, rather than bounds on how sharply it concentrates around its expectation. This exercise hints that the size of index set T can affect the expectation of a supremum.

In this lecture, we show how to quantify the complexity of index sets far more general than $\mathsf{T} = \{1, \dots, n\}$ through the notions of packing and covering in metric spaces. These tools will help us reason about a metric space associated with the random process $(X_t : t \in \mathsf{T})$. Studying the geometry of these metric spaces will, in turn, lead to upper and lower bounds for the suprema of random processes.

9.2 Metric spaces, nets, and separation

Recall that a psuedometric space is a pair $(\mathsf{T}, \mathrm{dist})$, where T is an abstract set. The distance function $\mathrm{dist} : \mathsf{T} \times \mathsf{T} \to \mathbb{R}$ satisfies the axioms

1. $\mathrm{dist}(x, y) \geq 0$ for all x, y in T.
2. $x = y$ implies $\mathrm{dist}(x, y) = 0$.
3. $\mathrm{dist}(x, y) \leq \mathrm{dist}(x, z) + \mathrm{dist}(z, y)$ for all x, y, z in T.

In a metric space, the converse relation also holds: $\mathrm{dist}(x, y) = 0$ implies $x = y$.

For a subset $\mathsf{A} \subseteq \mathsf{T}$, we define the distance from a point to the subset as $\mathrm{dist}(x, \mathsf{A}) := \inf_{y \in \mathsf{A}} \mathrm{dist}(x, y)$. The open and closed ball of radius ε centered at $x \in \mathsf{T}$ are defined, respectively, as

$$\mathsf{B}_\varepsilon(x) := \{y \in \mathsf{T} \,:\, \mathrm{dist}(y, x) < \varepsilon\};$$
$$\overline{\mathsf{B}}_\varepsilon(x) := \{y \in \mathsf{T} \,:\, \mathrm{dist}(y, x) \leq \varepsilon\}.$$

We are interested in subsets of the metric space with special properties. First, we consider a subset with the property that every point in the metric space is close to an element of the subset.

Definition 9.1 (Net). Fix a subset $\mathsf{K} \subseteq \mathsf{T}$ and a parameter $\varepsilon > 0$. A subset $\mathsf{N} \subseteq \mathsf{K}$ is called an ε-net for K if every point in K is within distance ε from N. That is, $\mathrm{dist}(y, \mathsf{N}) \leq \varepsilon$ for all y in K.

An equivalent definition of ε-net is that the union of closed ε-balls with centers in $\mathsf{N} \subseteq \mathsf{K}$ covers K. The figure at right shows a covering of a pentagon by 17 balls of various radii centered at black dots. If we take ε to be the *maximum* of these radii, then the black dots constitute an ε-net for the pentagon.

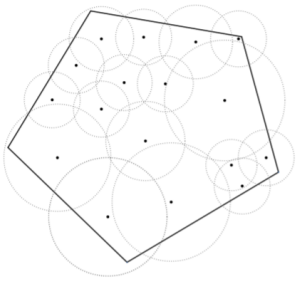


Figure 9.1 (An ε-net)

Next, we consider a complementary notion of a well-separated set of points in the metric space.

Definition 9.2 (Separation). A subset $\mathsf{N} \subseteq \mathsf{T}$ is said to be ε-separated if $\mathrm{dist}(x, y) > \varepsilon$ for all distinct points x, y in N.

The strict inequality in Definition 9.2 is important. For the equilateral triangle at right, the vertices are $(2 - \varepsilon)$-separated for all $\varepsilon > 0$, but they are not 2-separated.

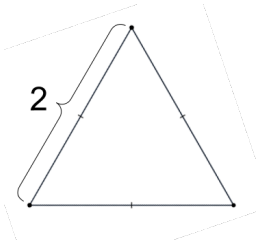


Figure 9.2 (Separation)

In the following lemma we see how the concepts of epsilon nets and epsilon separation are somewhat dual to each other.

Lemma 9.3 (ε-separation versus ε-nets). If N is a maximal ε-separated set in K, then N is an ε-net for K. Here, "maximal" means that adding any additional point to N would result in it no longer being ε-separated.

Proof. Consider an arbitrary point x in K. If x happens to belong to N, then there is nothing to show since $\mathsf{N} \subseteq \mathsf{K}$ tells us $\mathrm{dist}(x, \mathsf{K}) = 0$. If instead x belongs to $\mathsf{K} \setminus \mathsf{N}$, then $\mathsf{N}' = \mathsf{N} \cup \{x\}$ is *not* ε-separated by maximality of N, which implies the existence of some $y \in \mathsf{N}' \subseteq \mathsf{K}$ for which $\mathrm{dist}(x, y) \leq \varepsilon$. This shows that $\mathrm{dist}(x, \mathsf{K}) \leq \varepsilon$ for all x in N, which completes the proof. ∎

9.3 Covering problems

We are interested in understanding the complexity of a metric space. One approach to this problem is to compute how many metric balls we need to cover the set. This idea leads to the following important definition.

Definition 9.4 (Covering number). The covering number of a set $K \subseteq T$ at scale ε is the *minimum* cardinality of ε-net for K:

$$N(K, \varepsilon) := \min\{|N| : N \subseteq K \text{ and } N \text{ is an } \varepsilon\text{-net for } K.\}.$$

If we wish to emphasize the role of the metric, we may also write $N(K, \text{dist}; \varepsilon)$.

It is usually hard to find a covering number exactly; consider again the pentagon example (Figure 9.1). The new covering at right gets by with $|N| = 15$ for balls of the indicated radius. However it is not clear if this covering is optimal: the total space wasted by overlapping circles is greater than the area of an individual circle, and so it may be possible to cover the pentagon at this scale with fewer points.

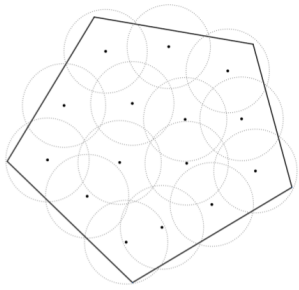


Figure 9.3 (Another ε-net)

Exercise 9.5 (Covering as a function of radius). Verify that $\varepsilon \mapsto N(K, \varepsilon)$ is *decreasing* in the covering radius ε.

Monotonicity does not quite hold for the set argument K, but we can still obtain bounds by way of the following proposition.

Proposition 9.6 If $K \subseteq L \subseteq T$, then $N(K, \varepsilon) \leq N(L, \varepsilon/2)$.

Proof. Let $N(L)$ be a δ-net for L where $|N(L)| = N(L, \delta)$. We have the inclusions

$$K \subseteq L \subseteq \bigcup_{x \in N(L)} \overline{B}_\delta(x).$$

Thus, for each x in $N(L)$, there exists a point y in the intersection of K and $\overline{B}_\delta(x)$. For such a point y, the ball $\overline{B}_{2\delta}(y)$ contains all of $\overline{B}_\delta(x)$. Indeed, this claim follows from the triangle inequality. For any z in $\overline{B}_\delta$, we can travel from y to z by first passing through x. Since both y and z belong to the δ-ball centered at x, we find $\text{dist}(y, z) \leq \text{dist}(y, x) + \text{dist}(x, z) \leq 2\delta$.

Let $N(K)$ denote a set populated by taking one such $y = y(x)$ for each point x from $N(L)$. Then

$$K \subseteq \bigcup_{x \in N(L)} \overline{B}_\delta(x) \subseteq \bigcup_{y \in N(K)} \overline{B}_{2\delta}(y)$$

and so $N(K)$ is a 2δ-net for K with cardinality $|N(L)|$. We obtain the desired result by replacing δ by $\varepsilon/2$ and observing that $N(K, \varepsilon) \leq |N(K)| = |N(L)|$. ∎

The scale $\varepsilon/2$ in Proposition 9.6 cannot be improved without making extra assumptions on the metric space (T, dist). The problem is that points in a δ-net for L may not belong to K. As a concrete example of this phenomenon, let $T = \mathbb{R}$ with the standard metric, $K = \{-1, 1\}$, and $L = \{-1, 0, 1\}$. To cover L at scale $\delta = 1$ we simply place $N(L) = \{0\}$, while covering K requires two elements ($N(K) = K$) whenever the scale parameter is less than two.

9.4 Packing problems

Next, we consider the concept of a *packing*, the maximum number of well-separated points we can insert in a set. We will say that $P \subseteq K$ is an ε-packing for K if P is an ε-separated subset of K.

Figure 9.4 displays a packing of our friend the pentagon. Note that the *centers* of the ε-balls are contained in the pentagon. These ε-balls may overlap, so long as no ball contains the center of another. On the other hand, the $\varepsilon/2$-balls centered at the points of an ε-packing must be pairwise disjoint sets. The latter fact admits a partial converse.

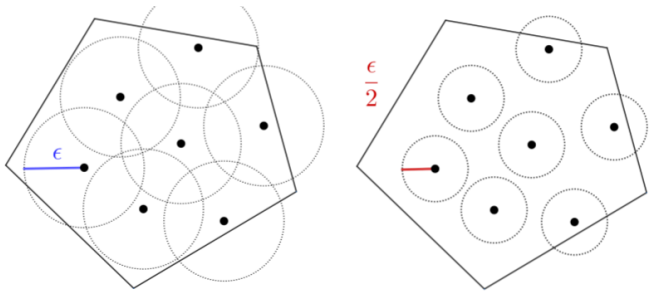


Figure 9.4 (Packings). [left] Balls around the points in a packing may overlap, but none contains the center of another. [right] Shrunken balls around the points of a packing must be disjoint.

Exercise 9.7 (Separated sets and disjointness). Suppose that the subset $\mathsf{P} \subseteq \mathsf{T}$ satisfies $\overline{\mathsf{B}}_{\varepsilon/2}(x) \cap \overline{\mathsf{B}}_{\varepsilon/2}(y) = \emptyset$ for all $x, y \in \mathsf{P}$. Assuming that $(\mathsf{T}, \mathrm{dist})$ is a normed linear space, prove that P is ε-separated.

To see why the exercise cannot be extended to all metric spaces, consider the two-point space $\mathsf{T} = \{x, y\}$ under the discrete metric $\mathrm{dist}(x, y) = 1$, and set the radius $\varepsilon = 3/2$.

We may now consider another measure of the complexity of a metric space, given by the maximum size of a packing.

Definition 9.8 (Packing number). The *packing number* of a set $\mathsf{K} \subseteq \mathsf{T}$ at scale ε is the *maximum* cardinality of an ε-separated subset:

$$\mathrm{P}(\mathsf{K}, \varepsilon) := \max\{|\mathsf{P}| : \mathsf{P} \subseteq \mathsf{K} \text{ and } \mathsf{P} \text{ is } \varepsilon\text{-separated}\}.$$

If we wish to emphasize the role of the metric, we may also write $\mathrm{P}(\mathsf{K}, \mathrm{dist}; \varepsilon)$.

Packing numbers, like covering numbers, are hard to compute. Some of their basic properties are easy enough to establish.

Exercise 9.9 (Packing as a function of radius). Verify that $\varepsilon \mapsto \mathrm{P}(\mathsf{K}, \varepsilon)$ is *decreasing* in the packing radius ε.

Exercise 9.10 (Packing number versus covering number). Prove that $\mathrm{P}(\mathsf{K}, \varepsilon) \geq \mathrm{N}(\mathsf{K}, \varepsilon)$ for all $\varepsilon > 0$.

9.5 Duality between packing and covering

The following proposition establishes that packing and covering are dual problems in a certain sense. This allows us to think in terms of whichever problem seems more tractable in a given situation.

Proposition 9.11 (Packing and covering: Duality). For all $\mathsf{K} \subseteq \mathsf{T}$ and $\varepsilon > 0$, we have

$$\mathrm{P}(\mathsf{K}, 2\varepsilon) \leq \mathrm{N}(\mathsf{K}, \varepsilon) \leq \mathrm{P}(\mathsf{K}, \varepsilon).$$

Proof. The upper bound was addressed at the end of Section 9.4. For the lower bound, let P be a *maximal* 2ε-packing for K, and let N be a *minimal* ε-covering for K. We need to argue that $|\mathsf{P}| \leq |\mathsf{N}|$.

First, observe that

$$\mathsf{P} \subseteq \mathsf{K} \subseteq \bigcup\nolimits_{y \in \mathsf{N}} \overline{\mathsf{B}}_\varepsilon(y).$$

Thus, each point in P is contained in *at least* one metric ball $\overline{\mathsf{B}}_\varepsilon(y)$ for $y \in \mathsf{N}$. Because P is 2ε-separated, no pair of distinct points in P can belong to a common ε-ball centered at a point in N. This observation tells us that each $y \in \mathsf{N}$ is associated to a single point $x \in \mathsf{P}$ by way of the relation $x \in \overline{\mathsf{B}}_\varepsilon(y)$. In other words, we have constructed a surjection from N to P.

We conclude that

$$\mathrm{P}(\mathsf{K}, 2\varepsilon) = |\mathsf{P}| \leq |\mathsf{N}| = \mathrm{N}(\mathsf{K}, \varepsilon).$$

This is the required result. ∎

It is natural to wonder how sensitive Proposition 9.11 is to our definitions for packing and covering. For example, we might instead have worked with an *external covering number* $\mathrm{N}^{\mathrm{ext}}(\mathsf{K}, \varepsilon)$ where the centers of the covering ε-balls can be anywhere in T rather than only K.

Exercise 9.12 (External covering). Let $\mathrm{N}^{\mathrm{ext}}$ be as defined above. Prove that

$$\mathrm{N}^{\mathrm{ext}}(\mathsf{K}, \varepsilon) \leq \mathrm{N}(\mathsf{K}, \varepsilon) \leq \mathrm{N}^{\mathrm{ext}}(\mathsf{K}, \varepsilon/2).$$

9.6 Volumetric bounds

In this section, we present the simplest example of a versatile method for bounding packing and covering numbers. This approach simply compares the "volume" of the set K with the total "volume" of a family of metric balls. We will develop this approach in a normed linear space, but the same arguments are valid in many common metric spaces.

Consider the metric space $(\mathsf{T}, \mathrm{dist})$ where $\mathsf{T} = \mathbb{R}^n$ and the metric $\mathrm{dist}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|$ is induced by an (arbitrary) norm. The closed unit ball in this normed space will be written as

$$\overline{\mathsf{B}} \coloneqq \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| \leq 1\}.$$

We will need some standard notions from the geometry of normed linear spaces. The *dilation* of a subset $\mathsf{X} \subseteq \mathsf{T}$ by a factor $\alpha \in \mathbb{R}$ is defined as

$$\alpha\mathsf{X} \coloneqq \{\alpha\boldsymbol{x} : \boldsymbol{x} \in \mathsf{X}\}.$$

The *Minkowski sum* of two sets $\mathsf{X}, \mathsf{Y} \subseteq \mathsf{T}$ is

$$\mathsf{X} + \mathsf{Y} \coloneqq \{\boldsymbol{x} + \boldsymbol{y} : \boldsymbol{x} \in \mathsf{X}, \boldsymbol{y} \in \mathsf{Y}\}.$$

In particular, we define the set translation $\mathsf{X} + \boldsymbol{y} \coloneqq \mathsf{X} + \{\boldsymbol{y}\}$ for $\boldsymbol{y} \in \mathsf{T}$. The function $\mathrm{Vol}(\mathsf{X})$ computes the Lebesgue measure of a (measurable) subset $\mathsf{X} \subseteq \mathbb{R}^n$.

In a normed linear space, the volumetric argument computes the covering number and packing number by comparing the ordinary volume of the set to the volume of the scaled unit ball.

Proposition 9.13 (Volumetric bounds). For $\mathsf{K} \subseteq \mathbb{R}^n$ and $\varepsilon > 0$, we have

$$\frac{\mathrm{Vol}(\mathsf{K})}{\mathrm{Vol}\left(\varepsilon\overline{\mathsf{B}}\right)} \leq \mathrm{N}(\mathsf{K}, \varepsilon) \leq \mathrm{P}(\mathsf{K}, \varepsilon) \leq \frac{\mathrm{Vol}\left(\mathsf{K} + (\varepsilon/2)\overline{\mathsf{B}}\right)}{\mathrm{Vol}\left((\varepsilon/2)\overline{\mathsf{B}}\right)}.$$

Proof. The middle inequality between was addressed in Proposition (9.11). We need to prove the lower bound on $\mathrm{N}(\mathsf{K}, \varepsilon)$ and the upper bound on $\mathrm{P}(\mathsf{K}, \varepsilon)$.

Consider first the lower bound, and begin by selecting a minimal ε-net N for K with cardinality $|\mathsf{N}| = \mathrm{N}(\mathsf{K}, \varepsilon)$. By the definition of ε-nets we have $\mathsf{K} \subseteq \bigcup_{\boldsymbol{x} \in \mathsf{N}} \overline{\mathsf{B}}_\varepsilon(\boldsymbol{x})$. In a normed linear space, the metric $\mathrm{dist}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|$ is translation invariant and homogeneous. Therefore, we can write the metric ball at a point with a given radius in the form $\overline{\mathsf{B}}_\varepsilon(\boldsymbol{x}) = \boldsymbol{x} + \varepsilon\overline{\mathsf{B}}$. Consequently,

$$\mathsf{K} \subseteq \bigcup\nolimits_{\boldsymbol{x} \in \mathsf{N}} \big(\boldsymbol{x} + \varepsilon\overline{\mathsf{B}}\big).$$

Take the volume of both sides of this inclusion. Using monotonicity and subadditivity of the volume,

$$\mathrm{Vol}(\mathsf{K}) \le \sum\nolimits_{\boldsymbol{x} \in \mathsf{N}} \mathrm{Vol}\,\big(\boldsymbol{x} + \varepsilon\overline{\mathsf{B}}\big) = \sum\nolimits_{\boldsymbol{x} \in \mathsf{N}} \mathrm{Vol}\,\big(\varepsilon\overline{\mathsf{B}}\big) = |\mathsf{N}| \cdot \mathrm{Vol}\,\big(\varepsilon\overline{\mathsf{B}}\big).$$

We have used the fact that the volume is translation invariant. This is the lower bound.

Now we address the upper bound. Let P denote a maximal ε-packing of K with cardinality $|\mathsf{P}| = \mathrm{P}(\mathsf{K}, \varepsilon)$. As an easy consequence of points in P being ε-separated, the $(\varepsilon/2)$-balls centered at distinct points in P must be pairwise disjoint. Combining this fact with additivity of the volume for a disjoint union and the translation-invariance of volume, we arrive at the identity

$$|\mathsf{P}| \cdot \mathrm{Vol}\,\big((\varepsilon/2)\overline{\mathsf{B}}\big) = \mathrm{Vol}\,\Big(\bigcup\nolimits_{\boldsymbol{x} \in \mathsf{P}} \big(\boldsymbol{x} + (\varepsilon/2)\overline{\mathsf{B}}\big)\Big) = \mathrm{Vol}\,\big(\mathsf{P} + (\varepsilon/2)\overline{\mathsf{B}}\big).$$

The second relation is an identity for the Minkowski sum. Since $\mathsf{P} \subseteq \mathsf{K}$, monotonicity of the volume implies that

$$|\mathsf{P}| \cdot \mathrm{Vol}\,\big((\varepsilon/2)\overline{\mathsf{B}}\big) \le \mathrm{Vol}\,\big(\mathsf{K} + (\varepsilon/2)\overline{\mathsf{B}}\big).$$

This is the desired upper bound. ∎

As a particular example, we can consider the problem of covering the unit ball in a normed space with scaled copies of itself.

Corollary 9.14 (Covering the unit ball). For any n and $\varepsilon > 0$, the covering numbers of the n-dimensional unit ball $\overline{\mathsf{B}}$ in the normed linear space $(\mathbb{R}^n, \|\cdot\|)$ admit the bounds

$$\left(\frac{1}{\varepsilon}\right)^n \le \mathrm{N}(\overline{\mathsf{B}}, \varepsilon) \le \left(1 + \frac{2}{\varepsilon}\right)^n.$$

Proof. We apply Proposition 9.13 with $\mathsf{K} = \overline{\mathsf{B}}$. The Lebesgue volume in $\mathbb{R}^n$ is homogeneous of degree n, so

$$\mathrm{Vol}\,\big(\varepsilon\overline{\mathsf{B}}\big) = \varepsilon^n \,\mathrm{Vol}\,\big(\overline{\mathsf{B}}\big).$$

This fact immediately yields the lower bound.

Next, we turn to the upper bound. Since the unit ball $\overline{\mathsf{B}}$ is convex,

$$\overline{\mathsf{B}} + (\varepsilon/2)\overline{\mathsf{B}} = (1 + \varepsilon/2)\overline{\mathsf{B}}.$$

Thus, $\mathrm{Vol}\,\big(\overline{\mathsf{B}} + (\varepsilon/2)\overline{\mathsf{B}}\big) = (1+\varepsilon/2)^n \,\mathrm{Vol}\,\big(\mathsf{B}\big)$. The result follows after a bit of algebra. ∎

The lower bound Corollary 9.14 is exponentially large for small values of ε. Fortunately, we can use exponential concentration inequalities to counteract the growth of the covering numbers.

The astute reader will realize that we have not used the full strength of the assumption that the metric space is a normed linear space. The same kinds of arguments apply in metric measure spaces where the measure of a metric ball does not depend on the location of the center of the ball. See Problem Set 3 (Application 1) for an example.

9.7 The empirical method

In this section, we will develop a probabilistic technique for bounding the covering number of a convex hull in a normed space. For simplicity, we will focus on Euclidean spaces, but the same argument is valid in a wider class of normed spaces. This argument was proposed by Bernard Maurey, but he never published his ideas. They first appeared in a paper of Carl on approximation theory.

Consider a finite-dimensional Euclidean space with the standard metric: $\mathsf{T} = \mathbb{R}^n$ and $\operatorname{dist}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2$. The following proposition is the main result of this section.

Proposition 9.15 (Empirical method). Consider a finite subset $\mathsf{A} = \{\boldsymbol{a}_1, \dots, \boldsymbol{a}_m\}$ of the ℓ_2 *unit ball* in $\mathbb{R}^n$. The covering numbers of the convex hull of the set A satisfy the bound $\mathrm{N}(\operatorname{conv}(\mathsf{A}), \varepsilon) \le m^{\lceil 4/\varepsilon^2 \rceil}$.

Observe that the dimension n of the Euclidean space does not play a role in the bound, and the number m of points in the set A enters polynomially. On the other hand, the ε^{-2} dependence leads to poor results for small ε. For very small ε, the volumetric method is better.

Proof. Fix an arbitrary point $\boldsymbol{a}$ in $\operatorname{conv}(\mathsf{A})$. By definition of the convex hull, we can write

$$\boldsymbol{a} = \sum\nolimits_{i=1}^{m} p_i \boldsymbol{a}_i \quad \text{where } p_i \ge 0 \text{ and } \sum\nolimits_{i=1}^{m} p_i = 1.$$

Using this probability vector $\boldsymbol{p}$, we can construct a random vector $\boldsymbol{x} \in \mathbb{R}^n$ with the distribution

$$\boldsymbol{x} = \boldsymbol{a}_i \quad \text{with probability } p_i.$$

It is clear that $\mathbb{E}\,\boldsymbol{x} = \boldsymbol{a}$. Draw an independent family $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_r)$ of r copies of the random vector $\boldsymbol{x}$, and form the empirical average

$$\boldsymbol{y} = \frac{1}{r} \sum\nolimits_{i=1}^{r} \boldsymbol{x}_i.$$

By linearity of expectation, $\mathbb{E}\,\boldsymbol{y} = \boldsymbol{a}$. We will obtain a bound on the number r of summands we need for a realization of $\boldsymbol{y}$ to approximate the point $\boldsymbol{a}$ with ℓ_2 error ε.

For a particular number $r = r(\varepsilon)$, suppose that

$$\mathbb{E}\,\|\boldsymbol{a} - \boldsymbol{y}\|_2^2 \le \varepsilon^2.$$

Then the probabilistic method implies the existence of an indexing scheme $\iota : \{1, \dots, r\} \to \{1, \dots, m\}$ for which $\boldsymbol{y}_\star = r^{-1} \sum_{j=1}^{r} \boldsymbol{a}_{\iota(j)}$ satisfies

$$\|\boldsymbol{a} - \boldsymbol{y}_\star\|_2^2 \le \varepsilon^2.$$

Observe that there are m^r such indexing schemes. Therefore, the vector $\boldsymbol{y}_\star$ takes at most m^r different values, one of which is an ε-approximation to the distinguished point $\boldsymbol{a}$. But $\boldsymbol{a}$ is an arbitrary element of the convex hull, so the m^r choices of $\boldsymbol{y}_\star$ compose an ε-net for $\operatorname{conv}(\mathsf{A})$.

It remains to determine how we large we must take the parameter $r = r(\varepsilon)$. Since $\mathbb{E}\,\boldsymbol{x}_i = \boldsymbol{a}$, we can begin by expressing

$$\mathbb{E}\,\|\boldsymbol{a} - \boldsymbol{y}\|_2^2 = r^{-2}\,\mathbb{E}\,\|\textstyle\sum_{i=1}^{r}(\boldsymbol{a} - \boldsymbol{x}_i)\|_2^2$$

Next, observe that $(\boldsymbol{a} - \boldsymbol{x}_i : i = 1, \dots, r)$ are independent, centered random vectors. By orthogonality,

$$\mathbb{E}\,\|\boldsymbol{a} - \boldsymbol{y}\|_2^2 = r^{-2} \textstyle\sum_{i=1}^{r} \mathbb{E}\,\|\boldsymbol{a} - \boldsymbol{x}_i\|_2^2.$$

Since the points $\boldsymbol{a}$ and $\boldsymbol{x}_i$ belong to $\operatorname{conv}(\mathsf{A})$ and A is contained in the ℓ_2 unit ball, we can bound $\|\boldsymbol{a} - \boldsymbol{x}_i\|_2^2 \leq 4$. It follows that

$$\mathbb{E}\,\|\boldsymbol{a} - \boldsymbol{y}\|_2^2 \leq \frac{4}{r}.$$

To achieve an error below ε^2, it suffices that r is an integer greater than $4/\varepsilon^2$. ∎

Figure 9.5 shows the ε-nets that would result from applying the proof's construction to the pentagon. From the illustration we see a qualitative cause for slow convergence of this construction as $\varepsilon \to 0$: for larger values of r they place many points near the pentagon's center that would be better used near the edges.

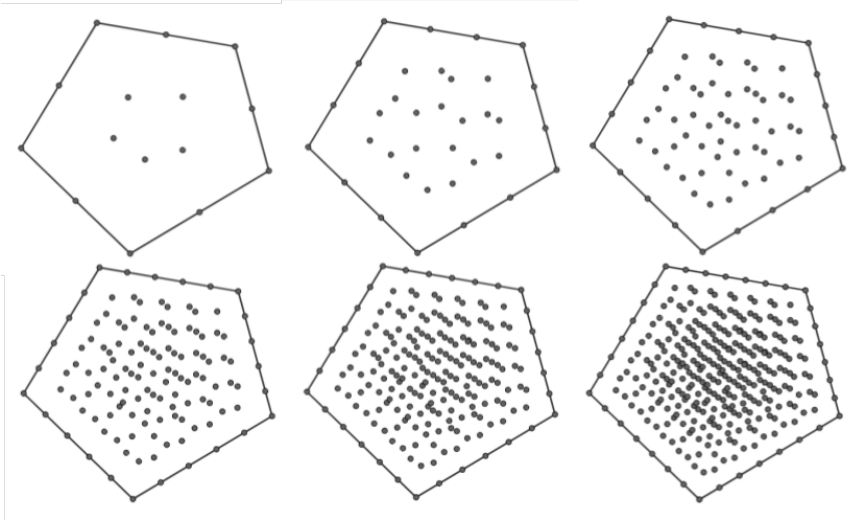


Figure 9.5 (Empirical nets). Empirical nets for a pentagon ($m = 5$ and $r = 2, \ldots, 7$).

Exercise 9.16 (Improved empirical method). The ε-net in the bottom right of Figure 9.5 cannot possibly be using $m^r = 5^7 = 78,125$ distinct points. Find a stronger upper bound on the number of unique sample averages $r^{-1}\sum_{j=1}^{r} \boldsymbol{a}_{\iota(j)}$ where $\iota : \{1, \ldots, r\} \to \{1, \ldots, m\}$.

Example 9.17 (Covering the probability simplex). Consider the case where $\mathsf{A} = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ consists of the standard basis vectors of $\mathbb{R}^n$, so that $\Delta_n := \operatorname{conv}(\mathsf{A})$ is the probability simplex. By Proposition 9.15, there exists an ε-net for Δ_n of size $n^{\lceil 4/\varepsilon^2 \rceil}$. ▪

Observe that the bound in the example can be written as

$$\log \mathrm{N}(\operatorname{conv}\mathsf{A}, \varepsilon) \leq \lceil 4/\varepsilon^2 \rceil \log n.$$

This estimate bears a similarity to the inequality $\mathbb{E}\max_{i=1,\ldots,n} X_i \leq \sqrt{2v \log n}$ for the maximum of a family $(X_i : i = 1, \ldots, n)$ of n centered v-subgaussian random variables. Later, we will appreciate that this similarity is no coincidence.

10. Gaussian Comparison Theorems

Date: 4 February 2021 **Scribe:** Taylan Kargin

Agenda:
1. Random processes
2. Gaussian processes
3. Slepian's lemma
4. Kahane's theorem
5. Gaussian interpolation
6. Proof of Kahane's theorem

In this lecture, we begin our discussion of random processes in earnest. We focus on a special class of random processes, called Gaussian processes (GPs). Gaussian processes are especially important for our discussion since they have additional properties that makes them easy to understand and work with.

10.1 Random processes and metric spaces

We start with some general definitions. In the next section, we will tighten our discussion to the class of Gaussian processes.

10.1.1 Random processes

Let us refresh our memory of the definition of a random process is.

Definition 10.1 (Random process). Let T be an abstract set. A *random process* on T is a family $(X_t : t \in \mathsf{T})$ of random variables. Generally, these random variables do *not* compose an independent family.

Historically, the symbol T stood for time, and it alludes to a setting where each random variable X_t gives the value of a process at a given time instant, $t \in \mathbb{R}$. We are explicitly interested in more general models. For example, when $\mathsf{T} \subseteq \mathbb{R}^n$, the process is often called a *random field*, and it might describe the spatial variation of a quantity such a pressure, temperature, etc.

Unless stated explicitly otherwise, all random processes in this course take real values. Some of the methods extend to more general random variables.

10.1.2 Increments

We need a way to capture the variation of a random process over the index set. The following basic definitions provide a useful way to think about this problem.

- We say that a random processes is *centered* if $\mathbb{E}\, X_t = 0$ for all $t \in \mathsf{T}$.
- The *covariance function* of a *centered* random process is

$$\Sigma(s,t) := \mathbb{E}[X_s X_t] \quad \text{for all} \quad s,t \in \mathsf{T}.$$

  This function packs up the (pairwise) correlations among the elements of the process.
- The *increments* of a random process are

$$d(s,t) := \|X_s - X_t\|_{\mathsf{L}_2} = \left(\mathbb{E}\,|X_s - X_t|^2\right)^{1/2} \quad \text{for all} \quad s,t \in \mathsf{T}$$

  The increments describe how much the random process changes from point $s \in \mathsf{T}$ to point $t \in \mathsf{T}$. For a centered process, we can interpret $d(s,t)^2$ as the variance of the difference $X_s - X_t$.

- The pair (T, d) is always a pseudo-metric space. Recall that in a pseudo-metric space, it is possible to have $d(s,t) = 0$ for $s \neq t$.

Exercise 10.2 (Covariance and increments). For a centered random process, explain how to compute the covariance function Σ from the increments and vice versa. **Hint:** Polarization.

10.1.3 Suprema

In this section of the course, we will be studying the supremum of a real-valued random process:

$$Z = \sup_{t \in \mathsf{T}} X_t.$$

Remarkably, the behavior of the supremum is intimately related to the geometry of the pseudo-metric space (T, d). This insight is captured by the following principle, which dates to work of Kolmogorov on the continuity of random processes.

> If the elements of a random process vary in a "smooth" way with the index, then the supremum of the process is controlled by the "complexity" of the index set.

Our task is to develop appropriate ways to quantify the complexity of a metric space. We have already seen one relevant notion, namely the covering numbers $\varepsilon \mapsto \mathrm{N}(\mathsf{T}, d; \varepsilon)$. Even for Gaussian processes, which are the simplest case, we will need more refined approaches to fully capture the behavior of the supremum.

10.1.4 Measurability

In order to properly define the supremum, we have to be careful about measurability issues because the index set can be an uncountable set. Indeed, we cannot take the supremum of an uncountable family of random variables without (possibly) breaking measurability. We will dispatch with these technical issues in this paragraph and then forget about them.

One way to overcome the concern about measurability is to work with the *lattice supremum*. We may *define* the expectation of the supremum as

$$\mathbb{E} \sup_{t \in \mathsf{T}} X_t := \sup \left\{ \mathbb{E} \max_{t \in \mathsf{T}'} X_t : \text{finite } \mathsf{T}' \subseteq \mathsf{T} \right\}.$$

Similar definitions can be given for the probability that the supremum takes particular values, as well as other types of integrals.

Another way to address this issue is to restrict our attention to separable random processes. A random process $(X_t : t \in \mathsf{T})$ is *separable* if there is a countable subset $\mathsf{T}_0 \subseteq \mathsf{T}$ where

$$X_t \in \lim_{\substack{s \to t \\ s \in \mathsf{T}_0}} X_s \quad \text{for all } t \in \mathsf{T}.$$

That is, we can realize the value of each element of the random process as a sequential limit of elements of the countable process $(X_t : t \in \mathsf{T}_0)$. Most random processes that you encounter are indeed separable. For a separable process, the supremum is indeed measurable, and it agrees with the lattice supremum.

10.2 Gaussian processes

We now move our discussion to a more specific class of random processes, namely Gaussian processes (GPs). GPs are important for applications in machine learning, statistics, numerical analysis, and other fields.

Definition 10.3 (Gaussian process). A real-valued random process $(X_t : t \in \mathsf{T})$ is called a *Gaussian process* if the family $(X_t : t \in \mathsf{T}')$ is jointly Gaussian for every finite subset $\mathsf{T}' \subseteq \mathsf{T}$.

Equivalently, the random process is Gaussian if the random variable $\sum_{t\in\mathsf{T}'} a_t X_t$ follows a normal distribution for each finite subset $\mathsf{T}' \subseteq \mathsf{T}$ and all coefficients $a_t \in \mathbb{R}$ with $t \in \mathsf{T}'$.

Exercise 10.4 (GPs: Equivalence). Explain why the two definitions of a Gaussian process are equivalent.

Exercise 10.5 (GPs: Description). Show that a GP is completely determined by its mean function $t \mapsto \mathbb{E}\, X_t$ and its covariance function $(s,t) \mapsto \Sigma(s,t)$. Show how to compute the covariance function from the increments and vice-versa. **Hint:** Polarization.

Warning 10.6 (Joint Gaussianity). It is possible to construct a family of random variables that are individually Gaussian but whose joint distribution is not Gaussian. According to our definition, these families are *not* Gaussian processes. ▪

10.2.1 Examples of GPs

As a first example, we can give a complete description of a GP comprising a finite number of random variables.

Example 10.7 (Finite GP). Consider a centered GP $(X_t : t \in \mathsf{T})$, where T is a *finite* set with cardinality $|\mathsf{T}| = N$. We can model this GP as an N-dimensional random vector:

$$\boldsymbol{x} := (X_1, \dots, X_N) = \boldsymbol{\Sigma}^{1/2}\boldsymbol{g}, \quad \text{where } \boldsymbol{g} \sim \textsc{normal}(\mathbf{0}, \mathbf{I}_N). \tag{10.1}$$

The psd matrix $(i,j) \mapsto \boldsymbol{\Sigma}(i,j)$ tabulates the covariance function of the GP. ▪

Next, let us give another explicit construction that yields a wide class of Gaussian processes. These processes are finite-dimensional, but they can contain an uncountable number of random variables.

Example 10.8 (Canonical GP). Consider a set $\mathsf{T} \subseteq \mathbb{R}^n$, and draw a standard normal vector $\boldsymbol{g} \sim \textsc{normal}(\mathbf{0}, \mathbf{I}_n)$. Form the real random variables

$$X_{\boldsymbol{t}} = \langle \boldsymbol{g},\ \boldsymbol{t} \rangle_{\ell_2} \quad \text{for } \boldsymbol{t} \in \mathsf{T}.$$

The family $(X_{\boldsymbol{t}} : \boldsymbol{t} \in \mathsf{T})$ is a centered GP, called a *canonical Gaussian process*. Let us emphasize that there is only one random vector $\boldsymbol{g}$ that goes into building the entire process. All the other random variables are derived from this unique source of randomness.

Let us compute the increments of the canonical GP defined on the set $\mathsf{T} \subseteq \mathbb{R}^n$. By a simple calculation,

$$d(X_{\boldsymbol{s}}, X_{\boldsymbol{t}}) = \|X_{\boldsymbol{s}} - X_{\boldsymbol{t}}\|_{\mathrm{L}_2} = \|\boldsymbol{s} - \boldsymbol{t}\|_{\ell_2}.$$

The last norm is simply the ℓ_2 norm on the space $\mathbb{R}^n$. In other words, the metric space (T, d) associated with a canonical Gaussian process is isomorphic to the metric subspace (T, ℓ_2) of the Euclidean space $(\mathbb{R}^n, \ell_2)$. As such, canonical GPs are particularly easy to visualize. ▪

Exercise 10.9 (Finite GPs). Compute the covariance function of a canonical GP on a finite-cardinality subset $\mathsf{T} \subseteq \mathbb{R}^n$.

10.3 Slepian's lemma and Kahane's theorem

As compared with general random processes, Gaussian processes are easier to analyze and have many additional properties. In this lecture, we will establish a comparison principle that allows us to relate the suprema of two Gaussian processes. This result is attributed to Slepian [Sle62].

Theorem 10.10 (Slepian). Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$ be finite centered GPs with covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{y}}$. Assume that

$$\begin{aligned} (\boldsymbol{\Sigma}_{\boldsymbol{x}})_{ij} = \mathbb{E}[X_i X_j] \le \mathbb{E}[Y_i Y_j] = (\boldsymbol{\Sigma}_{\boldsymbol{y}})_{ij} \qquad & \text{for all } i, j \in \{1, \dots, N\}; \\ (\boldsymbol{\Sigma}_{\boldsymbol{x}})_{ii} = \mathbb{E}[X_i^2] = \mathbb{E}[Y_i^2] = (\boldsymbol{\Sigma}_{\boldsymbol{y}})_{ii}, \qquad & \text{for all } i \in \{1, \dots, N\}. \end{aligned}$$

Then

$$\mathbb{P}\{\max_i Y_i > u\} \le \mathbb{P}\{\max_i X_i > u\} \quad \text{for all } u \in \mathbb{R}.$$

In particular,

$$\mathbb{E} \max_i Y_i \le \mathbb{E} \max_i X_i.$$

Aside: Theorem 10.10 is usually referred to as "Slepian's lemma," even though it is a major result in its own right.

To understand Theorem 10.10, notice that its hypotheses imply that the increments of the two GPs satisfy the relations

$$d_Y(i, j) = \|Y_i - Y_j\|_{L_2} \le \|X_i - X_j\|_{L_2} = d_X(i, j) \quad \text{for all } i, j. \tag{10.2}$$

In words, we consider two GPs have the same coordinate-wise variances, while the process $\boldsymbol{x}$ has bigger increments than the process $\boldsymbol{y}$. That is, the random variables in the process $\boldsymbol{x}$ are farther apart than the corresponding random variables in the process $\boldsymbol{y}$. In this case, Theorem 10.10 tells us that the maximum of the process $\boldsymbol{x}$ stochastically dominates the maximum of the process $\boldsymbol{y}$.

Exercise 10.11 (Minimum). Under the assumptions of Theorem 10.10, show that

$$\mathbb{E} \min_i Y_i \ge \mathbb{E} \min_i X_i.$$

This section contains an overview of the proof of Slepian's lemma. Some applications of Slepian's lemma appear in the next lecture.

10.3.1 Kahane's theorem

The main ingredient in the proof of Theorem 10.10 is an abstract theorem due to Kahane [Kah86]. In this section, we introduce the statement of Kahane's theorem and derive Slepian's lemma as a direct consequence. The proof of Kahane's theorem is given in Section 10.5.

Theorem 10.12 (Kahane). Assume that $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$ are finite centered GPs. Suppose that there is a pair of index sets $\mathsf{A}, \mathsf{B} \subseteq \{1, \dots, N\}^2$ for which

$$\begin{aligned} \mathbb{E}[X_i X_j] \le \mathbb{E}[Y_i Y_j] \qquad & \text{for all } (i, j) \in \mathsf{A}; \\ \mathbb{E}[X_i X_j] \ge \mathbb{E}[Y_i Y_j] \qquad & \text{for all } (i, j) \in \mathsf{B}; \\ \mathbb{E}[X_i X_j] = \mathbb{E}[Y_i Y_j] \qquad & \text{for all } (i, j) \notin \mathsf{A} \cup \mathsf{B}. \end{aligned}$$

Aside: Kahane's theorem isolates the abstract fact that underlies Slepian's lemma and related results. Slepian's original proof was based on similar principles, but it is more direct.

Let $f : \mathbb{R}^N \to \mathbb{R}$ be a function whose second (distributional) derivative satisfies

$$\partial_{ij} f \geq 0 \qquad \text{for all } (i,j) \in \mathsf{A};$$
$$\partial_{ij} f \leq 0 \qquad \text{for all } (i,j) \in \mathsf{B}.$$

Then

$$\mathbb{E}\, f(\boldsymbol{x}) \leq \mathbb{E}\, f(\boldsymbol{y}).$$

In the next section, we will use this result to derive Slepian's lemma. This argument helps us appreciate that Kahane's strange set of hypotheses is indeed useful. Turn to Section 10.5 for the proof of Theorem 10.12, where the source of the hypotheses will become clear.

Kahane's theorem has other applications beyond the proof of Slepian's lemma. In particular, as you will see on Problem Set 3, Kahane's result implies Gordon's minimax theorem [Gor88], which has significant implications for contemporary signal processing [TOH14].

10.3.2 Proof of Slepian from Kahane

Let us establish Theorem 10.10 as a consequence of Theorem 10.12. Introduce the sets $\mathsf{A} = \{(i,j) : i \neq j\}$ and $\mathsf{B} = \emptyset$. For this specific assignment, the hypotheses of Slepian's lemma align with the hypotheses of Kahane's theorem.

Next, we must choose an appropriate function $f : \mathbb{R}^N \to \mathbb{R}$. Fix a point $u \in \mathbb{R}$, and define

$$f(\boldsymbol{w}) := \mathbb{1}\{\max_i w_i \leq u\} = \prod\nolimits_{i=1}^{N} \mathbb{1}\{w_i \leq u\} \quad \text{for } \boldsymbol{w} \in \mathbb{R}^N.$$

Let us compute the second distributional derivative $\partial_{ij} f$ for a pair $(i,j) \in \mathsf{A}$. Write δ_u for the Dirac distribution at u with mass one. Then

$$(\partial_i f)(\boldsymbol{w}) = -\delta_u(w_i) \prod\nolimits_{k \neq i} \mathbb{1}\{w_k \leq u\};$$
$$(\partial_{ij} f)(\boldsymbol{w}) = \delta_u(w_i)\, \delta_u(w_j) \prod\nolimits_{k \neq i,j} \mathbb{1}\{w_k \leq u\} \geq 0.$$

See Figure 10.1 for an illustration.

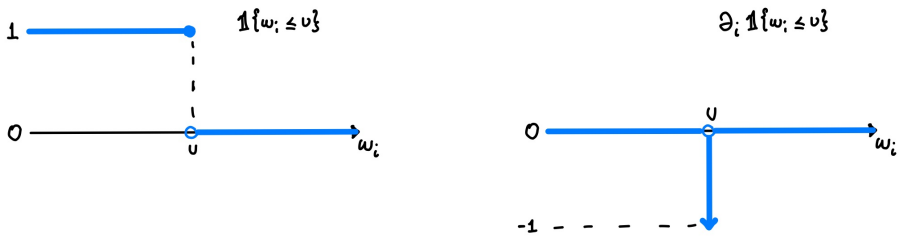


Figure 10.1 (Derivative of an indicator). An indicator and its distributional derivative.

We have now verified the assumptions of Kahane's theorem. Indeed, positivity of the second derivative $\partial_{ij} f$ for indices $(i,j) \in \mathsf{A}$ is the only thing that we need to check because B is empty. Kahane's theorem delivers

$$\mathbb{P}\{\max_i X_i \leq u\} = \mathbb{E}\, f(\boldsymbol{x})$$
$$\leq \mathbb{E}\, f(\boldsymbol{y}) = \mathbb{P}\{\max_i Y_i \leq u\}.$$

Taking the complements, we finally get

$$\mathbb{P}\{\max_i X_i > u\} \geq \mathbb{P}\{\max_i Y_i > u\}.$$

This is the conclusion of Slepian's lemma. ∎

Warning 10.13 (Distributional derivatives). To make this argument rigorous, we need to use the theory of distributions. We can also develop a more elementary, but messier, proof by smoothing the indicator function so it is differentiable. ▪

10.4 Gaussian interpolation

Before we can establish Kahane's theorem, we need to present some important facts about Gaussian integration by parts (IBP) and the interpolation of Gaussian random vectors.

10.4.1 Integration by parts

Although totally elementary, Gaussian integration by parts formulas play a fundamental role in Gaussian analysis.

Fact 10.14 (Gaussian IBP: Univariate case). Let $\gamma \sim \text{NORMAL}(0,1)$ and $f : \mathbb{R} \to \mathbb{R}$. Then

$$\mathbb{E}[\gamma f(\gamma)] = \mathbb{E}[f'(\gamma)].$$

Here, f' denotes the (distributional) derivative. ▪

Exercise 10.15 (Gaussian IBP). Prove Fact 10.14 by writing the expectation as an integral and invoking integration by parts. The formula is valid whenever we can make sense of both sides.

We will need a more general version of this result, which is an easy corollary.

Fact 10.16 (Multivariate Gaussian IBP). Let $\boldsymbol{x} \in \mathbb{R}^N$ be a finite centered GP with covariance matrix $\boldsymbol{\Sigma}$. For $f : \mathbb{R}^N \to \mathbb{R}$,

$$\mathbb{E}[X_i f(\boldsymbol{x})] = \sum\nolimits_{j=1}^{N} (\boldsymbol{\Sigma})_{ij}\, \mathbb{E}[(\partial_j f)(\boldsymbol{x})].$$

Here, $\partial_j f$ is the (distributional) derivative with respect to the jth coordinate. ▪

Proof sketch. Using the canonical model (10.1) for finite GPs, we may write $\boldsymbol{x} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{g}$ with $\boldsymbol{g} \sim \text{NORMAL}(\mathbf{0}, \mathbf{I}_N)$. Then

$$\begin{aligned}\mathbb{E}[X_i f(\boldsymbol{x})] &= \sum\nolimits_{k=1}^{N} (\boldsymbol{\Sigma}^{1/2})_{ik}\, \mathbb{E}[g_k f(\boldsymbol{\Sigma}^{1/2}\boldsymbol{g})] \\ &= \sum\nolimits_{k=1}^{N} (\boldsymbol{\Sigma}^{1/2})_{ik}\, \mathbb{E}[g_k h(\boldsymbol{g})],\end{aligned}$$

where $h(\boldsymbol{g}) = f(\boldsymbol{\Sigma}^{1/2}\boldsymbol{g})$. Apply the univariate Gaussian IBP to each component g_k of $\boldsymbol{g}$, and do the algebra to arrive at the desired result. ∎

10.4.2 Interpolation

The main idea in the proof of Kahane's theorem is a technique called Gaussian interpolation. We need to find a way of bridging the two GPs $\boldsymbol{x}$ and $\boldsymbol{y}$ that we are trying to compare in Kahane's theorem. The idea is to construct a continuous path from $\boldsymbol{y}$ to $\boldsymbol{x}$ along which we can control the behavior of functions of the random processes.

Proposition 10.17 (Gaussian interpolation). Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$ be finite centered GPs with covariance matrices $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_y$, and assume that $\boldsymbol{x}, \boldsymbol{y}$ are statistically independent from each other. Define

$$\boldsymbol{z}(\tau) := \sqrt{\tau}\,\boldsymbol{x} + \sqrt{1-\tau}\,\boldsymbol{y} \quad \text{for all } \tau \in [0,1]. \tag{10.3}$$

Then, for $f : \mathbb{R}^N \to \mathbb{R}$, we have that

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\,\mathbb{E}\,f(\boldsymbol{z}(\tau)) = \frac{1}{2}\sum\nolimits_{i,j=1}^{N} \big[(\boldsymbol{\Sigma}_x)_{ij} - (\boldsymbol{\Sigma}_y)_{ij}\big]\,\mathbb{E}[(\partial_{ij} f)(\boldsymbol{z}(\tau))].$$

To appreciate why we construct the interpolating process $\boldsymbol{z}(\tau)$ in this manner, observe that

$$\operatorname{Cov}(\boldsymbol{z}(\tau)) = \tau\,\boldsymbol{\Sigma}_x + (1-\tau)\,\boldsymbol{\Sigma}_y, \tag{10.4}$$

for $\tau \in [0,1]$. In other words, we have constructed a family $\boldsymbol{z}(\tau)$ of Gaussian vectors whose covariance matrices interpolate *linearly* between the covariances of $\boldsymbol{y}$ and $\boldsymbol{x}$.

Proof. The argument follows from a short calculation. Invoking the chain rule and the definition (10.3) of the interpolating process $\boldsymbol{z}(\tau)$,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}\,\mathbb{E}\,f(\boldsymbol{z}(\tau)) &= \sum\nolimits_{i=1}^{N} \mathbb{E}\left[(\partial_i f)(\boldsymbol{z}(\tau))\,\frac{\mathrm{d}}{\mathrm{d}\tau} z_i(\tau)\right] \\ &= \frac{1}{2}\sum\nolimits_{i=1}^{N} \mathbb{E}\left[(\partial_i f)(\boldsymbol{z}(\tau))\left(\frac{1}{\sqrt{\tau}}X_i - \frac{1}{\sqrt{1-\tau}}Y_i\right)\right].\end{aligned} \tag{10.5}$$

Applying Gaussian IBP (Fact 10.16) to each term, we obtain

$$\begin{aligned}\frac{1}{\sqrt{\tau}}\,\mathbb{E}[X_i\,(\partial_i f)(\boldsymbol{z}(\tau))] &= \sum\nolimits_{j=1}^{N} (\boldsymbol{\Sigma}_x)_{ij}\,\mathbb{E}[(\partial_{ij} f)(\boldsymbol{z}(\tau))]; \\ \frac{1}{\sqrt{1-\tau}}\,\mathbb{E}[Y_i\,(\partial_i f)(\boldsymbol{z}(\tau))] &= \sum\nolimits_{j=1}^{N} (\boldsymbol{\Sigma}_y)_{ij}\,\mathbb{E}[(\partial_{ij} f)(\boldsymbol{z}(\tau))].\end{aligned}$$

Combining the last display with (10.5), we get the desired result. ∎

10.4.3 Example: The Fernique–Sudakov comparison

Gaussian interpolation is a widely used and powerful tool. As a first illustration, we sketch how this method allows us to prove a comparison due to Fernique [Fer75] and Sudakov. This result is a refinement of Slepian's lemma that only involves the increments of the Gaussian processes.

Theorem 10.18 (Fernique–Sudakov). Consider finite centered GPs that satisfy

$$\|Y_i - Y_j\|_{\mathrm{L}_2} \le \|X_i - X_j\|_{\mathrm{L}_2}, \quad \text{for all } i, j.$$

Then

$$\mathbb{E}\max\nolimits_i Y_i \le \mathbb{E}\max\nolimits_i X_i.$$

In this result, we do *not assume* equal variances! We recover bounds for the expected maximum—but not tail probabilities.

Proof sketch. The proof is based on the fact that the maximum of a family of numbers $(w_i : i = 1, \dots, N)$ can be approximated by the soft-max:

$$\max\nolimits_i w_i \approx \frac{1}{\theta}\log\sum\nolimits_{i=1}^{N} \exp(\theta w_i) \quad \text{for } \theta > 0.$$

Apply Gaussian interpolation (Proposition 10.17) to the soft-max and optimize over θ to obtain the stated result. ∎

10.5 Proof of Kahane's theorem

Finally, we are prepared to establish Kahane's theorem. Without loss of generality, we may assume that the two finite centered GPs $\boldsymbol{x}$ and $\boldsymbol{y}$ are statistically independent. We use Gaussian interpolation to construct a path between the two GPs. Define

$$\boldsymbol{z}(\tau) = \sqrt{\tau}\,\boldsymbol{x} + \sqrt{1-\tau}\,\boldsymbol{y} \quad \text{for } \tau \in [0,1].$$

By Proposition 10.17,

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\,\mathbb{E}[f(\boldsymbol{z}(\tau))] = \frac{1}{2}\sum\nolimits_{i,j=1}^{N}\big[(\boldsymbol{\Sigma}_x)_{ij} - (\boldsymbol{\Sigma}_y)_{ij}\big]\,\mathbb{E}[(\partial_{ij}f)(\boldsymbol{z}(\tau))].$$

The hypothesis of Kahane's theorem 10.12 allow us to control each of the terms in the sum:

$$\begin{aligned}
(i,j) \in \mathsf{A}: &\qquad (\boldsymbol{\Sigma}_x)_{ij} \le (\boldsymbol{\Sigma}_y)_{ij} \text{ and } \partial_{ij}f \ge 0;\\
(i,j) \in \mathsf{B}: &\qquad (\boldsymbol{\Sigma}_x)_{ij} \ge (\boldsymbol{\Sigma}_y)_{ij} \text{ and } \partial_{ij}f \le 0;\\
(i,j) \notin \mathsf{A}\cup\mathsf{B}: &\qquad (\boldsymbol{\Sigma}_x)_{ij} = (\boldsymbol{\Sigma}_y)_{ij}.
\end{aligned}$$

We immediately deduce that

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\,\mathbb{E}[f(\boldsymbol{z}(\tau))] \le 0 \quad \text{for } \tau \in [0,1].$$

As a consequence,

$$\mathbb{E}\,f(\boldsymbol{x}) - \mathbb{E}\,f(\boldsymbol{y}) = \int_0^1 \mathrm{d}\tau \frac{\mathrm{d}}{\mathrm{d}\tau}\,\mathbb{E}\,f(\boldsymbol{z}(\tau)) \le 0.$$

This is the conclusion of Kahane's theorem. ∎

Lecture bibliography

[Fer75] X. Fernique. "Regularité des trajectoires des fonctions aléatoires gaussiennes". In: *Ecole d'Eté de Probabilités de Saint-Flour IV—1974*. Springer Berlin Heidelberg, 1975, pages 1–96.

[Gor88] Y. Gordon. "On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$". In: *Geometric Aspects of Functional Analysis*. Springer Berlin Heidelberg, 1988, pages 84–106.

[Kah86] J.-P. Kahane. "Une inegalité du type de Slepian et Gordon sur les processus gaussiens". In: *Israel Journal of Mathematics* 55.1 (Feb. 1986), pages 109–110. DOI: 10.1007/BF02772698.

[Sle62] D. Slepian. "The one-sided barrier problem for Gaussian noise". In: *The Bell System Technical Journal* 41.2 (1962), pages 463–501. DOI: 10.1002/j.1538-7305.1962.tb02419.x.

[TOH14] C. Thrampoulidis, S. Oymak, and B. Hassibi. "A Tight Version of the Gaussian min-max theorem in the Presence of Convexity". In: *CoRR* abs/1408.4837 (2014). URL: http://arxiv.org/abs/1408.4837. arXiv: 1408.4837.

11. Chevet and Sudakov

Date: 9 February 2021 **Scribe:** Ethan Epperly

In the last lecture, we discussed the Gaussian comparison results of Slepian and Fernique–Sudakov. Today, we shall discuss two applications of these result. The first is *Chevet's theorem*, which gives bounds on the expected supremum of a bilinear form defined by a Gaussian matrix. The second result is *Sudakov minoration*, which gives *lower bounds* on the supremum of a Gaussian process. Upper bounds for these suprema will be the subject of the next lecture.

Agenda:
1. Chevet's Theorem
2. Norm of a Gaussian matrix
3. Sudakov's minoration
4. Covering number bounds

11.1 Chevet's theorem

We begin with a result that describes the supremum of a bilinear form in a Gaussian matrix. This theorem was established by Simone Chevet in 1977 [Che77], with subsequent refinements by Yehoram Gordon [Gor85].

11.1.1 The Gaussian width

Before we begin, it is valuable to introduce special notation for the expected supremum of a canonical Gaussian process.

Definition 11.1 (Gaussian width). Let $\mathsf{T} \subset \mathbb{R}^n$ be a compact set. The *Gaussian width* of T is the quantity

$$w(\mathsf{T}) := \mathbb{E} \sup_{\boldsymbol{x} \in \mathsf{T}} \langle \boldsymbol{g}, \boldsymbol{x} \rangle \quad \text{where } \boldsymbol{g} \sim \textsc{normal}(\mathbf{0}, \mathbf{I}_n).$$

The Gaussian width can be viewed as a measure of the "size" of the set T. It has many beautiful properties, some of which are collected in the next exercise.

Exercise 11.2 (Gaussian width). Let $\mathsf{T} \subset \mathbb{R}^n$ be a compact set. Prove that the Gaussian width has the following properties.

1. **Rigid motion invariance.** $w(\boldsymbol{Q}\mathsf{T} + \boldsymbol{a}) = w(\mathsf{T})$ for each orthogonal matrix $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ and for each point $\boldsymbol{a} \in \mathbb{R}^n$.
2. **Bounds.** $0 \le w(\mathsf{T}) < \sqrt{n} \cdot \operatorname{rad}(\mathsf{T})$.
3. **Monotonicity.** If $\mathsf{S} \subseteq \mathsf{T}$, then $w(\mathsf{S}) \le w(\mathsf{T})$.
4. **Homogeneity.** $w(\alpha \mathsf{T}) = |\alpha| \, w(\mathsf{T})$ for $\alpha \in \mathbb{R}$.
5. **Convexity.** $w(\mathsf{T}) = w(\operatorname{conv}(\mathsf{T}))$.
6. ***Valuation.** If $\mathsf{S}, \mathsf{T}, \mathsf{S} \cup \mathsf{T}$ are convex, then
$$w(\mathsf{S}) + w(\mathsf{T}) = w(\mathsf{S} \cap \mathsf{T}) + w(\mathsf{S} \cup \mathsf{T}).$$
7. ***Continuity.** If $\mathsf{S}_n \to \mathsf{T}$ in Hausdorff metric, then $w(\mathsf{S}_n) \to w(\mathsf{T})$.
8. *****Uniqueness.** Up to scaling, the Gaussian width is the only rigid-motion-invariant, 1-homogeneous, continuous valuation on $\mathbb{R}^n$.

In this context, the radius of a set is defined as

$$\operatorname{rad}(\mathsf{T}) := \sup\{\|\boldsymbol{u}\|_2 : \boldsymbol{u} \in \mathsf{T}\}.$$

11.1.2 The supremum of a bilinear form

It is a remarkable fact that the supremum of a bilinear form in a Gaussian matrix is controlled by two linear forms over Gaussian vectors.

Theorem 11.3 (Chevet). Assume that $\mathsf{U} \subseteq \mathbb{R}^m$ and $\mathsf{V} \subseteq \mathbb{R}^n$ are compact subsets of the *unit spheres* in their respective spaces. Let $\boldsymbol{\Gamma} \in \mathbb{R}^{n \times m}$ and $\boldsymbol{g} \in \mathbb{R}^m$ and $\boldsymbol{h} \in \mathbb{R}^n$ have independent standard normal entries. Then

$$\mathbb{E} \max_{\substack{\boldsymbol{u} \in \mathsf{U} \\ \boldsymbol{v} \in \mathsf{V}}} \langle \boldsymbol{\Gamma} \boldsymbol{u},\ \boldsymbol{v} \rangle \le \mathbb{E} \max_{\substack{\boldsymbol{u} \in \mathsf{U} \\ \boldsymbol{v} \in \mathsf{V}}} [\langle \boldsymbol{g},\ \boldsymbol{u} \rangle + \langle \boldsymbol{h},\ \boldsymbol{v} \rangle] = w(\mathsf{U}) + w(\mathsf{V}).$$

Furthermore, for all $t \in \mathbb{R}$,

$$\mathbb{P} \left\{ \max_{\substack{\boldsymbol{u} \in \mathsf{U} \\ \boldsymbol{v} \in \mathsf{V}}} \langle \boldsymbol{\Gamma} \boldsymbol{u},\ \boldsymbol{v} \rangle > t \right\} \le 2\, \mathbb{P} \left\{ \max_{\substack{\boldsymbol{u} \in \mathsf{U} \\ \boldsymbol{v} \in \mathsf{V}}} [\langle \boldsymbol{g},\ \boldsymbol{u} \rangle + \langle \boldsymbol{h},\ \boldsymbol{v} \rangle] > t \right\}.$$

The beauty of Chevet's theorem is that it allows us to "decouple" the sets U and V that appear in the maximum. Another important feature is that the expectation bound is sharp up to and including the constants. In the next section, we will show how Chevet's theorem leads to a bound on the norm of a Gaussian matrix. There are many other elegant applications.

Proof. We shall apply Slepian's lemma to some carefully chosen GPs. Let $\gamma \sim \textsc{normal}(0, 1)$, independent of everything else. Consider the Gaussian processes $(X_{\boldsymbol{uv}} : \boldsymbol{u} \in \mathsf{U}, \boldsymbol{v} \in \mathsf{V})$ and $(Y_{\boldsymbol{uv}} : \boldsymbol{u} \in \mathsf{U}, \boldsymbol{v} \in \mathsf{V})$ defined by the relations

$$\begin{aligned} Y_{\boldsymbol{uv}} &= \langle \boldsymbol{\Gamma} \boldsymbol{u},\ \boldsymbol{v} \rangle + \gamma; \\ X_{\boldsymbol{uv}} &= \langle \boldsymbol{g},\ \boldsymbol{u} \rangle + \langle \boldsymbol{h},\ \boldsymbol{v} \rangle. \end{aligned}$$

Since everything is independent, we easily compute the covariances:

$$\begin{aligned} \mathbb{E}[Y_{\boldsymbol{uv}} Y_{\boldsymbol{u'v'}}] &= \mathbb{E}[\langle \boldsymbol{\Gamma} \boldsymbol{u},\ \boldsymbol{v} \rangle \langle \boldsymbol{\Gamma} \boldsymbol{u'},\ \boldsymbol{v'} \rangle] + 1 = \langle \boldsymbol{u},\ \boldsymbol{u'} \rangle \langle \boldsymbol{v},\ \boldsymbol{v'} \rangle + 1; \\ \mathbb{E}[X_{\boldsymbol{uv}} X_{\boldsymbol{u'v'}}] &= \mathbb{E}[\langle \boldsymbol{g},\ \boldsymbol{u} \rangle \langle \boldsymbol{g},\ \boldsymbol{u'} \rangle + \langle \boldsymbol{h},\ \boldsymbol{v} \rangle \langle \boldsymbol{h},\ \boldsymbol{v'} \rangle] = \langle \boldsymbol{u},\ \boldsymbol{u'} \rangle + \langle \boldsymbol{v},\ \boldsymbol{v'} \rangle. \end{aligned}$$

Thus, we have the comparison

$$\mathbb{E}[Y_{\boldsymbol{uv}} Y_{\boldsymbol{u'v'}}] - \mathbb{E}[X_{\boldsymbol{uv}} X_{\boldsymbol{u'v'}}] = (1 - \langle \boldsymbol{u},\ \boldsymbol{u'} \rangle)(1 - \langle \boldsymbol{v},\ \boldsymbol{v'} \rangle) \ge 0,$$

with equality when $\boldsymbol{u} = \boldsymbol{u'}$ or $\boldsymbol{v} = \boldsymbol{v'}$. Indeed, $\langle \boldsymbol{u},\ \boldsymbol{u'} \rangle \le 1$ by the Cauchy–Schwarz inequality because $\boldsymbol{u}$ and $\boldsymbol{u'}$ are unit vectors. Likewise, $\langle \boldsymbol{v},\ \boldsymbol{v'} \rangle \le 1$.

Thus, for any finite subsets $\mathsf{U'} \subseteq \mathsf{U}$ and $\mathsf{V'} \subseteq \mathsf{V}$, invoking Slepian's lemma yields

$$\begin{aligned} \mathbb{E} \max_{\substack{\boldsymbol{u} \in \mathsf{U'} \\ \boldsymbol{v} \in \mathsf{V'}}} \langle \boldsymbol{\Gamma} \boldsymbol{u},\ \boldsymbol{v} \rangle &= \mathbb{E} \max_{\substack{\boldsymbol{u} \in \mathsf{U'} \\ \boldsymbol{v} \in \mathsf{V'}}} [\langle \boldsymbol{\Gamma} \boldsymbol{u},\ \boldsymbol{v} \rangle + \gamma] \\ &= \mathbb{E} \max_{\substack{\boldsymbol{u} \in \mathsf{U'} \\ \boldsymbol{v} \in \mathsf{V'}}} Y_{\boldsymbol{uv}} \\ &\le \mathbb{E} \max_{\substack{\boldsymbol{u} \in \mathsf{U'} \\ \boldsymbol{v} \in \mathsf{V'}}} X_{\boldsymbol{uv}} = \mathbb{E} \max_{\substack{\boldsymbol{u} \in \mathsf{U} \\ \boldsymbol{v} \in \mathsf{V}}} [\langle \boldsymbol{g},\ \boldsymbol{u} \rangle + \langle \boldsymbol{h},\ \boldsymbol{v} \rangle]. \end{aligned}$$

The result follows by the definition of the lattice supremum or by an approximation argument. ∎

Exercise 11.4 (Chevet: Probability bound). Prove the probability bound in Theorem 11.3. **Hint:** Condition on the event $\{\gamma \ge 0\}$.

We conclude this section we some extensions of Chevet's theorem; see Section 8.7 of [Ver18] for further discussion. First, we note that Chevet's theorem admits a matching lower bound.

Exercise 11.5 (Chevet: Lower bound). Under the assumptions of Theorem 11.3,

$$\mathbb{E} \max_{\substack{\boldsymbol{u} \in \mathsf{U} \\ \boldsymbol{v} \in \mathsf{V}}} \langle \boldsymbol{\Gamma u},\ \boldsymbol{v} \rangle \geq \max\{w(\mathsf{U}),\ w(\mathsf{V})\}.$$

Hint: Use Jensen's inequality and the rotational invariance of the standard normal distribution.

Next, we consider what happens for sets U and V that may not contained in the unit sphere.

Exercise 11.6 (Chevet: General sets). For general compact sets $\mathsf{U} \subseteq \mathbb{R}^m$ and $\mathsf{V} \subseteq \mathbb{R}^n$, show that

$$\mathbb{E} \max_{\substack{\boldsymbol{u} \in \mathsf{U} \\ \boldsymbol{v} \in \mathsf{V}}} \langle \boldsymbol{\Gamma u},\ \boldsymbol{v} \rangle \leq w(\mathsf{U})\ \operatorname{rad}(\mathsf{V}) + w(\mathsf{V})\ \operatorname{rad}(\mathsf{U}),$$

where $\operatorname{rad}(\mathsf{U}) := \sup_{\boldsymbol{u} \in \mathsf{U}} \|\boldsymbol{u}\|_2$. Develop a matching lower bound.

Chevet's theorem extends to matrices with independent subgaussian entries by way of a difficult comparison argument.

Theorem 11.7 (Subgaussian Chevet). Let $\mathsf{U} \subseteq \mathbb{R}^m$ and $\mathsf{V} \subseteq \mathbb{R}^n$ be compact. Let $\boldsymbol{\Gamma} \in \mathbb{R}^{n \times m}$ be a matrix whose entries are independent, centered σ^2-subgaussian random variables. Then

$$\mathbb{E} \max_{\substack{\boldsymbol{u} \in \mathsf{U} \\ \boldsymbol{v} \in \mathsf{V}}} \langle \boldsymbol{\Gamma u},\ \boldsymbol{v} \rangle \leq \text{Const} \cdot \sigma \cdot \left[w(\mathsf{U})\ \operatorname{rad}(\mathsf{V}) + w(\mathsf{V})\ \operatorname{rad}(\mathsf{U}) \right].$$

The proof depends on Exercise 11.6 and Talagrand's comparison inequality, which will be discussed in Lecture 13.

11.2 Spectral norm of a Gaussian matrix

As a first application of Chevet's theorem, we can compute sharp bounds for the spectral norm of a Gaussian matrix. These bounds are very important in algorithmic applications of Gaussian matrices, including randomized SVD algorithms [HMT11].

Corollary 11.8 (Norm of a standard Gaussian matrix). Suppose that $\boldsymbol{\Gamma} \in \mathbb{R}^{n \times m}$ has iid standard normal entries. Then

$$\mathbb{E} \|\boldsymbol{\Gamma}\|_{\ell_2 \to \ell_2} \leq \sqrt{m} + \sqrt{n}.$$

Here, $\|\cdot\|_{\ell_2 \to \ell_2}$ is the ℓ_2 operator norm, also known as the spectral norm.

Proof. Let $\boldsymbol{g} \in \mathbb{R}^m$ and $\boldsymbol{h} \in \mathbb{R}^n$ have iid standard normal entries. By definition of the operator norm and by Chevet's theorem,

$$\begin{aligned} \mathbb{E} \|\boldsymbol{\Gamma}\|_{\ell_2 \to \ell_2} &= \mathbb{E} \max_{\|\boldsymbol{u}\|_{\ell_2} = \|\boldsymbol{v}\|_{\ell_2} = 1} \langle \boldsymbol{\Gamma u},\ \boldsymbol{v} \rangle \\ &\leq \mathbb{E} \max_{\|\boldsymbol{u}\|_{\ell_2} = \|\boldsymbol{v}\|_{\ell_2} = 1} \left[\langle \boldsymbol{g},\ \boldsymbol{u} \rangle + \langle \boldsymbol{h},\ \boldsymbol{v} \rangle \right] = \mathbb{E} \left[\|\boldsymbol{g}\|_{\ell_2} + \|\boldsymbol{h}\|_{\ell_2} \right]. \end{aligned}$$

By Lyapunov's inequality, $\mathbb{E} \|\boldsymbol{g}\|_{\ell_2} \leq (\mathbb{E} \|\boldsymbol{g}\|_{\ell_2}^2)^{1/2} = \sqrt{m}$. Likewise, $\mathbb{E} \|\boldsymbol{h}\|_{\ell_2} \leq \sqrt{n}$. These observations complete the proof. ∎

Problem 11.9 (Optimality). Using the Marčenko–Pastur theorem, prove that Corollary 11.8 is sharp—up to and including the constants.

Exercise 11.10 (Spectral norm via matrix concentration). One might wonder how the bound we obtained using Chevet's theorem, which is specialized for standard normal matrices, compares with bounds developed with more general matrix concentration tools. Show that matrix concentration inequalities (e.g., the noncommutative Khintchine inequality) only yield the bound

$$\mathbb{E}\,\|\mathbf{\Gamma}\|_{\ell_2\to\ell_2} \le \text{const}\cdot\sqrt{(m+n)\log(m+n)}.$$

Chevet's bound improves on this result in two important ways: it removes the logarithm and it replaces the square root of the sum by the sum of the square roots.

Exercise 11.11 (Inhomogeneous norm bound). Use the generalized Chevet theorem (Exercise 11.6) to develop the following spectral norm bound for the product of a Gaussian matrix with fixed matrices:

Here, $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm.

$$\mathbb{E}\,\|\boldsymbol{B}\mathbf{\Gamma}\boldsymbol{C}\|_{\ell_2\to\ell_2} \le \|\boldsymbol{B}\|_{\ell_2\to\ell_2}\|\boldsymbol{C}\|_{\mathrm{F}} + \|\boldsymbol{B}\|_{\mathrm{F}}\|\boldsymbol{C}\|_{\ell_2\to\ell_2}$$

where $\boldsymbol{B}\in\mathbb{R}^{k\times n}$ and $\boldsymbol{C}\in\mathbb{R}^{m\times p}$ are fixed matrices and $\mathbf{\Gamma}\in\mathbb{R}^{n\times m}$ is a matrix with iid standard normal entries. Taking $\boldsymbol{B}=\mathbf{I}_n$ and $\boldsymbol{C}=\mathbf{I}_m$, this result implies Corollary 11.8.

11.3 Sudakov's minoration

In this section, we shall see a second application of Gaussian comparison inequalities: a *lower bound* for the expected supremum of a Gaussian process. As discussed, we anticipate that the size of the supremum will be related to the "complexity" or "geometry" of the index set. As we shall see, the covering number provides a natural quantity to characterize this complexity. In this section, we shall see how to obtain lower bounds on the expectation of suprema using covering numbers; upper bounds shall be the subject of next lecture.

11.3.1 Gaussian processes and geometry

Let us quickly review the geometry of Gaussian processes. Consider a centered GP $(X_t : t\in\mathsf{T})$. The Gaussian process naturally equips the index set T with a pseudometric

$$d(s,t) := \|X_s - X_t\|_{\mathrm{L}_2} = (\mathbb{E}(X_s-X_t)^2)^{1/2}.$$

By the polarization identities, the covariance function $\Sigma(s,t)=\mathbb{E}[X_sX_t]$ is a function of the increments $d(s,t)$. A Gaussian process is determined by its expectation and covariances, so the pseudometric space (T,d) contains all there is to know about a centered Gaussian process.

The polarization identity states that for a real inner product space $(\mathsf{X},\langle\cdot,\cdot\rangle)$ with induced norm $\|\boldsymbol{x}\|=\sqrt{\langle\boldsymbol{x},\boldsymbol{x}\rangle}$ the inner product between vectors $\boldsymbol{x},\boldsymbol{y}\in\mathsf{X}$ is given by $\langle\boldsymbol{x},\boldsymbol{y}\rangle=\frac{1}{4}(\|\boldsymbol{x}+\boldsymbol{y}\|^2-\|\boldsymbol{x}-\boldsymbol{y}\|^2)$. Thus, given a black box to compute norms, one can compute inner products as well.

We wish to study the supremum $\sup_{t\in\mathsf{T}}X_t$ of a centered Gaussian process. The deviations of $\sup_{t\in\mathsf{T}}X_t$ from its mean value are well-characterized by Gaussian concentration. Specifically, we have

$$\mathbb{P}\left\{|\sup_{t\in\mathsf{T}}X_t - \mathbb{E}\sup_{t\in\mathsf{T}}X_t| > u\right\} \le 2\exp\left(-\frac{u^2}{2\sigma^2}\right),$$

where $\sigma^2 := \sup_{t\in\mathsf{T}}\mathbb{E}\,X_t^2$. Consequently, our main goal is to understand the expected value, $\mathbb{E}\sup_{t\in\mathsf{T}}X_t$, around which $\sup_{t\in\mathsf{T}}X_t$ concentrates.

11.3.2 Minoration

We shall provide a *lower bound* on $\mathbb{E}\sup_{t\in\mathsf{T}}X_t$ in terms of the *metric entropy* $\log_2\mathrm{N}(\mathsf{T},d;\varepsilon)$ of the space (T,d) on the scale ε. As usual, $\mathrm{N}(\mathsf{T},d;\varepsilon)$ is the covering number of (T,d). The following bound is due to V. N. Sudakov.

Recall that the covering number $\mathrm{N}(\mathsf{T},d;\varepsilon)$ is the minimum number of ε-balls in the (pseudo)metric d that suffice to cover T.

Theorem 11.12 (Sudakov's minoration). Let $(X_t : t \in \mathsf{T})$ be a centered GP. For each $\varepsilon > 0$,

$$\mathbb{E} \sup_{t\in\mathsf{T}} X_t \geq \text{const} \cdot \varepsilon \sqrt{\log \mathrm{N}(\mathsf{T}, d; \varepsilon)},$$

where d is the canonical (pseudo)metric of the process.

Since $\log_2(a) = \log(a)/\log(2)$ for every $a > 0$, we could just as well use the base-2 logarithm.

Sudakov minoration can be used in two distinct ways. First, we can convert lower bounds on covering numbers into lower bounds on the suprema of Gaussian processes. Second, we can convert upper bounds on suprema of Gaussian processes into upper bounds on covering numbers. We develop some examples after proving the theorem.

Proof. The proof uses the Fernique–Sudakov comparison inequality. Without loss, we may assume that $\mathrm{N}(\mathsf{T}, \varepsilon; d) < +\infty$. If not, then $\mathbb{E} \sup_{t\in\mathsf{T}} X_t = +\infty$. (Why?)

Let N_ε be a maximal ε-separated set in T. Then N_ε is an ε-net for T with cardinality $|\mathsf{N}_\varepsilon| \geq \mathrm{N}(\mathsf{T}, d; \varepsilon)$. See Figure 11.1 for an illustration.

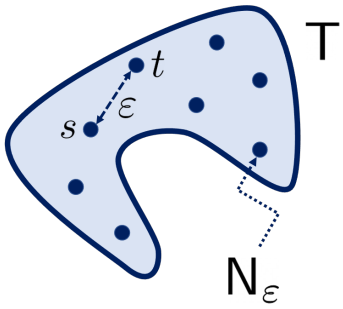


Figure 11.1 Illustration of a maximal ε-separated subset $\mathsf{N}_\varepsilon \subseteq \mathsf{T}$.

Let us compare $(X_t : t \in \mathsf{N}_\varepsilon)$ with another Gaussian process $(Y_t : t \in \mathsf{N}_\varepsilon)$ consisting of independent Gaussian variables. Specifically, let $Y_t = \frac{1}{\sqrt{2}} \varepsilon g_t$, where $(g_t : t \in \mathsf{N}_\varepsilon)$ consists of iid standard normal random variables. We can compare the increments of these processes:

$$\begin{aligned} \mathbb{E}(X_s - X_t)^2 &= [d(s,t)]^2 \geq \varepsilon^2; \\ \mathbb{E}(Y_s - Y_t)^2 &= \frac{\varepsilon^2}{2} \mathbb{E}(g_s - g_t)^2 = \varepsilon^2 \end{aligned} \qquad \text{for } s, t \in \mathsf{N}_\varepsilon.$$

To handle the process (X_t), we used the fact that N_ε is an ε-net with respect to d. To handle the process (Y_t), we simply apply the fact that it consists of independent standard normal variables.

We have shown that the increments of $(X_t : t \in \mathsf{N}_\varepsilon)$ dominate those of $(Y_t : t \in \mathsf{N}_\varepsilon)$. By the Sudakov–Fernique comparison theorem,

$$\mathbb{E} \sup_{t\in\mathsf{T}} X_t \geq \mathbb{E} \sup_{t\in\mathsf{N}_\varepsilon} X_t \geq \mathbb{E} \sup_{t\in\mathsf{N}_\varepsilon} Y_t = \frac{\varepsilon}{\sqrt{2}} \mathbb{E} \max_{t\in\mathsf{N}_\varepsilon} g_t.$$

To complete the argument, we employ a basic lower bound for the expected maximum of iid standard normal variables:

$$\mathbb{E} \max_{t\in\mathsf{N}_\varepsilon} g_t \geq \text{const} \cdot \sqrt{\log |\mathsf{N}_\varepsilon|} \geq \text{const} \cdot \sqrt{\log \mathrm{N}(\mathsf{T}, d; \varepsilon)}.$$

The last relation holds because the covering number is the minimum cardinality of an ε-net. Combine the last two displays to complete the proof. ∎

Problem 11.13 (Maximum of Gaussians). Let $(g_i : i = 1, \dots, m)$ be an iid family of standard normal random variables. Confirm that

$$\text{const} \cdot \sqrt{\log m} \leq \mathbb{E} \max_{1\leq i\leq m} g_i \leq \sqrt{2 \log m}.$$

Prove the asymptotic formula

$$\lim_{m\to\infty} \frac{\mathbb{E} \max_{1\leq i\leq m} g_i}{\sqrt{2 \log m}} = 1.$$

11.3.3 From lower bounds for covering numbers to lower bounds for suprema

As a first illustration of Sudakov's minoration, we will show how to use volumetric lower bounds for covering numbers to obtain lower bounds for Gaussian processes.

For a set $\mathsf{T} \subseteq \mathbb{R}^n$, construct the canonical Gaussian process $X_{\boldsymbol{t}} = \langle \boldsymbol{g}, \boldsymbol{t} \rangle$, where $\boldsymbol{g} \in \mathbb{R}^n$ is a standard normal vector. The canonical metric d coincides with the ℓ_2 distance: $d(\boldsymbol{s}, \boldsymbol{t}) = \|\boldsymbol{s} - \boldsymbol{t}\|_{\ell_2}$. The volumetric bound for covering numbers gives

$$\mathrm{N}(\mathsf{T}, \ell_2; \varepsilon) \geq \frac{\mathrm{Vol}(\mathsf{T})}{\mathrm{Vol}(\mathsf{B}_2)} \cdot \varepsilon^{-n},$$

where B_2 is the ℓ_2 unit ball in $\mathbb{R}^n$. As a consequence,

$$\log \mathrm{N}(\mathsf{T}, \ell_2; \varepsilon) \geq n \log(1/\varepsilon) + \log \frac{\mathrm{Vol}(\mathsf{T})}{\mathrm{Vol}(\mathsf{B}_2)}.$$

Combining this estimate with Sudakov's minoration leads to lower bounds on the expected supremum of the canonical GP:

$$\mathbb{E} \sup_{\boldsymbol{t} \in \mathsf{T}} X_{\boldsymbol{t}} \geq \mathrm{const} \cdot \sup_{\varepsilon > 0} \varepsilon \sqrt{n \log(1/\varepsilon) + \log \mathrm{Vol}(\mathsf{T}) - \log \mathrm{Vol}(\mathsf{B}_2)}.$$

For any set T, this bound is qualitatively correct for very small values of ε.

Example 11.14 (Norm of a standard Gaussian vector). We use the lower bound on metric entropy from the previous example to obtain a lower bound for the norm of a standard Gaussian vector. With the notation of this section, we consider the set $\mathsf{T} = \mathsf{B}_2$. Then Sudakov's minoration yields

$$\mathbb{E} \|\boldsymbol{g}\|_{\ell_2} = \mathbb{E} \sup_{\boldsymbol{t} \in \mathsf{B}_2} \langle \boldsymbol{g}, \boldsymbol{t} \rangle \geq \mathrm{const} \cdot \sup_{\varepsilon > 0} \varepsilon \sqrt{n \log(1/\varepsilon)} \geq \mathrm{const} \cdot \sqrt{n}.$$

This bound is qualitatively correct because $\mathrm{const} \cdot \sqrt{n} \leq \mathbb{E} \|\boldsymbol{g}\|_{\ell_2} \leq \sqrt{n}$. ▪

11.3.4 From upper bounds for suprema to upper bounds for covering numbers

Now, let us show how upper bounds on the supremum of a Gaussian process can be used to derive upper bounds for covering numbers.

We consider the convex hull of m points in the unit ball of $\mathbb{R}^n$. Define $\mathsf{A} = \{\boldsymbol{a}_1, \dots, \boldsymbol{a}_m\} \subseteq \mathsf{B}_2$, and form $\mathsf{T} = \mathrm{conv}(\mathsf{A})$. Let $\boldsymbol{g} \in \mathbb{R}^n$ be a standard normal vector. By the Gaussian maximal inequality (Problem II.13),

$$\mathbb{E} \sup_{\boldsymbol{t} \in \mathsf{T}} \langle \boldsymbol{g}, \boldsymbol{t} \rangle = \mathbb{E} \max_{\boldsymbol{a} \in \mathsf{A}} \langle \boldsymbol{g}, \boldsymbol{a} \rangle \leq \sqrt{2 \log m}.$$

Indeed, a linear function on a compact convex set achieves its maximum on an extreme point! Then, for all $\varepsilon > 0$, by Sudakov minoration,

$$\mathrm{const} \cdot \varepsilon \sqrt{\log \mathrm{N}(\mathsf{T}, \ell_2; \varepsilon)} \leq \sqrt{2 \log m}$$

which implies

$$\log \mathrm{N}(\mathsf{T}, \ell_2; \varepsilon) \leq m^{\mathrm{const}/\varepsilon^2}.$$

Up to the value of the constant, this matches the empirical bound for the covering number of a convex hull due to Maurey.

Lecture bibliography

[Che77] S. Chevet. "Séries de variables aléatoires Gaussiennes à valeurs dans $E\hat{\otimes}_\varepsilon F$. Application aux produits d'espaces de Wiener abstraits". In: *Séminaire sur ia Géométrie des Espaces de Banach (1977-1978), Exp. No. 19, 15, École Polytechnique, Palaiseau* (1977).

[Gor85] Y. Gordon. "Some Inequalities for Gaussian Processes and Applications". In: *Israel Journal of Mathematics* 50.4 (1985), pages 265–289.

[HMT11] N. Halko, P. G. Martinsson, and J. A. Tropp. "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions". In: *SIAM Review* 53.2 (Jan. 2011), pages 217–288. DOI: 10.1137/090771806.

[Ver18] R. Vershynin. *High-dimensional probability*. An introduction with applications in data science, With a foreword by Sara van de Geer. Cambridge University Press, Cambridge, 2018. DOI: 10.1017/9781108231596.

12. Dudley's Inequality

Date: 11 February 2020 **Scribe:** Hamed Hamze

Agenda:
1. Dudley's inequality
2. Chaining
3. Proof of Dudley
4. Extensions
5. Examples

In the last lecture, we obtained a lower bound on the supremum of a centered Gaussian process by using Sudakov's minoration. More precisely, let $(X_t : t \in \mathsf{T})$ be a centered GP with canonical metric $d(s,t) := \|X_s - X_t\|_{L_2}$. Sudakov's minoration shows that

$$\mathbb{E} \sup_{t\in\mathsf{T}} X_t \geq \text{const} \cdot \sup_{\varepsilon>0} \varepsilon\sqrt{\log \mathrm{N}(\mathsf{T}, d; \varepsilon)}. \tag{12.1}$$

Sudakov's bound replaces a probabilistic question by a geometric one. We can use the metric entropy $\log \mathrm{N}(\mathsf{T}, d; \varepsilon)$, which reflects the complexity of the metric space (T, d), to produce a lower bound for the supremum of a centered GP. Hence, whenever the metric entropy is large, the supremum must also be large.

In the previous lecture, we saw two distinct ways to use Sudakov's minoration:

- A lower bound on the metric entropy gives a lower bound on the supremum.
- An upper bound on the supremum gives an upper bound on the metric entropy.

Today, we are going to show how to obtain an upper bound on the supremum of the process in terms of metric entropy.

12.1 Dudley's inequality

In this section, we introduce an upper bound on the supremum of a GP. This result is known as *Dudley's inequality*.

Theorem 12.1 (Dudley's inequality). Let $(X_t : t \in \mathsf{T})$ be a centered Gaussian process with canonical metric d. Then

$$\mathbb{E} \sup_{t\in\mathsf{T}} X_t \leq \text{Const} \int_0^\infty \mathrm{d}\varepsilon \sqrt{\log \mathrm{N}(\mathsf{T}, d; \varepsilon)}. \tag{12.2}$$

The upper limit of the integral can be set as $\operatorname{diam}(\mathsf{T}, d) := \max\{d(s,t) : s, t \in \mathsf{T}\}$.

The only technical assumption is the separability of the process; that is, T contains a countable subset that approximates the GP arbitrarily well.

Figure 12.1 illustrates the relation between Dudley's inequality and Sudakov's minoration. As we can see in the figure, the bounds have graphical interpretations.

- **Sudakov.** The supremum is bounded below by the area of the largest rectangle under the curve.
- **Dudley.** The supremum is bounded above by the area under the curve.

Recall that the metric entropy $\varepsilon \mapsto \mathrm{N}(\mathsf{T}, d; \varepsilon)$ is a decreasing function.

Since the metric entropy is positive, it is clear that Sudakov's minoration is smaller than the Dudley's inequality. Together, the two results show that the supremum of the process lies somewhere in between these two extremes.

The quantity on the right-hand of the bound (12.2) is called an *entropy integral*. It accumulates the total metric entropy of the space (T, d) across all scales, so Dudley's bound is a multiscale estimate. By using a Riemann–Darboux sum, we can approximate the entropy integral using discrete scales. In other words, proving the discrete version of Dudley's bound is equivalent to proving the integral formulation.

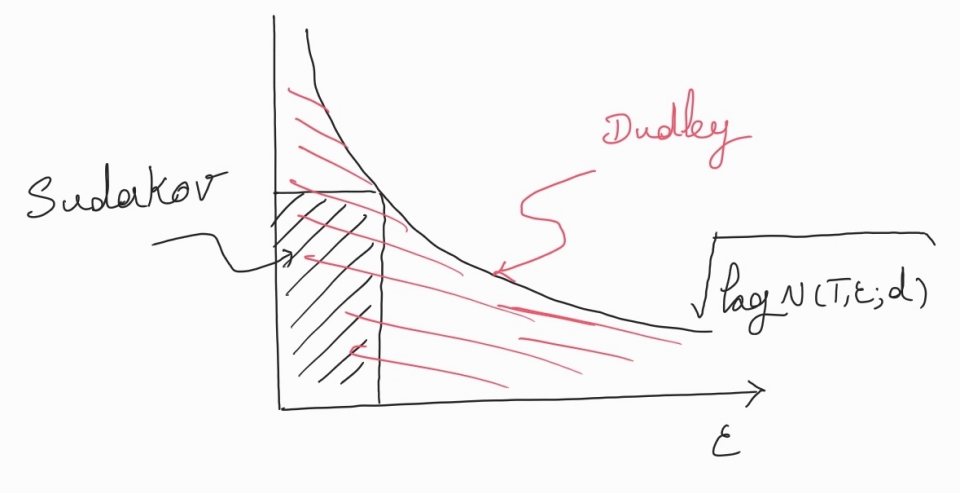


Figure 12.1 (Sudakov versus Dudley). Comparison of Sudakov's minoration (12.1) and Dudley's bound (12.2).

Theorem 12.2 (Dudley's inequality: Discrete version). Let $(X_t : t \in \mathsf{T})$ be a centered Gaussian process with canonical metric d. Then

$$\mathbb{E} \sup_{t\in\mathsf{T}} X_t \le \text{Const} \sum\nolimits_{i\in\mathbb{Z}} 2^{-i} \sqrt{\log \mathrm{N}(\mathsf{T}, d; 2^{-i})}.$$

In the next section, we give an intuitive presentation of the main idea behind the proof of Theorem 12.2. The detailed proof appears in Section 12.3.

12.2 Chaining

The main idea behind the proof of Theorem 12.2 is a technique called *chaining*. This method allows us to combine ε-nets on different scales. Here is an overview.

For each $\varepsilon > 0$, let N_ε be an ε-net for (T, d). For each index $t \in \mathsf{T}$ and each scale ε, we may approximate t by a closest point $\pi_\varepsilon(t)$ in N_ε. That is,

$$\pi_\varepsilon(t) \in \arg\min\{d(s,t) : s \in \mathsf{N}_\varepsilon\}.$$

(We may assume that the ε-nets are finite, so ties are broken deterministically.) Since N_ε is an ε-net with respect to the canonical metric, we have $d(t, \pi_\varepsilon(t)) \le \varepsilon$. As a consequence,

$$\|X_t - X_{\pi_\varepsilon(t)}\|_{\mathrm{L}_2} = d(t, \pi_\varepsilon(t)) \le \varepsilon.$$

This controls the correlation between X_t and its approximation $X_{\pi_\varepsilon(t)}$ at scale ε.

We may now decompose the supremum over the index set by approximating each point using the ε-net at a given scale:

$$\mathbb{E} \sup_{t\in\mathsf{T}} X_t \le \mathbb{E} \sup_{t\in\mathsf{N}_\varepsilon} X_{\pi_\varepsilon(t)} + \mathbb{E} \sup_{t\in\mathsf{T}} (X_t - X_{\pi_\varepsilon(t)}).$$

The inequality holds because the supremum is subadditive. Since the projection $\pi_\varepsilon(t)$ only takes values in N_ε, we may restrict the first supremum to this net, whose cardinality increases as the scale ε decreases.

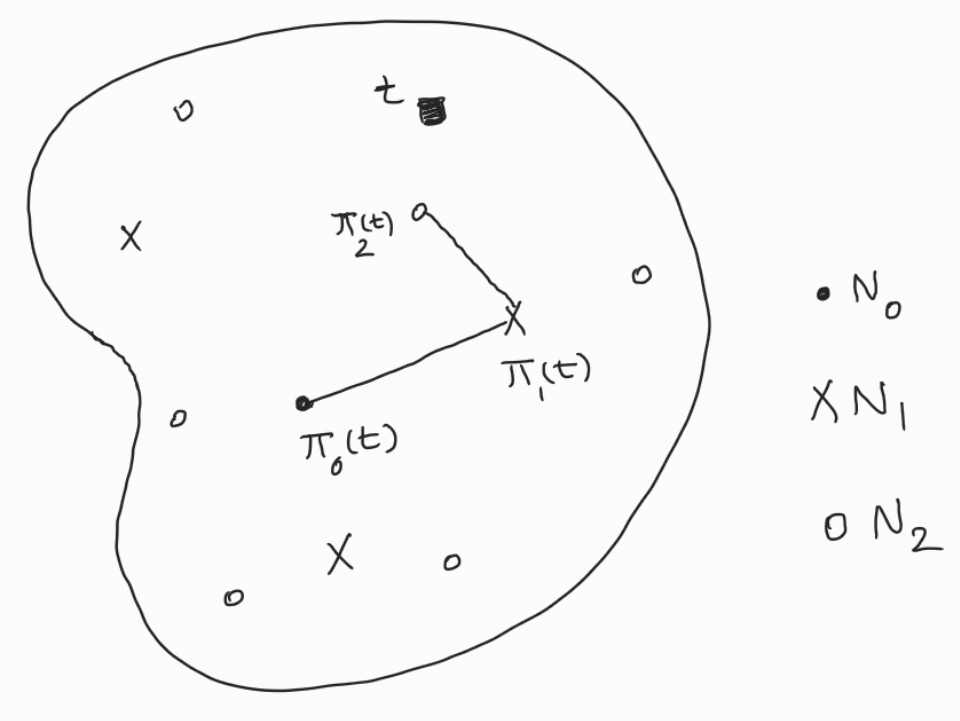


Figure 12.2 (Chaining). As the nets become finer, we have a better approximation of t. However, the number of possible links in the chain is increasing.

The last display suggests an approach to controlling the supremum. We can use a maximal inequality for finite-cardinality Gaussian processes (which is essentially a union bound) to control the first term. The second term still ranges over the whole index set T, but the random variables $X_t - X_{\pi_\varepsilon(t)}$ have smaller variances because $d(t, \pi_\varepsilon(t)) \leq \varepsilon$. To handle the second term, we will iterate the argument by introducing another net on a finer scale. The key is to balance the cardinality of the net against the variance of the random variables appearing in the supremum.

Figure 12.2 gives a clearer illustration of what we plan to do. We increase the number of points in the net, while approximating the index set more finely. As in Figure 12.2, let us consider a fixed point $t \in \mathsf{T}$ and develop a multiscale decomposition.

- First, we approximate t by an initial net, labeled N_0, containing only one point. The closest point $\pi_0(t)$ may not a good approximation of t, but the cardinality of the net is minimal.
- Next, we approximate $\pi_0(t)$ by a point $\pi_1(t)$ in the next net, labeled N_1. The net N_1 has larger cardinality, so $\pi_1(t)$ yields a better approximation of t.
- Again, we approximate $\pi_1(t)$ by a point $\pi_2(t)$ in the next net, labeled N_2. The net N_2 is still larger, and the approximation $\pi_2(t)$ of t is still better.
- We repeat this process indefinitely so that $\pi_i(t) \to t$ as i increases.
- As we do so, the length $d(\pi_i(t), \pi_{i-1}(t))$ of each link in the chain grows shorter, which reflect the increasing similarity of the random variables $X_{\pi_{i-1}(t)}$ and $X_{\pi_i(t)}$. At the same time, the number of possible links from level $i-1$ to level i is growing. These two factors must counterbalance each other.

In the next section, we fill out this sketch to establish Dudley's inequality.

12.3 Proof of Dudley's inequality

In this section, we first prove the discrete version of Dudley's theorem (Theorem 12.2). Afterward, by approximating the integral using Riemann–Darboux sum, we obtain the integral version (Theorem 12.1).

Proof. By making an approximation argument, we may assume without the loss of generality that $|\mathsf{T}| < \infty$. This step requires the separability of the Gaussian process.

Step 1: Chaining

For each $i \in \mathbb{Z}$, we define the length scale $\varepsilon_i := 2^{-i}$. At each scale, we fix an ε_i-net T_i with $|\mathsf{T}_i| = \mathrm{N}(\mathsf{T}, d; \varepsilon_i)$. Since the cardinality of the index set T is finite, there is a coarsest scale $k \in \mathbb{Z}$ and a finest scale $K \in \mathbb{Z}$ with $k \le K$ for which

$$\mathsf{T}_k = \{t_0\} \quad \text{and} \quad \mathsf{T}_K = \mathsf{T}.$$

Define the nearest-point functions

$$\pi_i(t) := \arg\min\{d(s,t) : s \in \mathsf{T}_i\},$$

with ties broken in a deterministic fashion. We may assume that $\pi_K(t) = t$ for all $t \in \mathsf{T}$. By construction,

$$d(t, \pi_i(t)) \le \varepsilon_i = 2^{-i} \quad \text{for all } t \in \mathsf{T} \text{ and all } k \le i \le K.$$

These functions allow us to construct a sequence of increasingly accurate approximations to each point t in the index set T.

To begin the chaining argument, we use the fact that the GP is centered to see that

$$\mathbb{E}\sup_{t\in\mathsf{T}} X_t = \mathbb{E}\sup_{t\in\mathsf{T}} (X_t - X_{t_0}) \quad \text{where} \quad \mathsf{T}_k = \{t_0\}.$$

We can decompose each increment of this process into a telescoping sum:

$$X_t - X_{t_0} = X_{\pi_K(t)} - X_{\pi_k(t)} = \sum_{i=k+1}^{K} (X_{\pi_i(t)} - X_{\pi_{i-1}(t)}).$$

Using subadditivity of the supremum, we obtain the simple bound

$$\mathbb{E}\sup_{t\in\mathsf{T}} (X_t - X_{t_0}) \le \sum_{i=k+1}^{K} \mathbb{E}\sup_{t\in\mathsf{T}} (X_{\pi_i(t)} - X_{\pi_{i-1}(t)}). \tag{12.3}$$

As it happens, the main source of error in the chaining argument is our decision to pass the sum through the supremum. In the next lecture, we will obtain better bounds by maintaining the supremum outside of the chaining decomposition.

Step 2: Bounding the increments

The next step in the argument is to obtain a bound for the expected supremum at each length scale. The argument depends on the maximal inequality for standard normal random variables.

Let us consider the contribution to the sum at a fixed scale i:

$$\mathbb{E}\sup_{t\in\mathsf{T}} (X_{\pi_i(t)} - X_{\pi_{i-1}(t)})$$

Since $\pi_i(t) \in \mathsf{T}_i$ and $\pi_{i-1}(t) \in \mathsf{T}_{i-1}$, we realize that the number of terms in this supremum is bounded by

$$|\mathsf{T}_i||\mathsf{T}_{i-1}| \le |\mathsf{T}_i|^2 = \mathrm{N}(\mathsf{T}, d; \varepsilon_i)^2.$$

Indeed, the nets are increasing in cardinality as i increases. Furthermore, we can control the variance of each term:

$$\begin{aligned}\|X_{\pi_i(t)} - X_{\pi_{i-1}(t)}\|_{\mathsf{L}_2} &= d(\pi_i(t), \pi_{i-1}(t)) \\ &\le d(\pi_i(t), t) + d(t, \pi_{i-1}(t)) \le \varepsilon_i + \varepsilon_{i-1} = 3\varepsilon_i.\end{aligned}$$

We have used the triangle inequality for the canonical metric d and the definition of the length scales ε_i. Therefore,

$$\mathbb{E} \sup_{t \in \mathsf{T}} (X_{\pi_i(t)} - X_{\pi_{i-1}(t)}) \le 6\varepsilon_{i-1} \sqrt{\log \mathrm{N}(\mathsf{T}, d; \varepsilon_i)}. \tag{12.4}$$

This is simply the maximal inequality for $|\mathsf{T}_i|^2$ Gaussian random variables, each with variance bounded by $(3\varepsilon_{i-1})^2$.

Step 3: Summing the increments

We may now complete the argument by introducing our bound (12.4) for the increments into the chaining inequality (12.3). Thus,

$$\mathbb{E} \sup_{t \in \mathsf{T}} (X_t - X_{t_0}) \le \sum\nolimits_{i=k+1}^{K} 6\varepsilon_i \sqrt{\log \mathrm{N}(\mathsf{T}, d; \varepsilon_i)}.$$

Recall that $\varepsilon_i = 2^{-i}$. Furthermore, if we extend the range of the sum to all $i \in \mathbb{Z}$, it only increases. This is the statement of Theorem 12.2.

Finally, we explain how to derive the integral formulation in Theorem 12.1. Recall that $\varepsilon \mapsto \mathrm{N}(\mathsf{T}, d; \varepsilon)$ is a decreasing function. Therefore, we have the chain of relations

$$\begin{aligned}\int_0^\infty \mathrm{d}\varepsilon \sqrt{\log \mathrm{N}(\mathsf{T}, d; \varepsilon)} &= \sum\nolimits_{i \in \mathbb{Z}} \int_{2^{-i}}^{2^{-i+1}} \mathrm{d}\varepsilon \sqrt{\log \mathrm{N}(\mathsf{T}, d; \varepsilon)} \\ &\le \sum\nolimits_{i \in \mathbb{Z}} 2^{-i} \sqrt{\log \mathrm{N}(\mathsf{T}, d; 2^{-i})} \\ &= 2 \sum\nolimits_{i \in \mathbb{Z}} 2^{-(i+1)} \sqrt{\log \mathrm{N}(\mathsf{T}, d; 2^{-i})} \\ &= 2 \sum\nolimits_{i \in \mathbb{Z}} \int_{2^{-(i+1)}}^{2^{-i}} \mathrm{d}\varepsilon \sqrt{\log \mathrm{N}(\mathsf{T}, d; \varepsilon)} \\ &= 2 \int_0^\infty \mathrm{d}\varepsilon \sqrt{\log \mathrm{N}(\mathsf{T}, d; \varepsilon)}.\end{aligned}$$

We deduce that the sum and integral presentations of Dudley's inequality are equivalent up to the precise constant. ∎

Exercise 12.3 (Subgaussian maximal inequality). Let $(X_i : i = 1, \dots, m)$ be a family of centered random variables that are all ν-subgaussian. Prove that

$$\mathbb{E} \max\nolimits_{1 \le i \le m} X_i \le \nu \sqrt{2 \log m}.$$

Recall that a centered ν-subgaussian random variable X has cgf $\xi_X(\theta) \le \nu\theta^2/2$ for all $\theta \in \mathbb{R}$.

12.4 Extensions

Dudley's inequality admits a number of refinements. The most important observation is that we did not use any special properties of standard normal variables in the argument. Indeed, we only used the fact that increments of the process have subgaussian tails. As a consequence, Dudley's inequality applies to a far wider class of random processes.

Definition 12.4 (Subgaussian process). Let $(\mathsf{T}, \mathrm{dist})$ be a (pseudo)metric space. Consider a centered random process $(X_t : t \in \mathsf{T})$ defined on the metric space. We say that the process is *subgaussian* with respect to the metric, dist, if each increment satisfies

$$\log \mathbb{E}\, \mathrm{e}^{\theta(X_s - X_t)} \leq \mathrm{dist}^2(s,t) \cdot \theta^2/2 \quad \text{for all } s, t \in \mathsf{T} \text{ and } \theta \in \mathbb{R}.$$

Theorem 12.5 (Dudley's inequality: Subgaussian process). Let $(X_t : t \in \mathsf{T})$ be a centered random process that is subgaussian with respect to the metric dist. Then

$$\mathbb{E} \sup\nolimits_{t \in \mathsf{T}} X_t \leq \mathrm{Const} \int_0^\infty \mathrm{d}\varepsilon \sqrt{\log \mathrm{N}(\mathsf{T}, \mathrm{dist}; \varepsilon)}.$$

We may take the upper limit of the integral as $\mathrm{diam}(\mathsf{T}, \mathrm{dist})$.

Exercise 12.6 (Dudley: Subgaussian processes). Prove Theorem 12.5.

Dudley's inequality can also be extended to obtain tail bounds for subgaussian random processes. In general, these estimates are not comparable with the bounds that follow from computing the expectation of the supremum and applying a concentration inequality.

Exercise 12.7 (Dudley: Tail bound). Consider a centered, subgaussian random process $(X_t : t \in \mathsf{T})$ on a metric space $(\mathsf{T}, \mathrm{dist})$. Prove that

$$\mathbb{P}\left\{ \sup\nolimits_{t \in \mathsf{T}} X_t \geq u + \mathrm{Const} \int_0^\infty \mathrm{d}\varepsilon \sqrt{\log \mathrm{N}(\mathsf{T}, \mathrm{dist}; \varepsilon)} \right\} \leq \mathrm{Const} \cdot \mathrm{e}^{-\mathrm{const} \cdot u^2 / \mathrm{diam}(\mathsf{T})^2}.$$

Exercise 12.8 (Dudley: Uncentered processes). We can also extend Dudley's inequality to subgaussian random processes that are not necessarily centered. In this case, the result gives a bound for the quantity

$$\mathbb{E} \sup\nolimits_{s,t \in \mathsf{T}} (X_s - X_t).$$

Frame appropriate hypotheses, and establish a version of Dudley's inequality without assuming that the random process $(X_t : t \in \mathsf{T})$ is centered.

Finally, the chaining method also works with somewhat weaker control on the increments. For example, we can also obtain bounds for random processes whose increments have subexponential tails.

Definition 12.9 (Subexponential process). Let $(\mathsf{T}, \mathrm{dist})$ be a metric space. Consider a centered random process $(X_t : t \in \mathsf{T})$ defined on the metric space. We say that the process is *subexponential* with respect to the metric dist if each increment satisfies

$$\log \mathbb{E}\, \mathrm{e}^{\theta(X_s - X_t)} \leq \mathrm{dist}(s,t) \cdot \theta^2 \quad \text{for all } s, t \in \mathsf{T} \text{ and } |\theta| < \theta_0.$$

Problem 12.10 (Dudley: Subexponential processes). Suppose that $(X_t : t \in \mathsf{T})$ is a centered, subexponential random process on a metric space $(\mathsf{T}, \mathrm{dist})$. Prove that

$$\mathbb{E} \sup\nolimits_{t \in \mathsf{T}} X_t \leq \mathrm{Const} \int_0^\infty \mathrm{d}\varepsilon \log \mathrm{N}(\mathsf{T}, \mathrm{dist}; \varepsilon).$$

Problem 12.11 (Dudley: Mixed tails). Another common scenario occurs for centered random processes that have Bernstein-type tails. In this case, we assume that the increments

are controlled by two different metrics.

$$\log \mathbb{E}\, \mathrm{e}^{\theta(X_s - X_t)} \le (\mathrm{dist}_1(s,t) + \mathrm{dist}_2^2(s,t)) \cdot \theta^2 \quad \text{for } s,t \in \mathsf{T} \text{ and } |\theta| < \theta_0.$$

Prove that

$$\mathbb{E} \sup_{t \in \mathsf{T}} X_t \le \text{Const} \int_0^\infty \mathrm{d}\varepsilon \left[\log \mathrm{N}(\mathsf{T}, \mathrm{dist}_1; \varepsilon) + \sqrt{\log \mathrm{N}(\mathsf{T}, \mathrm{dist}_2; \varepsilon)} \right].$$

12.5 Elementary examples

In this section, we develop a few simple examples of Dudley's inequality, as applied to canonical Gaussian processes. These results demonstrate how we can use upper bounds for covering numbers to obtain upper bounds for Gaussian processes. They also point toward some of the limitations of Dudley's inequality, which we seek to address in the next lecture.

For an index set $\mathsf{T} \subseteq \mathbb{R}^n$, we consider the centered canonical GP:

$$X_{\boldsymbol{t}} = \langle \boldsymbol{g}, \boldsymbol{t} \rangle \quad \text{where} \quad \boldsymbol{g} \sim \textsc{normal}(\mathbf{0}, \mathbf{I}_n).$$

We will consider several choices for the index set T.

12.5.1 The Euclidean unit ball

We begin with the simplest choice of index set: $\mathsf{T} = \mathsf{B}_2$, the Euclidean unit ball in $\mathbb{R}^n$. In this case,

$$\mathbb{E} \sup_{\boldsymbol{t} \in \mathsf{B}_2} X_{\boldsymbol{t}} = \mathbb{E} \, \|\boldsymbol{g}\|_{\ell_2}.$$

It is well-known that the norm of a standard normal vector satisfies the inequalities

$$\sqrt{n-1} \le \mathbb{E} \, \|\boldsymbol{g}\|_{\ell_2} \le \sqrt{n}.$$

In the last lecture, we used Sudakov's minoration to match the lower bound up to a constant. We now use Dudley's inequality to reproduce the upper bound.

Using the volumetric argument, we know that

$$\mathrm{N}(\mathsf{B}_2, \ell_2; \varepsilon) \le (1 + 2/\varepsilon)^n \le (3/\varepsilon)^n \quad \text{for all } \varepsilon > 0.$$

Of course, $\mathrm{diam}(\mathsf{B}_2, \ell_2) = 2$. Therefore, Dudley's inequality yields

$$\begin{aligned} \mathbb{E} \sup_{\boldsymbol{t} \in \mathsf{B}_2} X_{\boldsymbol{t}} &\le \text{Const} \int_0^2 \mathrm{d}\varepsilon \sqrt{\log \mathrm{N}(\mathsf{B}_2, \ell_2; \varepsilon)} \\ &\le \text{Const} \int_0^2 \mathrm{d}\varepsilon \sqrt{n \log(3/\varepsilon)} \le \text{Const} \cdot \sqrt{n}. \end{aligned}$$

Indeed, the singularity of the function $\varepsilon \mapsto \sqrt{\log(1/\varepsilon)}$ at $\varepsilon = 0$ is integrable. Roughly,

$$\int_0^a \mathrm{d}\varepsilon \sqrt{\log(1/\varepsilon)} \le \text{Const} \cdot a\sqrt{\log(1/a)}.$$

This is the correct answer.

12.5.2 Weighted orthonormal basis

Next, we consider a discrete index set:

$$\mathsf{T} = \left\{ \frac{\mathbf{e}_i}{\sqrt{\log(1+i)}} : i = 1, \dots, n \right\} \subset \mathbb{R}^n.$$

Using a more refined maximal inequality, one may verify that

$$\mathbb{E} \max_{t \in \mathsf{T}} X_t \approx \text{Const}.$$

Sudakov's minoration gives a matching lower bound because we can trivially cover T with one ℓ_2 ball of radius $\varepsilon = \sqrt{2}$.

On the other hand, Dudley's inequality gives the wrong answer for this simple example. Indeed,

$$\mathrm{N}(\mathsf{T}, \ell_2; \varepsilon) \geq 1 + i \quad \text{when} \quad \varepsilon \leq \frac{1}{\sqrt{\log(1+i)}}.$$

After some careful estimates, we may verify that the entropy integral satisfies

$$\int_0^\infty \mathrm{d}\varepsilon \sqrt{\log \mathrm{N}(\mathsf{T}, \ell_2; \varepsilon)} \geq \text{const} \cdot \log \log n.$$

This bound is not terrible, but it does not capture the correct behavior of the supremum.

12.5.3 Dudley versus Sudakov

With this last example in mind, one may wonder about the possible discrepancy between the lower bound from Sudakov's minoration and the upper bound from Dudley's inequality. The next statement gives an explicit estimate.

Proposition 12.12 (Two-sided Sudakov). Consider a canonical centered Gaussian process indexed by a set $\mathsf{T} \subseteq \mathbb{R}^n$. Then

$$\begin{aligned} \text{const} \cdot \sup_{\varepsilon > 0} \varepsilon \sqrt{\log \mathrm{N}(\mathsf{T}, \ell_2; \varepsilon)} &\leq \mathbb{E} \sup_{t \in \mathsf{T}} X_t \\ &\leq \text{Const} \cdot (\log n) \cdot \sup_{\varepsilon > 0} \varepsilon \sqrt{\log \mathrm{N}(\mathsf{T}, \ell_2; \varepsilon)}. \end{aligned}$$

Problem 12.13 (Two-sided Sudakov). Establish Proposition 12.12. **Hint:** Use the volumetric estimate to control the entropy integral for $\varepsilon \approx 0$.

In other words, for a canonical GP, the gap between Sudakov's minoration and Dudley's inequality cannot never be worse than the logarithm of the dimension of the GP. In fact, neither Sudakov nor Dudley captures the actual behavior of the supremum of a GP. In the next lecture, we will introduce a refinement of the chaining methodology which yields a full geometric characterization of the supremum of a GP.

13. Generic Chaining

Date: 16 Feburary 2020 **Scribe:** Roy Wang

Agenda:
1. Dudley reformulated
2. Generic chaining functional
3. Generic chaining theorem
4. Majorizing measure theorem
5. Talagrand's comparison

In this lecture, we introduce the generic chaining functional, which gives a geometric characterization of the supremum of a centered Gaussian process. We shall begin with intuitions on the derivation of the functional, and then present and prove the generic chaining theorem, which shows that the generic chaining functional gives an *upper bound* on the supremum of a subgaussian process. We demonstrate that this upper bound is qualitatively better than that of Dudley's theorem. Next, we state the majorizing measure theorem, which shows that the generic chaining functional gives a *lower bound* for the supremum of a Gaussian process. Finally, we give some implications of these theorems, including the Talagrand comparison and its application toward the estimation of the spectral norm of a random matrix with iid entries.

For positive numbers a, b, we shall use the notation $a \lesssim b$ to mean that there exists a positive universal constant such that $a \le \text{Const} \cdot b$. This constant does not depend on any parameters of the problem, but its exact value is unimportant.

13.1 Dudley and Sudakov, revisited

In this section, we recall the bounds that we have established for the suprema of Gaussian processes. The theorems of Sudakov and Dudley respectively give lower bounds and upper bounds of centered Gaussian processes. We encapsulate both these results in the next statement.

Proposition 13.1 (Sudakov + Dudley). Let $(X_t : t \in \mathsf{T})$ be a centered Gaussian process, endowed with canonical metric $d(s,t) = \|X_s - X_t\|_{\mathsf{L}_2}$. Then we have the following inequalities:

$$\sup_{\varepsilon>0} \varepsilon\sqrt{\log \mathrm{N}(\mathsf{T}, d; \varepsilon)} \quad \lesssim \quad \mathbb{E}\sup_{t\in\mathsf{T}} X_t \quad \lesssim \quad \int_0^\infty \sqrt{\log \mathrm{N}(\mathsf{T}, d; \varepsilon)}\,. \qquad (13.1)$$

As usual, $\mathrm{N}(\mathsf{T}, d; \varepsilon)$ is the covering number of (T, d) on the scale ε.

Several remarks are in order. Dudley's inequality, the *upper bound* in Proposition 13.1, extends to a centered *subgaussian* process. That is, for a metric space $(\mathsf{T}, \mathrm{dist})$, we assume that

$$\log \mathbb{E}\, \mathrm{e}^{\theta(X_s - X_t)} \le \mathrm{dist}^2(s,t) \cdot \theta^2/2 \quad \text{for all } s, t \in \mathsf{T} \text{ and all } \theta \in \mathbb{R}.$$

Of course, a GP is subgaussian with respect to its canonical metric. This extension follows from the same argument we used to establish the upper bound for a GP using the maximal inequality for subgaussian random variables. In contrast, Sudakov's minoration, the *lower bound* in Proposition 13.1, depends on special properties of GPs.

For the canonical GP on an index set $\mathsf{T} \subset \mathbb{R}^n$, it can be shown that the ratio between the estimates of Sudakov and Dudley does not exceed $\text{Const} \cdot \log n$. (See Exercise 2.b in Problem Set 3.) Therefore, a natural question arises:

Can we finding matching lower and upper bounds for the supremum of a centered Gaussian process in terms of the geometry of the metric space (T, d)?

Fortunately, the answer is *yes*! However, we need complexity measures that are more refined than the humble covering number. The results in this lecture originated in 1970s work of Fernique [Fer75]. In the 1990s, these results were revisited by Talagrand [Tal+96], who developed the generic chaining machinery described in this section. For more recent developments, see the 2016 papers by van Handel [Han18a; Han18b].

Aside: For the simple case of Gaussian processes, characterizing the supremum is already shockingly difficult. For other types of random processes, the situation can be even worse. For instance, it is also natural to study Rademacher processes. Let $\boldsymbol{\epsilon} \in \mathbb{R}^n$ have iid Rademacher entries, and consider the random process with members

$$X_{\boldsymbol{t}} = \sum_{i=1}^{n} \varepsilon_i t_i \quad \text{where } \boldsymbol{t} \in \mathsf{T} \subseteq \mathbb{R}^n.$$

Even for this basic example, the characterization of the supremum was only completed in 2013 by Bednorz and Latała [BL14].

13.2 Dudley: What went wrong?

Before we introduce the functional that characterizes the supremum of a GP, we first examine Dudley's inequality to see where we might have sacrificed the optimality of the bound. Recall that the proof constructs a multiscale family (T_i) of ε-nets, and we use the (sub)gaussian maximal inequality to obtain

$$\mathbb{E} \sup_{t \in \mathsf{T}} X_t \lesssim \sum_{i=k+1}^{\infty} \varepsilon_i \sqrt{\log |\mathsf{T}_i|}. \tag{13.2}$$

Here, $\varepsilon_i = 2^{-i}$ is the scale of the ith ε-net, and T_i is an ε_i-net with cardinality $\mathrm{N}(\mathsf{T}, d; \varepsilon_i)$. We choose k such that $|\mathsf{T}_k| = 1$, so the scales coarser than the kth scale make no contributions in the sum.

To understand where the bound is loose, it is helpful to reformulate the inequality. Instead of choosing a dyadic sequence of distance scales, instead we will select coverings whose log-cardinalities increase dyadically.

Definition 13.2 (Admissible sequence). A family $(\mathsf{T}_i : i \in \mathbb{Z}_+)$ of (nested) subsets of T is called admissible if $|\mathsf{T}_0| = 1$ and $|\mathsf{T}_i| < 2^{2^i}$ for $i \in \mathbb{N}$.

Note that $\sqrt{\log |T_i|} \le 2^{i/2}$.

Therefore, if we set $\varepsilon_i = \sup_{t \in \mathsf{T}} d(t, \mathsf{T}_i)$, then T_i forms an ε_i-net for the metric space (T, d). With this notation, we can rewrite the Dudley bound (13.2) using this set of scales:

$$\mathbb{E} \sup_{t \in \mathsf{T}} X_t \lesssim \sum_{i=0}^{\infty} 2^{i/2} \sup_{t \in \mathsf{T}} d(t, \mathsf{T}_i)\,. \tag{13.3}$$

The formula (13.3) highlights the shortcoming of Dudley's inequality. The supremum over points $t \in \mathsf{T}$ appears *inside the sum*, so we are always choosing the point that has the worst approximation by the net T_i at scale i. We might hope to obtain a similar bound with the supremum *outside the sum*, so that we are choosing the point that $t \in \mathsf{T}$ that has the worst overall approximation.

We can indeed develop a similar bound with the supremum outside the sum. As it happens, this simple modification addresses the fundamental shortcoming in the version (13.2) of Dudley's inequality.

13.3 Generic chaining functional

We are now prepared to present the definition of the generic chaining functional, which is a measure of the complexity of a metric space that is more refined than the covering numbers.

Definition 13.3 (Generic chaining functional). Let $(\mathsf{T}, \mathrm{dist})$ be a metric space, and let $(\mathsf{T}_i : i \in \mathbb{N} \cup \{-1\})$ be an admissible sequence. Then the (subgaussian) generic chaining functional is defined as

$$\gamma_2(\mathsf{T}, d) = \inf_{(\mathsf{T}_i) \text{ admissible}} \sup_{t \in \mathsf{T}} \sum_{i=0}^{\infty} 2^{i/2} d(t, \mathsf{T}_i) \,. \tag{13.4}$$

Modulo constants, the generic chaining functional is always in between the covering number bound in Sudakov's minoration and the entropy integral in Dudley's inequality. These claims follows more or less directly from the discussion in the last section.

Exercise 13.4 (Generic chaining versus entropy). Prove that

$$\sup_{\varepsilon > 0} \varepsilon \sqrt{\log \mathrm{N}(\mathsf{T}, \mathrm{dist}; \varepsilon)} \quad \lesssim \quad \gamma_2(\mathsf{T}, \mathrm{dist}) \quad \lesssim \quad \int_0^\infty \mathrm{d}\varepsilon \sqrt{\log \mathrm{N}(\mathsf{T}, \mathrm{dist}; \varepsilon)}.$$

In particular, Exercise 13.4 shows that the generic chaining functional results in a smaller estimate for the complexity of a metric space than the entropy integral. On the other hand, we have more tools for estimating covering numbers than for estimating γ_2. For instance, volumetric bounds, the empirical method, and Sudakov's minoration all give upper bounds for covering numbers. In contrast, it can be quite difficult to build good admissible sequences that realize the infimum in (13.4). As a consequence, it can be challenging to deploy the generic chaining functional in new settings. For some recent advances on bounding the generic chaining functional, see the papers [Han18a; Han18b].

13.4 Generic chaining theorem

We are now prepared to develop a bound on the supremum of a Gaussian process in terms of the generic chaining functional.

Theorem 13.5 (Generic chaining). Let $(X_t : t \in \mathsf{T})$ be a centered Gaussian process with canonical metric d. Then

$$\mathbb{E} \sup_{t \in \mathsf{T}} X_t \lesssim \gamma_2(\mathsf{T}, d). \tag{13.5}$$

As implied by Exercise 13.4, Theorem 13.5 yields a bound for the supremum that is at least as good as Dudley's. Here is an explicit example where Dudley's inequality is suboptimal, while the generic chaining theorem gives a correct bound.

Example 13.6 (Weighted orthogonal systems). Consider a canonical GP indexed by the set

$$\mathsf{T} = (\mathbf{e}_i / \sqrt{1 + \log i} : i = 1, 2, \cdots, n) \subset \mathbb{R}^n.$$

The expected supremum of this GP has constant order. Dudley's inequality only gives a upper bound of order $\log \log n$, while the generic chaining theorem implies the correct constant upper bound. ∎

Let us establish the generic chaining theorem. The proof is quite similar in spirit to the proof of Dudley's theorem.

Theorem 13.5. By approximation, we can assume that $|\mathsf{T}| < +\infty$. Consider an admissible sequence $(\mathsf{T}_i : i \in \mathbb{Z}_+)$ with $\mathsf{T}_0 = \{t_0\}$. Since T is a finite set, we can identify a scale $K \in \mathbb{Z}_+$ such that $\mathsf{T}_K = \mathsf{T}$.

As before, define the projection $\pi_i(t) := \arg\min\{d(s,t) : s \in \mathsf{T}_i\}$ for each $t \in \mathsf{T}$. We build the chain

$$t_0 = \pi_0(t) \to \pi_1(t) \to \pi_1(t) \to \cdots \to \pi_K(t) = t.$$

For each $t \in \mathsf{T}$, we have the chaining identity:

$$X_t - X_{t_0} = \sum\nolimits_{i=1}^{K} (X_{\pi_i(t)} - X_{\pi_{i-1}(t)}). \tag{13.6}$$

We need to bound the increments within the chain. With high probability,

$$|X_{\pi_i(t)} - X_{\pi_{i-1}(t)}| \le 2^{i/2} d(t, \mathsf{T}_i) \qquad \text{for all } t \in \mathsf{T} \text{ and } i \in \mathbb{N}.$$

More precisely, for a fixed point $t \in \mathsf{T}$ and scale $i \in \mathbb{N}$, we have the (sub)gaussian tail bound

$$\mathbb{P}\left\{|X_{\pi_i(t)} - X_{\pi_{i-1}(t)}| > u 2^{i/2} d(\pi_i(t), \pi_{i-1}(t))\right\} \le \exp(-u^2 2^i/2) \quad \text{for all } u > 0.$$

This result follows from the subgaussian maximal inequality because the increment is subgaussian with a variance proxy of $d^2(\pi_i(t), \pi_{i-1}(t))$.

Now, let us collect the probability across all of the possible pairs $(\pi_i(t), \pi_{i-1}(t)) \in \mathsf{T}_i \times T_{i-1}$ for $i \in \mathbb{N}$. At scale i, the number of pairs is at most

$$|\mathsf{T}_i||\mathsf{T}_{i-1}| \le |\mathsf{T}_i|^2 \le 2^{2^{i+1}}.$$

Applying the union bound, we arrive at

$$\begin{aligned}\mathbb{P}\left\{\exists i, \exists t : |X_{\pi_i(t)} - X_{\pi_{i-1}(t)}| > u 2^{i/2} d(\pi_i(t), \pi_{i-1}(t))\right\} \\ \le \sum\nolimits_{i=0}^{\infty} 2^{2^{i+1}} \exp(-u^2 2^i/2) \lesssim \exp(-\text{const} \cdot u^2).\end{aligned}$$

Indeed, for $u \ge 2$, the summand $2^{2^{i+1}} \exp(-u^2 2^i/2) \le (\mathrm{e}/2)^{-u^2 2^i/2}$, and we can adjust constants to make the bound hold for all $u > 0$. In other terms,

$$|X_{\pi_i(t)} - X_{\pi_{i-1}(t)}| \le u 2^{i/2} d(\pi_i(t), \pi_{i-1}(t)) \quad \text{for all } i, t, u$$

with high probability.

Returning to the chaining identity (13.6), we invoke the triangle inequality for the absolute value and the metric to arrive at

$$\begin{aligned}|X_t - X_{t_0}| &\le \sum\nolimits_{i=1}^{K} |X_{\pi_i(t)} - X_{\pi_{i-1}(t)}| \\ &\le u \sum\nolimits_{i=1}^{K} 2^{i/2} d(\pi_i(t), \pi_{i-1}(t)) \\ &\le u \sum\nolimits_{i=1}^{K} 2^{i/2} (d(\pi_i(t), t) + d(t, \pi_{i-1}(t))) \\ &\lesssim u \gamma_2(\mathsf{T}, d)\,.\end{aligned}$$

We deduce that

$$\mathbb{P}\left\{\sup\nolimits_{t \in \mathsf{T}} |X_t - X_{t_0}| \gtrsim u \gamma_2(\mathsf{T}, d)\right\} \lesssim \exp(-\text{const} \cdot u^2).$$

That is to say, $\sup_{t\in\mathsf{T}} |X_t - X_{t_0}|$ is subgaussian with variance proxy on the order of $\gamma_2(\mathsf{T}, d)$. We conclude that

$$\mathbb{E}\sup_{t\in\mathsf{T}} X_t = \mathbb{E}\sup_{t\in\mathsf{T}} (X_t - X_{t_0}) \lesssim \gamma_2(\mathsf{T}, d).$$

This is what we needed to show. ∎

As with Dudley's inequality, the upper bound from Theorem 13.5 is also valid for subgaussian random processes.

Exercise 13.7 (Generic chaining: Subgaussian processes). Let $(X_t : t \in \mathsf{T})$ be a subgaussian process on the metric space $(\mathsf{T}, \mathrm{dist})$. Prove that

$$\mathbb{E}\sup_{t\in\mathsf{T}} X_t \lesssim \gamma_2(\mathsf{T}, \mathrm{dist}).$$

13.5 Majorizing measure theorem

As we have indicated, it can be difficult to calculate the mysterious quantity γ_2. So why should we bother? The reason is that γ_2 gives a full geometric characterization of the supremum of a Gaussian process. This result, called the majorizing measure theorem, is essentially due to Fernique; the formulation in terms of the generic chaining functional was developed by Talagrand.

Theorem 13.8 (Majorizing measure). Let $(X_t : t \in \mathsf{T})$ be a centered Gaussian process with canonical metric d. Then

$$\gamma_2(\mathsf{T}, d) \quad \lesssim \quad \mathbb{E}\sup_{t\in\mathsf{T}} X_t \quad \lesssim \quad \gamma_2(\mathsf{T}, d). \tag{13.7}$$

We will prove Theorem 13.8 in Lecture 14.

Even though it may be hard to calculate the geometric quantity γ_2, the majorizing measure theorem still has some interesting applications. The next result gives one powerful consequence, which is known as Talagrand's subgaussian comparison principle.

Corollary 13.9 (Talagrand's subgaussian comparison). Let $(Y_t : t \in \mathsf{T})$ be a centered Gaussian process on the index set T with canonical metric d. Suppose that $(X_t : t \in \mathsf{T})$ is a centered subgaussian process on the metric space (T, d). Then

$$\mathbb{E}\sup_{t\in\mathsf{T}} X_t \lesssim \mathbb{E}\sup_{t\in\mathsf{T}} Y_t.$$

Proof. We exploit the fact that the generic chaining functional provides an *upper bound* for any subgaussian process, while it provides a *lower bound* for a Gaussian process. Indeed,

$$\mathbb{E}\sup_{t\in\mathsf{T}} X_t \lesssim \gamma_2(\mathsf{T}, d) \lesssim \mathbb{E}\sup_{t\in\mathsf{T}} Y_t.$$

The first inequality is Exercise 13.7, and the second is Theorem 13.8. ∎

Corollary 13.9 is a valuable tool because we can reduce the problem of bounding the supremum of a general subgaussian process above to the easier problem of bounding the supremum of a Gaussian process. We have many additional tools for studying GPs, including the comparison theorems of Slepian, Chevet, and Sudakov–Fernique. The next exercise describes a problem that can be addressed by this approach.

Exercise 13.10 (Subgaussian Chevet). Suppose $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ is a random matrix with iid 1-subgaussian entries. Using Theorem 13.8 and Chevet's theorem, prove that

$$\mathbb{E}\,\|\boldsymbol{X}\|_{\ell_2 \to \ell_2} \lesssim \sqrt{m} + \sqrt{n}.$$

As usual, $\|\cdot\|_{\ell_2 \to \ell_2}$ is the ℓ_2 operator norm. More generally, if $\boldsymbol{S}$ and $\boldsymbol{T}$ are fixed conformal matrices, prove that

$$\mathbb{E}\,\|\boldsymbol{S}\boldsymbol{X}\boldsymbol{T}\|_{\ell_2 \to \ell_2} \lesssim \|\boldsymbol{S}\|_{\ell_2 \to \ell_2}\|\boldsymbol{T}\|_{\mathrm{F}} + \|\boldsymbol{S}\|_{\mathrm{F}}\|\boldsymbol{T}\|_{\ell_2 \to \ell_2}.$$

Here, $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm.

Lecture bibliography

[BL14] W. Bednorz and R. Latała. "On the boundedness of Bernoulli processes". In: *Annals of Mathematics* 180.3 (2014), pages 1167–1203.

[Fer75] X. Fernique. "Regularité des trajectoires des fonctions aléatoires gaussiennes". In: *Ecole d'Eté de Probabilités de Saint-Flour IV—1974*. Springer Berlin Heidelberg, 1975, pages 1–96.

[Han18a] R. van Handel. "Chaining, interpolation, and convexity". In: *J. Eur. Math. Soc. (JEMS)* 20.10 (2018), pages 2413–2435. DOI: 10.4171/JEMS/815.

[Han18b] R. van Handel. "Chaining, interpolation and convexity II: The contraction principle". In: *Ann. Probab.* 46.3 (2018), pages 1764–1805. DOI: 10.1214/17-AOP1214.

[Tal+96] M. Talagrand et al. "Majorizing measures: the generic chaining". In: *The Annals of Probability* 24.3 (1996), pages 1049–1103.

14. Majorizing Measure Theorem

Date: 18 February 2021 **Scribe:** Yifan Chen

Agenda:
1. Gaussian width
2. Generic chaining functional
3. Majorizing measure theorem
4. Growth functional
5. Contraction
6. Admissible sequence

The majorizing measure theorem is the deepest result about the suprema of Gaussian processes. It shows that the supremum of a Gaussian process admits an alternative characterization in terms of the geometry of the metric space induced by the canonical metric of the process. This lecture provides a proof of this result, developed by van Handel in 2016 [Han18b].

> We will frequently use the notations $\lesssim$ and $\gtrsim$ in this lecture. These relations suppress a universal constant that does not depend on anything else.

14.1 Gaussian width

For a set $\mathsf{T} \subset \mathbb{R}^d$, we may construct a canonical centered Gaussian process: $(X_{\boldsymbol{t}} : \boldsymbol{t} \in \mathsf{T})$. The elements of this process are defined by

$$X_{\boldsymbol{t}} = \langle \boldsymbol{g}, \boldsymbol{t} \rangle \quad \text{where } \boldsymbol{g} \sim \textsc{normal}(\mathbf{0}, \mathbf{I}_d).$$

Each random variable is zero mean, and in general, they have a nontrivial covariance structure determined by the set T. We emphasize that the random vector $\boldsymbol{g}$ is the unique source of randomness in the process.

We are interested in the expected supremum of the canonical Gaussian process indexed by T. The expected supremum depends only on the index set T, and it is a fundamental measure of the size of the set. Let us remind the reader of this important definition, which already appeared in Lecture 11.

Definition 14.1 (Gaussian width). The *Gaussian width* of T is defined as

$$w(\mathsf{T}) := \mathbb{E} \sup_{\boldsymbol{t} \in \mathsf{T}} X_{\boldsymbol{t}} = \mathbb{E} \sup_{\boldsymbol{t} \in \mathsf{T}} \langle \boldsymbol{g}, \boldsymbol{t} \rangle.$$

The Gaussian width has a number of valuable properties, which we enumerate for future reference.

1. **Invariance.** The functional w is translation and rotation invariant.
2. **Lower bound.** The translation invariance property implies that $w(\mathsf{T}) \geq 0$ since we can always translate the set T to contain the origin. A fortiori,
$$w(\mathsf{T}) \gtrsim \operatorname{diam}(\mathsf{T}).$$
Recall that $\operatorname{diam}(\mathsf{T}) := \sup\{\|\boldsymbol{s} - \boldsymbol{t}\|_2 : \boldsymbol{s}, \boldsymbol{t} \in \mathsf{T}\}$.
3. **Monotonicity.** The Gaussian width is an increasing set function. If $\mathsf{S} \subset \mathsf{T}$, then $w(\mathsf{S}) \leq w(\mathsf{T})$.

Exercise 14.2 (Gaussian width: Properties). Verify the properties of the Gaussian width listed above.

14.2 Generic chanining functional

Next, we introduce the generic chaining functional. The construction today is slightly different from the previous lecture. This change of notation helps us organize the proof better. Instead of approximating T with nets, we approximate it with set partitions (think of Voronoi cells). To that end, let us reformulate the definition of an admissible sequence in terms of partitions.

Definition 14.3 (Admissible sequence). An *admissible sequence* for T is a sequence $(\mathscr{A}_n)$ of nested partitions of T with $|\mathscr{A}_n| < 2^{2^n}$ for $n \in \mathbb{Z}_+$.

Example 14.4 (Admissible sequence). Let us give an illustration of an admissible sequence (Figure 14.1). Suppose the index set T is an ellipsoid. Then $\mathscr{A}_0$ contains a single element, namely T. The cardinality of $\mathscr{A}_1$ is less than $2^{2^1} = 4$, so it chops the set T into three pieces. The partition $\mathscr{A}_1$ is the collection of these pieces. A similar construction can be applied to $\mathscr{A}_2$ and beyond. ▪

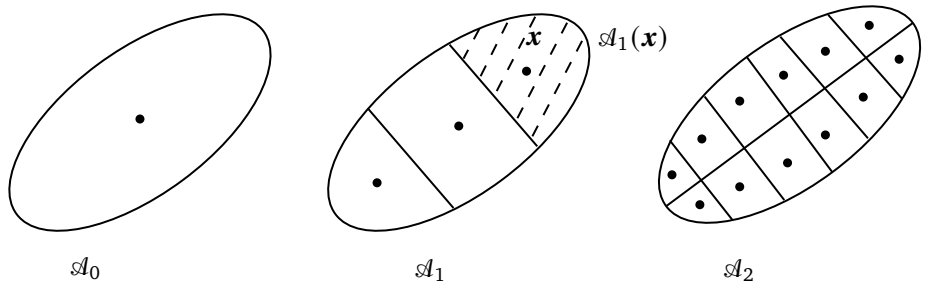


Figure 14.1 (An admissible sequence). The first three partitions $\mathscr{A}_0$, $\mathscr{A}_1$, $\mathscr{A}_2$ in an admissible sequence for the ellipsoid. The second panel illustrates the function $\boldsymbol{x} \mapsto \mathscr{A}_1(\boldsymbol{x})$, which returns the element of the partition $\mathscr{A}_1$ that contains $\boldsymbol{x}$.

Next, let us define a function that connects points in T and partitions in the admissible sequence.

Definition 14.5 (Neighborhood function). For any $\boldsymbol{x} \in \mathsf{T}$, the function $\mathscr{A}_n(\boldsymbol{x})$ returns the unique set $\mathsf{A} \in \mathscr{A}_n$ that contains the point $\boldsymbol{x}$.

As an example, in the middle of Figure 14.1, we have shown a point $\boldsymbol{x}$ and the corresponding neighborhood $\mathscr{A}_1(\boldsymbol{x})$, the element of the partition marked with dotted lines.

Using these constructions, we are ready to define the generic chaining functional.

Definition 14.6 (Generic chaining functional). The generic chaining functional is defined as

$$\gamma_2(\mathsf{T}) := \inf_{(\mathscr{A}_n) \text{ admissible}} \sup_{\boldsymbol{x} \in \mathsf{T}} \sum_{n \geq 0} 2^{n/2} \operatorname{diam}(\mathscr{A}_n(\boldsymbol{x})).$$

Exercise 14.7 (Generic chaining functional). Show that Definition 14.6 is equivalent to the definition in the last lecture modulo a universal constant.

The majorizing measure theorem states that the Gaussian width $w(\mathsf{T})$ is bounded *below* by the generic chaining functional $\gamma_2(\mathsf{T})$.

Theorem 14.8 (Majorizing measure). For a compact set $\mathsf{T} \subset \mathbb{R}^d$, it holds that

$$w(\mathsf{T}) \gtrsim \gamma_2(\mathsf{T}).$$

The reverse inequality $w(\mathsf{T}) \lesssim \gamma_2(\mathsf{T})$ is the generic chaining bound, which we proved in Lecture 13. The proof of the upper bound is straightforward; it is based on a simple reorganization of the proof of Dudley's inequality. In contrast, the lower bound of $w(\mathsf{T})$ stated in Theorem 14.8 is a much deeper result. A version of this result was first obtained by Fernique in the 1970s; in the 1990s, Talagrand obtained a reformulation in terms of the generic chaining functional. The proof here was recently proposed by van Handel [Han18b].

The proof contains three main steps. First, we introduce a growth functional that controls the local complexity of the Gaussian process (Section 14.3). Second, we establish a contraction lemma, which shows how to bound the entropy numbers of a set using the growth functional (Section 14.4). Last, we show how to construct an admissible sequence whose elements are bounded in terms of the growth functional (Section 14.5). Putting these pieces together, we obtain the result.

14.3 Growth functional

The first step in the argument is to introduce a *growth functional*, which describes the local complexity of the canonical GP indexed by the set T. This function reflects the supremum of the process over Euclidean balls centered at points in the set.

To begin, we introduce the truncated Euclidean balls

$$\mathsf{B}(\boldsymbol{x}, s) := \{\boldsymbol{y} \in \mathsf{T} : \|\boldsymbol{y} - \boldsymbol{x}\|_2 \le s\} \subseteq \mathsf{T}. \tag{14.1}$$

This truncated ball reflects the structure the set T at the point $\boldsymbol{x}$ on the scale s. The Gaussian width $w(\mathsf{B}(\boldsymbol{x}, s))$ of the truncated ball measures the contribution to the supremum of the GP due to points near $\boldsymbol{x}$ on the scale s. The function $s \mapsto w(\mathsf{B}(\boldsymbol{s}, s))$ is increasing, and it is bounded above by $w(\mathsf{T})$. See Figure 14.2 for illustrations.

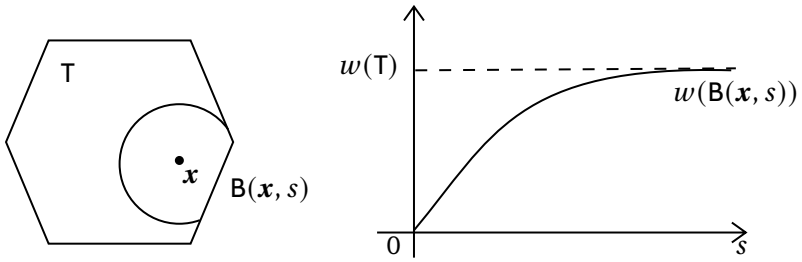


Figure 14.2 Illustration of $\mathsf{B}(\boldsymbol{x}, s)$ and $w(\mathsf{B}(\boldsymbol{x}, s))$.

What is important is the scale s at which the width $w(\mathsf{B}(\boldsymbol{x}, s))$ is increasing at a specified rate r. We can capture this effect by means of an interpolation function.

Definition 14.9 (Interpolation function). For each $\boldsymbol{x} \in \mathsf{T}$ and $r > 0$, the *interpolation function* is defined as

$$K(\boldsymbol{x}, r) := \inf_{s>0} (rs - w(\mathsf{B}(\boldsymbol{x}, s))) . \tag{14.2}$$

Aside: The terminology stems from an interpretation of K as the K-functional that appears in real interpolation theory [Han18a, Remark 6.2]. It can be viewed as a smoothing of the Gaussian width of local balls.

In the definition of the interpolation function, the argument of the infimum is a lower semi-continuous function, so the minimizer always exists. The minimizer itself is the quantity of primary interest.

Definition 14.10 (Growth functional). The *growth functional* $s(\boldsymbol{x}, r)$ is defined as the

(smallest) minimizer of the optimization problem (14.2). In other terms,

$$K(\boldsymbol{x}, r) =: rs(\boldsymbol{x}, r) - w(\mathsf{B}(\boldsymbol{x}, s(\boldsymbol{x}, r))).$$

The formula (14.2) may be familiar to you as the Legendre transform of the concave function $s \mapsto w(\mathsf{B}(\boldsymbol{x}, s))$. When this function is differentiable,

$$\frac{\mathrm{d}}{\mathrm{d}s} w(\mathsf{B}(\boldsymbol{x}, s))\Big|_{s=s(\boldsymbol{x},r)} = r.$$

Therefore, $s(\boldsymbol{x}, r)$ is the scale at which r is the growth rate of the supremum of the GP on a local ball centered at $\boldsymbol{x}$.

The interpolation function $K(\boldsymbol{x}, r)$ has several properties that we will use in the argument:

1. **Limits.** Asymptotically, $\lim_{r\to\infty} K(\boldsymbol{x}, r) = 0$.
2. **Lower bound.** It holds that $K(\boldsymbol{x}, r) \geq -w(\mathsf{T})$.
3. **Increments.** The interpolation function changes at a rate determined by the growth functional:

   $$K(\boldsymbol{x}, r) - K(\boldsymbol{x}, \alpha r) \geq (1-\alpha) rs(\boldsymbol{x}, r) \quad \text{for any } \alpha \leq 1.$$

   This property can be verified as follows:

   $$\begin{aligned} K(\boldsymbol{x}, r) - K(\boldsymbol{x}, \alpha r) &= rs(\boldsymbol{x}, r) - w(\mathsf{B}(\boldsymbol{x}, s(\boldsymbol{x}, r))) - \inf_{s>0} (\alpha r s - w(\mathsf{B}(\boldsymbol{x}, s))) \\ &\geq (1-\alpha) rs(\boldsymbol{x}, r). \end{aligned}$$

   We have used the definition of infimum and the fact that $s(\boldsymbol{x}, r)$ is feasible for the optimization problem.
4. **Monotonicity.** The function $r \mapsto K(\boldsymbol{x}, r)$ is increasing.

Using these facts, we may prove an interpolation lemma. This result allows us to bound the Gaussian width $w(\mathsf{T})$ below by a multiscale sum involving the growth function.

Lemma 14.11 (Interpolation). For each $\alpha \in (0, 1)$, it holds that

$$\sup_{\boldsymbol{x}\in\mathsf{T}} \sum\nolimits_{n\geq 0} 2^{n/2} s(\boldsymbol{x}, \alpha 2^{n/2}) \lesssim \frac{w(\mathsf{T})}{\alpha}.$$

Proof. Fix an arbitrary point $\boldsymbol{x} \in \mathsf{T}$. We calculate that

$$\begin{aligned} w(\mathsf{T}) &\geq \lim_{r\to\infty} K(\boldsymbol{x}, r) - K(\boldsymbol{x}, \alpha/2) \\ &= \sum\nolimits_{n\geq 0} (K(\boldsymbol{x}, \alpha 2^{n/2}) - K(\boldsymbol{x}, \alpha 2^{(n-1)/2})) \\ &\gtrsim \sum\nolimits_{n\geq 0} \alpha 2^{n/2} s(\boldsymbol{x}, \alpha 2^{n/2}). \end{aligned} \tag{14.3}$$

We have used the first three properties of the interpolation function in sequence. Last, take the supremum over $\boldsymbol{x} \in \mathsf{T}$. ∎

To reiterate, we can obtain a lower bound for the Gaussian width $w(\mathsf{T})$ in terms of the growth functional. Observe that the left-hand side of the bound resembles the form of the generic chaining functional. The question then is how to use the growth functional to control the size of the pieces of an (optimal) admissible sequence.

14.4 Contraction

In this section, we will prove a result which shows that the growth function provides a bound on how accurately we can quantize a set. In the next section, we will use this fact to convert an optimal admissible sequence into an admissible sequence that is controlled by the growth functions.

First, we introduce a classic quantity from approximation theory.

Definition 14.12 (Entropy number). The *entropy number* of a set $\mathsf{A} \subset \mathbb{R}^d$ is defined as

$$e_n(\mathsf{A}) := \inf_{|\hat{\mathsf{A}}|<2^{2^n}} \sup_{\boldsymbol{x}\in\mathsf{A}} \operatorname{dist}(\boldsymbol{x}, \hat{\mathsf{A}}).$$

In this expression, the set $\hat{\mathsf{A}} \subseteq \mathbb{R}^n$, but it is not necessarily contained in A.

The entropy number $e_n(\mathsf{A})$ is the error in the best quantization of the set A using 2^{2^n} points. These points can be labeled using 2^n bits. Let us emphasize that the entropy number should be regarded as a *distance*, the scale on which we can approximate the set A using 2^n bits.

The next lemma that connects the entropy number with the growth functional.

Lemma 14.13 (Contraction). For $n \geq 0$ and $\mathsf{A} \subseteq \mathsf{T}$ and $\alpha \in (0, 1)$, it holds that

$$e_n(\mathsf{A}) \lesssim \alpha \operatorname{diam}(\mathsf{A}) + \sup_{\boldsymbol{x}\in\mathsf{A}} s(\boldsymbol{x}, \alpha 2^{n/2}).$$

Before proving this result, let us discuss the interpretation. The quantity $\operatorname{diam}(\mathsf{A})$ is the worst possible error one could suffer when quantizing the set A. The first term on the right-hand side is *smaller* than this worst-case error by a factor of α. To achieve this reduction, we also have to pay for the maximum value of the growth functional $s(\boldsymbol{x}, \alpha 2^{n/2})$ over the set. This quantity reflects the length scale at which local balls are growing at the rate $\alpha 2^{n/2}$. The parameter α negotiates a tradeoff between these two effects.

In the next section, we will use Lemma 14.13 to partition sets into smaller pieces whose diameters are controlled by the growth functional. Since the growth functional bounds the width from below, this procedure will help us prove the majorizing measure theorem.

Before moving to that stage, let us establish the contraction lemma. We require another minimum principle for GPs, which is drawn from [Tal14, Prop 2.4.9].

Proposition 14.14 (Super-Sudakov). Let $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \subseteq \mathsf{T}$ be an ε-separated set in T. For all $\sigma > 0$, it holds that

$$w\big(\textstyle\bigcup_{i\leq N} \mathsf{B}(\boldsymbol{x}_i, \sigma)\big) - \min_{i\leq N} w(\mathsf{B}(\boldsymbol{x}_i, \sigma)) \geq \text{const} \cdot (\varepsilon - \sigma)\sqrt{\log N}.$$

For comparison, the simplest form of Sudakov's minoration only has the initial term on each side. Super-Sudakov enhances this bound. We can reduce the left-hand side by the Gaussian width of the smallest truncated ball $\mathsf{B}(\boldsymbol{x}_i, \sigma)$, provided that we also decrease the right-hand side by $\sigma\sqrt{\log N}$.

Exercise 14.15 (Super-Sudakov). Prove Proposition 14.14. **Hint:** Consider the random variables $Y_i = \sup_{\boldsymbol{t}\in\mathsf{B}(\boldsymbol{x}_i,\sigma)} X_{\boldsymbol{t}} - X_{\boldsymbol{x}_i}$.

With Proposition 14.14 at hand, we may establish the contraction lemma.

Proof of Lemma 14.13. Fix $n \geq 0$. By definition of the entropy number $e_n(\mathsf{A})$, we may quantize the set A by means of an $(e_n(\mathsf{A})/2)$-separated set $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \subset \mathsf{A}$ with $N = 2^{2^n}$. A simple proof of this fact can be found in [Han18b, Lemma 2.2].

Define the quantities

$$\sigma := \sup_{\boldsymbol{x}\in\mathsf{A}} s(\boldsymbol{x}, \alpha 2^{n/2}) \quad \text{and} \quad r := \operatorname{diam}(\mathsf{A}) + \sigma.$$

For each $i \in \{1, 2, \ldots, N\}$, using the monotonicity of the Gaussian width, we find that

$$\begin{aligned} -w(\mathsf{B}(\boldsymbol{x}_i, \sigma)) &\le -w(\mathsf{B}(\boldsymbol{x}_i, s(\boldsymbol{x}_i, \alpha 2^{n/2}))) \\ &\le K(\boldsymbol{x}_i, \alpha 2^{n/2}) \\ &\le r\alpha 2^{n/2} - w(\mathsf{B}(\boldsymbol{x}_i, r)) \\ &\le r\alpha 2^{n/2} - w(\textstyle\bigcup_{i\le N} \mathsf{B}(\boldsymbol{x}_i, \sigma)). \end{aligned}$$

In the second and third inequalities, we used the definition of the interpolation function and the growth functional. The last inequality exploits the fact that $\bigcup_{i\le N} \mathsf{B}(\boldsymbol{x}_i, \sigma) \subset \mathsf{B}(\boldsymbol{x}_i, r)$ beause $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subset \mathsf{A}$ and $r = \operatorname{diam}(\mathsf{A}) + \sigma$.

Choose the index $i \in \{1, 2, \ldots, N\}$ for which $w(\mathsf{B}(\boldsymbol{x}_i, r))$ is minimized. Rearrange the last display, and invoke Proposition 14.14:

$$\begin{aligned} 0 &\ge w(\textstyle\bigcup_{i\le N} \mathsf{B}(\boldsymbol{x}_i, \sigma)) - \min_{i\le N} w(\mathsf{B}(\boldsymbol{x}_i, r)) - r\alpha 2^{n/2} \\ &\ge \text{const} \cdot ((e_n(\mathsf{A})/2) - \sigma) \cdot \sqrt{\log 2^{2^n}} - r\alpha 2^{n/2}. \end{aligned}$$

Rearranging the above inequality and recalling the definitions of σ and r,

$$e_n(\mathsf{A}) \lesssim \alpha \operatorname{diam}(\mathsf{A}) + \sup_{\boldsymbol{x}\in\mathsf{A}} s(\boldsymbol{x}, \alpha 2^{n/2}).$$

This is the desired result. ∎

14.5 Admissible sequence

The last step of proving the majorizing measure theorem is to perform some surgery on an admissible sequence to obtain a new admissible sequence whose pieces have diameters controlled by the local values of the growth functional. More precisely, we will take an optimal sequence for which the minimum in $\gamma_2(\mathsf{T})$ is achieved. Then we will modify the sequence so that the diameter of each piece $\mathscr{A}_n(\boldsymbol{x})$ in each partition is controlled by the growth functional $s(\boldsymbol{x}, r)$. To complete the proof, we invoke the comparison between the growth functional and the width $w(\mathsf{T})$ from Lemma 14.11.

Proof of Theorem 14.8. Let $(\mathscr{A}_n)$ be an optimal admissible sequence such that the infimum in $\gamma_2(\mathsf{T})$ is attained. Fix the scale $n \in \mathbb{Z}_+$. We will subdivide each set $\mathsf{A}_i \in \mathscr{A}_n$ into n disjoint pieces such that

For simplicity, we assume that the infimum is achieved. If not, we can use a routine approximation argument.

$$\begin{aligned} (j < n): &\quad \mathsf{A}_i^j = \{\boldsymbol{x} \in \mathsf{A}_i : 2^{-i}\operatorname{diam}(\mathsf{T}) < s(\boldsymbol{x}, \alpha 2^{n/2}) \le 2^{-i+1}\operatorname{diam}(\mathsf{T})\} \\ (j = n): &\quad \mathsf{A}_i^j = \{\boldsymbol{x} \in \mathsf{A}_i : s(\boldsymbol{x}, \alpha 2^{n/2}) \le 2^{-n+1}\operatorname{diam}(\mathsf{T})\}. \end{aligned}$$

By construction, for each set A_i, each index $j \le n$ and each point $\boldsymbol{x} \in \mathsf{A}_i^j$, it holds that

$$\sup_{\boldsymbol{y}\in\mathsf{A}_i^j} s(\boldsymbol{y}, \alpha 2^{n/2}) \le 2s(\boldsymbol{x}, \alpha 2^{n/2}) + 2^{-n+1}\operatorname{diam}(\mathsf{T}).$$

This formula means that, in each set A_i^j, the maximum value of the growth functional is controlled by the local value of the growth functional, plus a tiny quantity that depends on the diameter of T.

Now, by the above property and the contraction result (Lemma 14.13), we can further partition A_i^j into fewer than 2^{2^n} pieces A_i^{jk}, each with the property that

$$\operatorname{diam}(\mathsf{A}_i^{jk}) \lesssim \alpha \operatorname{diam}(\mathsf{A}_i) + s(\boldsymbol{x}, \alpha 2^{n/2}) + 2^{-n}\operatorname{diam}(\mathsf{T}).$$

The latter bound holds uniformly for all $\boldsymbol{x} \in \mathsf{A}_i^{jk}$.

We may now reassemble these pieces to make a new admissible sequence. Concretely, define

$$\begin{aligned}\mathscr{C}_0 = \mathscr{C}_1 = \mathscr{C}_2 &= \{\mathsf{T}\};\\ \mathscr{C}_{n+3} &= \{\mathsf{A}_i^{jk} \in \mathscr{A}_n : 1 \le i, k < 2^{2^n}, 1 \le j \le n\}.\end{aligned}$$

We can easily verify that $|\mathscr{C}_n| < 2^{2^n}$ so $(\mathscr{C}_n)$ is also an admissible sequence.

This admissible sequence allows us to bound the generic chaining functional in terms of itself. Indeed,

$$\begin{aligned}\gamma_2(\mathsf{T}) &\le \sup_{\boldsymbol{x}\in\mathsf{T}} \sum\nolimits_{n\ge 0} 2^{n/2}\operatorname{diam}(\mathscr{C}_n(\boldsymbol{x}))\\ &\lesssim \sup_{\boldsymbol{x}\in\mathsf{T}} \sum\nolimits_{n\ge 0} 2^{n/2}[\alpha\operatorname{diam}(\mathscr{A}_n(\boldsymbol{x})) + s(\boldsymbol{x}, \alpha 2^{n/2}) + 2^{-n}\operatorname{diam}(\mathsf{T})]\\ &=: \text{(i)} + \text{(ii)} + \text{(iii)}.\end{aligned}$$

Let us examine each of the three terms on the right-hand side. First,

$$\text{(i)} := \alpha \cdot \sup_{\boldsymbol{x}\in\mathsf{T}} \sum\nolimits_{n\ge 0} 2^{n/2}\operatorname{diam}(\mathscr{A}_n(\boldsymbol{x})) = \alpha\gamma_2(T)$$

because γ_2 is attained for the admissible sequence $(\mathscr{A}_n)$. Second, according to Lemma 14.11,

$$\text{(ii)} := \sup_{\boldsymbol{x}\in\mathsf{T}} \sum\nolimits_{n\ge 0} 2^{n/2} s(\boldsymbol{x}, \alpha 2^{n/2}) \lesssim \frac{w(\mathsf{T})}{\alpha}.$$

Third, since the Gaussian width is larger than the diameter,

$$\text{(iii)} := \sup_{\boldsymbol{x}\in\mathsf{T}} \sum\nolimits_{n\ge 0} 2^{n/2}\cdot 2^{-n}\operatorname{diam}(\mathsf{T}) \lesssim \operatorname{diam}(\mathsf{T}) \lesssim w(\mathsf{T}).$$

Combining these displays, we arrive at the bound

$$\gamma_2(\mathsf{T}) \lesssim \alpha\gamma_2(\mathsf{T}) + (1+\alpha^{-1})w(\mathsf{T}).$$

Choose a sufficiently small positive number α to complete the argument. ∎

Lecture bibliography

[Han18a] R. van Handel. "Chaining, interpolation, and convexity". In: *J. Eur. Math. Soc. (JEMS)* 20.10 (2018), pages 2413–2435. DOI: 10.4171/JEMS/815.

[Han18b] R. van Handel. "Chaining, interpolation and convexity II: The contraction principle". In: *Ann. Probab.* 46.3 (2018), pages 1764–1805. DOI: 10.1214/17-AOP1214.

[Tal14] M. Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*. Springer Science & Business Media, 2014.

III. empirical processes

15. Uniform Law of Large Numbers

Date: 23 February 2020 **Scribe:** Kevin Yu

In the previous section of the course, we investigated the suprema of random processes in a somewhat abstract manner. We covered Gaussian comparison theorems, and we established the Sudakov minoration for Gaussian processes and Dudley's chaining inequality for subgaussian processes, both of which are suboptimal. We introduced the generic chaining method and obtained a geometric characterization of the supremum of a Gaussian process using the generic chaining functional.

In the next section of the course, we turn our attention to *empirical processes*, an important class of random processes that arise in statistics and statistical learning. These processes capture the properties of random elements sampled from a population. Dudley's chaining inequality and covering number bounds will serve as key techniques for controlling the behavior of empirical processes.

Agenda:

1. Monte Carlo integration
2. The uniform LLN
3. Covering Lipschitz functions
4. Empirical Processes

15.1 The uniform law of large numbers

Before we give the formal definition of an empirical process, we begin with an illustrative application to Monte Carlo integration. As you know, we can approximate (high-dimensional) integrals by evaluating the integrand at random points. We will pursue the question about whether we can approximate a family of integrals in a similar manner.

15.1.1 Monte Carlo Integration

First, we review Monte Carlo integration. Suppose that $\Omega \subseteq \mathbb{R}^d$ is a domain, that μ is a probability measure supported on Ω, and that $f : \Omega \mapsto \mathbb{R}$ is a measurable function on the domain. Our goal is to approximate the integral

$$\int_\Omega f \, \mathrm{d}\mu.$$

This problem may be challenging when the integrand f is very irregular or when the domain Ω is very high dimensional.

The idea behind Monte Carlo integration is to approximate the integral by evaluating the integrand at a small collection of random points in the domain. Consider a sample $(X_1, \ldots, X_n)$ of points drawn from the measure: $X_i \sim \mu$ iid. We approximate the integral by an empirical average:

$$\int_\Omega f \, \mathrm{d}\mu \approx \frac{1}{n} \sum_{i=1}^n f(X_i). \tag{15.1}$$

Figure 15.1 illustrates the procedure. Observe that we tend to draw more sample points from regions where the measure μ places more mass. On the other hand, the sample points are drawn without reference to the integrand f.

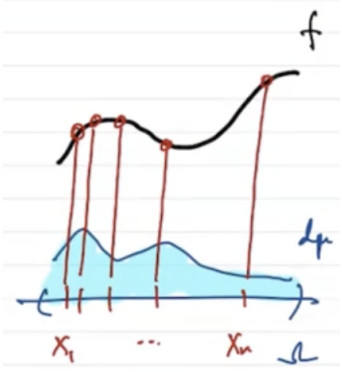


Figure 15.1 A sample $(X_1, \ldots, X_n)$ from a measure μ allows us to approximate the integral of a function f against the measure.

Why is the estimate in equation (15.1) reasonable? Assuming that $f \in \mathsf{L}_1(\mu)$, we can express the integral as an expectation:

$$\mathbb{E}\, f(X) = \int_\Omega f \,\mathrm{d}\mu \quad \text{where } X \sim \mu.$$

As a result, the Monte Carlo estimator provides an *unbiased* estimate for the integral:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n f(X_i)\right] = \int_\Omega f \,\mathrm{d}\mu.$$

Thus, the Monte Carlo estimator is accurate on average.

Now many samples we need to draw from μ to obtain an approximation of a given quality? Assuming now that $f \in \mathsf{L}_2(\mu)$, we may compute the variance of the estimator:

$$\operatorname{Var}\left[\frac{1}{n}\sum_{i=1}^n f(X_i)\right] = \frac{1}{n}\left[\int_\Omega f^2 \,\mathrm{d}\mu - \left(\int_\Omega f \,\mathrm{d}\mu\right)^2\right].$$

We can see that the variance decreases in proportion to $n^{-1}\operatorname{Var}_\mu(f)$. In other words, to achieve an approximation that satisfies

$$\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \int_\Omega f \,\mathrm{d}\mu\right| \le \varepsilon\sqrt{\operatorname{Var}_\mu(f)} \quad \text{with probability at least } 2/3,$$

it suffices to draw a sample with size $n = O(\varepsilon^{-2})$. Furthermore, by the strong law of large numbers (SLLN), the Monte Carlo estimator converges to the true integral almost surely as $n \to \infty$.

15.1.2 Uniform Monte Carlo integration

A natural question that arises is whether we can use a fixed sample $(X_1, \dots, X_n)$ from the measure μ to approximate the μ-integral of *every* integrable function f. Unfortunately, this is impossible, because we can select a function that vanishes on the sample points while taking nontrivial values elsewhere.

Let us refine the question. Observe that the counterexample involves a function that oscillates rapidly. We may ask whether we can use a fixed sample from the measure μ to approximate the integral of every μ-integral function *that is sufficiently regular*. In some settings, the answer is yes, provided that we use an appropriate form of regularity.

The main result of today's lecture addresses the simplest case We will consider a fixed sample of points from a probability measure supported on the unit interval in the real line. We will prove that this sample simultaneously allows us to approximate the integral of every 1-Lipschitz function on the interval.

Recall that a function $f : \Omega \to \mathbb{R}$ is L-Lipschitz if

$$|f(x) - f(y)| \le L|x - y|$$

for all $x, y \in \Omega$. The Lipschitz constant of f is defined as

$$\|f\|_{\mathrm{Lip}} := \inf\{L : f \text{ is } L\text{-Lipschitz}\}.$$

Theorem 15.1 (Uniform LLN). Let μ be a probability measure on $[0, 1]$, and let $(X_1, \dots, X_n)$ be an iid sample from μ. Consider the class of functions

$$\mathscr{F} := \{f : [0, 1] \to \mathbb{R} : \|f\|_{\mathrm{Lip}} \le 1\}.$$

Then,

$$\mathbb{E}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{I=1}^n f(X_i) - \int_\Omega f \,\mathrm{d}\mu\right| \le \frac{\text{Const}}{\sqrt{n}} \tag{15.2}$$

In other words, for a fixed sample of size n, the Monte Carlo methods can integrate every 1-Lipschitz function on the interval $[0,1]$ up to an error of $O(1/\sqrt{n})$. This is comparable with the error produced by the Monte Carlo method when applied to a single function $f \in \mathsf{L}_2(\mu)$. We have restricted our attention to a much smaller class $\mathscr{F}$ of integrands, but we can achieve the same sample complexity. We do not lose much by making the Monte Carlo method uniform.

Exercise 15.2 (ULLN: Extensions). By formal arguments, extend Theorem 15.1 to a slightly wider setting. Consider the case of a probability measure μ supported on a compact interval $[a,b] \subset \mathbb{R}$. Consider the class of L-Lipschitz functions where L is a fixed constant.

15.2 Covering Lipschitz functions

In order to prove the uniform LLN, we will introduce a random process indexed by the class $\mathscr{F}$ of Lipschitz functions. Then we will cover the class $\mathscr{F}$ with respect to an appropriate metric structure, which will allow us to use Dudley's inequality to bound the supremum. In anticipation, let us developed the required estimate for the covering numbers of $\mathscr{F}$.

Proposition 15.3 (Uniform covering of Lipschitz functions). Consider the metric space of functions

$$\mathscr{F}_0 := \{f : [0,1] \to [0,1] : \|f\|_{\mathrm{Lip}} \le 1\}$$

equipped with the uniform norm:

$$\|f\|_\infty := \sup_{x\in[0,1]} |f(x)|.$$

For all $\varepsilon > 0$,

$$N(\mathscr{F}_0, \|\cdot\|_\infty; \varepsilon) \le \exp\left(\frac{\mathrm{Const}}{\varepsilon}\right).$$

Proof. Consider the metric space $\mathscr{B} := \{f : [0,1] \to [0,1]\}$ equipped with $\|\cdot\|_\infty$, and observe that $\mathscr{F}_0 \subset \mathscr{B}$. It is more convenient to compute external covering numbers of $\mathscr{F}_0$ as a subset of $\mathscr{B}$. According to Exercise 9.13, the external covering numbers are comparable with the ordinary covering numbers:

$$\mathrm{N}_{\mathrm{ext}}(\mathscr{F}_0, \|\cdot\|_\infty; \varepsilon) \le \mathrm{N}(\mathscr{F}_0, \|\cdot\|_\infty; \varepsilon) \le \mathrm{N}_{\mathrm{ext}}(\mathscr{F}_0, \|\cdot\|_\infty; \varepsilon/2).$$

The result will follow by adjusting constants.

The idea is to approximate functions in $\mathscr{F}_0$ by piecewise constant functions, adapted to a grid. The proof concept is best understood graphically, as we see from Figure 15.2.

Since f is 1-Lipschitz, there is such a piecewise constant function $g \in \mathscr{B}$, adapted to the grid, such that $\|f - g\|_\infty \le \varepsilon$. (Why?) Therefore, adapted piecewise constant functions form an ε-cover with respect to the metric $\|\cdot\|_\infty$.

To bound the (external) covering number of $\mathscr{F}_0$ above, it suffices to count these piecewise constant functions. Here is a first effort: There are at most $1 + 1/\varepsilon$ intervals on the domain, and the function g takes at most $1 + 1/\varepsilon$ possible values. These observations give an initial bound of $(1 + 1/\varepsilon)^{1+1/\varepsilon}$ functions. We can improve this estimate by more careful reasoning.

The piecewise constant function g can take at most $(1+1/\varepsilon)$ values at the left-hand endpoint of the domain. Given the value of g on one interval in the domain, there are

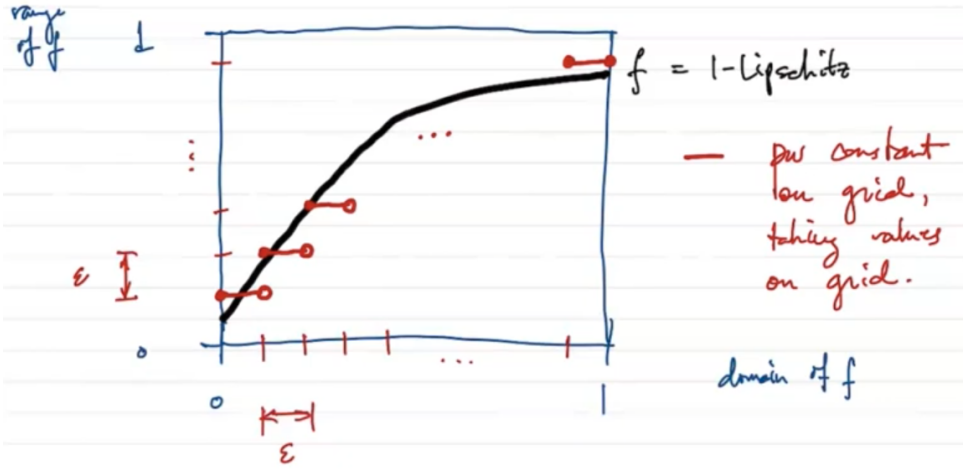


Figure 15.2 (Graphical Proof of Proposition 15.3). Here, f is a 1-Lipschitz function. Divide both the domain and range of f into a uniform grid of size ε. Approximate f by a piecewise constant function, adapted to the grid in the domain and taking values on the grid in the range.

only 3 possible values for g on the subsequent interval because f is 1-Lipschitz. This observation yields refined bound of $(1 + 1/\varepsilon) \cdot 3^{1/\varepsilon}$. It is easy to verify that this bound is less than $e^{\text{const}/\varepsilon}$, completing the proof. ∎

> **Aside:** The Arzelà–Ascoli theorem already tells us that $\mathcal{F}_0$ has finite covering numbers with respect to the uniform norm. The quantitative estimate from Proposition 15.3 is needed to invoke Dudley's chaining inequality.

Problem 15.4 (Covering Lipschitz functions on $\mathbb{R}^d$). Consider the class of 1-Lipschitz functions on the unit cube $[0,1]^d$ in $\mathbb{R}^d$, with value $f(\mathbf{0}) = 0$, and equipped with the uniform norm $\|\cdot\|_\infty$. Find a bound for the covering numbers.

15.3 Uniform LLN: Proof

Let us continue with the proof of Theorem 15.1. Without loss, we can replace $\mathcal{F}$ by the class $\mathcal{F}_0$. Indeed, the integration error is shift invariant, so we can shift each 1-Lipschitz function $f \in \mathcal{F}$ so that it takes values in $[0,1]$.

A random process

The first step is to construct a random process on the metric space $(\mathcal{F}_0, \|\cdot\|_\infty)$. For each $f \in \mathcal{F}_0$, define

$$Z_f = \frac{1}{n} \sum_{i=1}^{n} [f(X_i) - \mathbb{E} f(X_i)] \tag{15.3}$$

The random variable Z_f is the (random) error in approximating the integral of the function f by the Monte Carlo estimate induced by the (random) sample (X_i). We need to control these integration errors *uniformly*. That is, we must bound

$$\mathbb{E} \sup_{f \in \mathcal{F}} |Z_f|.$$

The key idea is to show that $(Z_f : f \in \mathcal{F}_0)$ is a centered, subgaussian random process that varies slowly as we change the function f. Therefore, Dudley's chaining inequality allows us to produce an upper bound for the supremum in terms of the covering numbers of $\mathcal{F}_0$.

Subgaussian increments

The first step is to show that the random process is subgaussian with respect to the uniform metric. For functions $f, g \in \mathcal{F}_0$, consider the increment:

$$Z_f - Z_g = \frac{1}{n} \sum\nolimits_{i=1}^{n} [(f-g)(X_i) - \mathbb{E}(f-g)(X_i)] = Z_{f-g}.$$

This is an iid sum of centered, bounded random variables. Indeed, regardless of the sample values (X_i),

$$|(f-g)(X_i) - \mathbb{E}(f-g)(X_i)| \le 2\|f-g\|_\infty.$$

Applying Hoeffding's inequality, we see that $Z_f - Z_g$ is subgaussian with variance proxy

$$v = \sum\nolimits_{i=1}^{n} \frac{1}{n^2} (2\|f-g\|_\infty)^2 = \frac{4}{n}\|f-g\|_\infty^2.$$

Therefore, we may invoke Dudley's inequality for the metric space $(\mathcal{F}_0, \|\cdot\|)$.

Dudley's inequality

To invoke the chaining bound, it is perhaps conceptually simpler to rescale the random process. Define

$$\tilde{Z}_f := \frac{\sqrt{n}}{2} Z_f \quad \text{for } f \in \mathcal{F}_0.$$

The rescaled process is centered and 1-subgaussian with respect to the uniform norm $\|\cdot\|_\infty$. Dudley's chaining inequality shows that

$$\begin{aligned}
\mathbb{E} \sup\nolimits_{f \in \mathcal{F}_0} |Z_f| &= \frac{2}{\sqrt{n}} \mathbb{E} \sup\nolimits_{f \in \mathcal{F}_0} |\tilde{Z}_f| \\
&= \frac{2}{\sqrt{n}} \int_0^{\operatorname{diam}(\mathcal{F}_0, \|\cdot\|_\infty)} \mathrm{d}\varepsilon \sqrt{\log \mathrm{N}(\mathcal{F}_0, \|\cdot\|_\infty; \varepsilon)} \\
&\le \frac{2}{\sqrt{n}} \int_0^1 \mathrm{d}\varepsilon \sqrt{\log \exp(\mathrm{Const}/\varepsilon)} \\
&= \frac{\mathrm{Const}}{\sqrt{n}} \int_0^1 \mathrm{d}\varepsilon \, \varepsilon^{-1/2} = \frac{\mathrm{Const}}{\sqrt{n}}.
\end{aligned}$$

We have used the fact that the diameter of the metric space $(\mathcal{F}_0, \|\cdot\|_\infty)$ equals 1. Afterward, we applied the covering number bound (Proposition 15.3). This calculation completes the argument.

> **Aside: (Higher dimensions?).** Versions of Theorem 15.1 can be established for Lipschitz functions on the unit cube $[0,1]^d$ in $\mathbb{R}^d$, but they exhibit a dimensional dependency. Unfortunately, to prove these results, it is not enough to obtain a bound on the covering numbers with respect to the uniform norm because the resulting entropy integral does not converge. Instead, we need more delicate tools to control the supremum of the random process.

15.4 Empirical processes

In this section, we will introduce the concept of an empirical process, and we will reinterpret the uniform LLN (Theorem 15.1) as a result about empirical processes.

15.4.1 Empirical measures

Let μ be a probability measure on a domain Ω, and let $(X_1, \dots, X_n)$ be an iid sample from the population measure μ. Define the *empirical measure*

$$\mu_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i} \quad \text{on the domain } \Omega.$$

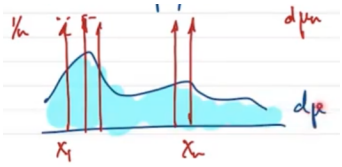


Figure 15.3 The population measure μ and the empirical measure μ_n induced by a sample of n points.

It is easy to see that μ_n is a positive measure with total mass one, so it is a probability measure. The empirical measure is atomic, even when the population measure has a density. Nevertheless, the empirical measure tends to place atoms in locations where the population measure has a lot of mass.

Here is the key question about empirical measures:

> To what extent can the empirical measure μ_n stand in for the population measure μ?

Unlike measures you may have encountered before, the empirical measure μ_n is a *random measure* because it depends on the random sample (X_i). We will not go into the details of how to define a measure-valued random variable in full generality. For the empirical measures arising in this course, no particular difficulties arise.

15.4.2 Plug-in estimators

More explicitly, there are many statistical questions we can answer completely if we know the population measure. For example, we can compute the exact expectation and variance of the distribution μ. Both the expectation and variance are moments of the measure, namely functionals of the form

$$\mu(f) = \int_{\Omega} f \, \mathrm{d}\mu \quad \text{for integrable } f : \Omega \to \mathbb{R}.$$

In statistical problems, we typically have access to data, which we can model as a sample (X_i) from the population measure. The goal is to use the observed data to make inferences about population quantities, such as the expectation and variance. A natural way to estimate a population moment $\mu(f)$ is via the *plug-in estimator*

$$\mu_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \quad \text{for integrable } f : \Omega \to \mathbb{R}.$$

At the simplest level, we would like to understand how well this approach works. Can we obtain bounds for the error

$$|\mu_n(f) - \mu(f)|?$$

Keep in mind that the error is a random quantity that depends on the sample, so the error has a distribution.

Observe that the Monte Carlo integration method (Section 15.1.1) can be described as a question about the error in approximating a population moment by an empirical moment. Our initial analysis of Monte Carlo gives some simple answers to the question.

15.4.3 Uniform bounds

We can also ask more sophisticated questions about empirical measures. In particular, we can try to understand how well they allow us to approximate an entire family of moments.

To formulate this question, let us introduce a class $\mathcal{F}$ that consists of real-valued functions $f : \Omega \to \mathbb{R}$. The class $\mathcal{F}$ usually does not contain all such functions, but only a distinguished subset (e.g., 1-Lipschitz functions).

We can ask how well the empirical measure μ_n allows us to approximate the population measure μ for the worst choice of function f in the class $\mathcal{F}$. This leads us to write down the quantity

$$\sup_{f\in\mathcal{F}} |\mu_n(f) - \mu(f)| = \sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum\nolimits_{i=1}^{n} [f(X_i) - \mathbb{E}\, f(X_i)] \right|.$$

As before, this is a random quantity that depends on the sample. As we have seen, to understand the uniform approximation error, it is natural to apply tools from the theory of random processes.

We can also recognize the uniform error bound as a statement about the distance between the empirical measure μ_n and the population measure μ with respect to the integral probability metric induced by the class $\mathcal{F}$. That is,

$$\mathrm{dist}_{\mathcal{F}}(\mu_n, \mu) = \sup_{f\in\mathcal{F}} |\mu_n(f) - \mu(f)|.$$

If $\mathcal{F}$ separates points, then the integral probability metric is a true metric.

Example 15.5 (Uniform LLN). As a simple example, we may return to the uniform law of large numbers. In this case, the class $\mathcal{F}$ consists of all 1-Lipschitz functions on the unit interval $[0, 1]$. Theorem 15.1 can be restated as

$$\mathbb{E} \sup_{f\in\mathcal{F}} |\mu_n(f) - \mu(f)| \le \frac{\mathrm{Const}}{\sqrt{n}}.$$

Thus, the ULLN is a uniform error bound for an empirical process.

The integral probability metric generated by 1-Lipschitz functions is nothing other than the Kantorovich transportation distance, also known as the Wasserstein distance. For this reason, the ULLN is sometimes called the Wasserstein law of large numbers. ∎

15.4.4 Prospects

In the next lecture, we will continue our study of uniform bounds for empirical processes. We will develop a symmetrization method that provides another way to bring forward the subgaussian nature of an empirical process. This perspective will show that the uniform norm is not always the best way to control the increments of an empirical process, and we can obtain much tighter bounds using covering numbers with respect to other metrics.

16. VC Dimension

Date: 25 February 2021 **Scribe:** Joe Slote

In the last lecture, we established the uniform law of large numbers. This result states that we can use a single sample of points from a probability measure on the unit interval $[0, 1]$ to approximate the integral of every 1-Lipschitz function. The proof relies on Dudley's chaining inequality and a covering number bound for Lipschitz functions with respect to the uniform norm. We also saw that the uniform law of large numbers can be understood as a statement about the supremum of an empirical process indexed by the class of Lipschitz functions.

In this lecture, we will continue our study of empirical processes. We will introduce a symmetrization method that allows us to bring forward the subgaussian nature of an empirical process in a way that is easier to exploit. To simplify our work, we will focus on the problem of estimating the probabilities of a family of events, rather than more general integrals. This formulation will lead us to a combinatorial notion of the complexity of a class of events, called the VC dimension, and we will learn how to bound empirical processes in terms of the VC dimension.

Agenda:

1. Recap: Empirical processes
2. Empirical error: First look
3. Giné–Zinn symmetrization
4. Estimating probabilities
5. VC dimension
6. Sauer–Shelah theorem

16.1 Empirical measures and empirical processes

An empirical process is associated with a sample of points from a measure. We begin with a formal definition of the probability measure induced by a random sample.

Definition 16.1 (Empirical measure). Consider an iid family $(X_1, \dots, X_n)$ of samples drawn from a probability measure μ on a measurable domain Ω. The *empirical probability measure* associated with the sample (X_i) is

$$\mu_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},$$

where δ_{X_i} is the Dirac measure at the sample point X_i, defined by $\delta_{X_i}(\mathsf{A}) := \mathbb{1}_{\mathsf{A}}(X_i)$ for every event A.

The empirical probability measure is a probability measure, but it is a *random* measure because it depends on the random sample (X_i). The distribution of the points

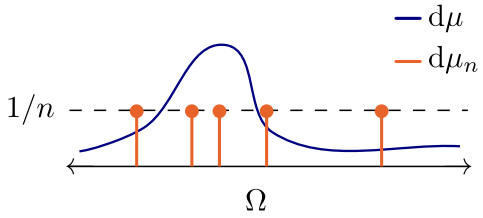


Figure 16.1 The distributions of the empirical measure μ_n and the population measure μ.

masses in the empirical measure μ_n reflects the way that the *population measure* μ distributes mass. So μ_n can roughly be thought of as a sort of histogram approximation to μ. See Figure 16.1 for an illustration.

In fact, the empirical measure is an unbiased estimator for the population measure μ. Indeed, for any event A,

$$\begin{aligned}(\mathbb{E}_{(X_i)}\,\mu_n)(\mathsf{A}) &= \mathbb{E}_{(X_i)}\left(\frac{1}{n}\sum\nolimits_{i=1}^{n}\delta_{x_i}\right)(\mathsf{A}) \\ &= \mathbb{E}_{(X_i)}\,\frac{1}{n}\sum\nolimits_{i=1}^{n}\mathbb{1}_{\mathsf{A}}(X_i) \\ &= \frac{1}{n}\sum\nolimits_{i=1}^{n}\mathbb{P}(\mathsf{A}) = \mathbb{P}(\mathsf{A}).\end{aligned}$$

This calculation is formal because we have not defined what it means to compute an expectation over the space of (signed) measures. Nevertheless, it is reasonable to assume that the expectation is linear with respect to mixtures and the action of measures on events.

16.1.1 Uniform empirical errors

Here is the central motivation for the next couple lectures:

> When may we substitute μ_n for μ? And when we do, how well does it perform?

One way to articulate this question mathematically is to compare the action of μ_n and μ on a fixed class of functions.

Definition 16.3 (Uniform empirical error). Let μ be a probability measure on a domain Ω, and let μ_n be an empirical measure associated with μ. For a class $\mathcal{F}$ of real-valued functions on the domain, the *uniform empirical error* is the quantity

$$\mathbb{E}\sup\nolimits_{f\in\mathcal{F}}\left|\mu_n(f)-\mu(f)\right| = \mathbb{E}_{(X_i)}\sup\nolimits_{f\in\mathcal{F}}\left|\mathbb{E}_{X\sim\mu_n} f(X) - \mathbb{E}_{X'\sim\mu} f(X')\right|.$$

Warning 16.2 The term "uniform empirical error" is nonstandard and is introduced for convenience. ▪

We can interpret the uniform empirical error as the distance between the empirical measure μ_n and the population measure μ under the integral probability metric induced by $\mathcal{F}$. That is,

$$\mathrm{dist}_{\mathcal{F}}(\mu_n,\mu) := \sup\nolimits_{f\in\mathcal{F}}\left|\mu_n(f)-\mu(f)\right|.$$

This construction always yields a pseudometric on probability measures, and it is a true metric whenever the class $\mathcal{F}$ separates points.

16.1.2 Empirical processes

The form of the the uniform empirical error suggests a connection to our study of suprema of random processes. Indeed, expanding the probability measures gives us

$$\begin{aligned}\mathbb{E}\sup\nolimits_{f\in\mathcal{F}}\left|\mu_n(f)-\mu(f)\right| &= \mathbb{E}_{(X_i)}\sup\nolimits_{f\in\mathcal{F}}\left|\frac{1}{n}\sum\nolimits_{i=1}^{n} f(X_i) - \mathbb{E}_{X\sim\mu} f(X)\right| \\ &= \mathbb{E}_{(X_i)}\sup\nolimits_{f\in\mathcal{F}}\left|\frac{1}{n}\sum\nolimits_{i=1}^{n}\left[f(X_i) - \mathbb{E}\,f(X_i)\right]\right| \qquad (16.1) \\ &=: \mathbb{E}_{(X_i)}\sup\nolimits_{f\in\mathcal{F}}\left|Z_f\right|.\end{aligned}$$

That is, the uniform empirical error is the expected supremum of a centered random process $(Z_f : f\in\mathcal{F})$ indexed by the class $\mathcal{F}$.

16.1.3 Chaining bounds

Let us review our initial approach to bounding the uniform empirical error using a chaining method. This approach resulted in covering numbers with respect to uniform norm on the index class. In the next section, we will develop a more powerful approach based on symmetrization.

First, let us rescale the random process:

$$\mathbb{E}_{(X_i)} \sup_{f\in\mathcal{F}} |Z_f| = \frac{1}{\sqrt{n}} \mathbb{E}_{(X_i)} \sup_{f\in\mathcal{F}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \big[f(X_i) - \mathbb{E}\, f(X_i) \big] \right|.$$

For functions $f, g \in \mathcal{F}$, the increments of the process take the form

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \big[(f-g)(X_i) - \mathbb{E}(f-g)(X_i) \big].$$

Observe that the summands are independent and centered. Therefore, Hoeffding's inequality shows that the increment is subgaussian with variance proxy

$$\frac{1}{n} \sum_{i=1}^n \|(f-g) - \mathbb{E}(f-g)\|_\infty^2 \le \frac{2}{n} \sum_{i=1}^n \|f-g\|_\infty^2 = 2\|f-g\|_\infty^2.$$

In other words, the rescaled process is subgaussian with respect to the uniform norm $\|\cdot\|_\infty$. We may apply Dudley's chaining inequality to obtain an upper bound for the supremum.

Proposition 16.4 (Empirical process: Uniform bound). With the foregoing notation,

$$\mathbb{E} \sup_{f\in\mathcal{F}} |\mu_n(f) - \mu(f)| \le \frac{\text{Const}}{\sqrt{n}} \int_0^\infty \mathrm{d}\varepsilon \sqrt{\log \mathrm{N}(\mathcal{F}, \|\cdot\|_\infty; \varepsilon)}.$$

In the last lecture, we applied this argument to establish the uniform LLN. In that case, $\mathcal{F}$ contains the 1-Lipschitz functions on the unit interval $[0, 1]$. We were able to complete the argument using a simple estimate for the covering numbers of the class $\mathcal{F}$ with respect to the uniform norm. In this particular case, the argument results in a sharp bound.

Nevertheless, in most instances, it is wasteful to bound the increments using the uniform norm. Instead, we will use an alternative method, based on symmetrization, that allows us to see that the random process is subgaussian with respect to weaker metrics.

16.2 Symmetrization of empirical processes

To continue, we develop a symmetrization method for empirical processes. We will show that this approach leads to more refined bounds for the uniform error.

16.2.1 The Giné–Zinn argument

We introduced the idea of symmetrization to obtain bounds on the moments of an independent sum. A similar approach applies to empirical processes. This formulation is due to Evarist Giné and Joel Zinn.

Proposition 16.5 (Empirical processes: Symmetrization). Let $(X_i : i = 1, \dots, n)$ be an independent sequence of random variables. For each (separable) class $\mathcal{F}$ of measurable

real-valued functions,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\big[f(X_i)-\mathbb{E}\,f(X_i)\big]\right| \le \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\big[f(X_i)-f(X_i')\big]\right| \le 2\,\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|.$$

Here, $(X_i' : i = 1,\dots,n)$ is an independent copy of the original sequence, and $(\varepsilon_i : i = 1,\dots,n)$ is a Rademacher sequence, independent from everything.

Proof. To begin, we write the expectations in terms of the independent sequence (X_i'):

$$\begin{aligned} E &:= \mathbb{E}_{(X_i)}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\big[f(X_i)-\mathbb{E}\,f(X_i)\big]\right| \\ &= \mathbb{E}_{(X_i)}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\big[f(X_i)-\mathbb{E}\,f(X_i')\big]\right| \\ &\le \mathbb{E}_{(X_i),(X_i')}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\big[f(X_i)-f(X_i')\big]\right|. \end{aligned}$$

The inequality is Jensen's. Observe that the distribution of the quantity

$$\frac{1}{n}\sum_{i=1}^{n}\big[f(X_i)-f(X_i')\big]$$

is invariant under exchange of the pair (X_j, X_j') for each index j. Therefore, it has the same distribution as

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\big[f(X_i)-f(X_i')\big].$$

In particular, under mild technical conditions,

$$E \le \mathbb{E}_{(X_i),(X_i')}\,\mathbb{E}_{(\varepsilon_i)}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\big[f(X_i)-f(X_i')\big]\right|.$$

The last bound follows from the triangle inequality. ∎

16.2.2 Chaining bound

The power of the symmetrization method derives from the fact that we now have two independent sources of randomness in the process: (X_i) and (ε_i). Therefore, we may condition on the choice (X_i) of the sample, and we may average over the Rademacher variables (ε_i).

Fix the sample points (X_i), and consider the symmetrized random process

$$\tilde{Z}_f := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i f(X_i) \quad \text{for } f\in\mathcal{F}.$$

Conditional on (X_i), this random process is a Rademacher series. Applying Hoeffding's inequality (conditionally!), we see that the increments $\tilde{Z}_f - \tilde{Z}_g$ are subgaussian with variance proxy

$$\frac{1}{n}\sum_{i=1}^{n}(f(X_i)-g(X_i))^2 = \mu_n(f-g)^2 = \|f-g\|_{\mathsf{L}_2(\mu_n)}.$$

Therefore, when we apply the chaining inequality to control the conditioned random process, we obtain a bound in terms of the covering numbers of the class $\mathcal{F}$ with respect to the *random* metric $\mathsf{L}_2(\mu_n)$.

Proposition 16.6 (Empirical process: Symmetrized bound). With the prevailing notation,

$$\mathbb{E} \sup_{f \in \mathcal{F}} |\mu_n(f) - \mu(f)| \leq \mathbb{E}_{(X_i)} \frac{\text{Const}}{\sqrt{n}} \int_0^\infty d\varepsilon \sqrt{\log \mathrm{N}(\mathcal{F}, \mathsf{L}_2(\mu_n); \varepsilon)}.$$

Comparing the bounds from Propositions 16.4 and 16.6, we see that a miracle has occurred. We have passed from the $\mathsf{L}_\infty(\Omega)$ norm on the entire domain to the $\mathsf{L}_2(\mu_n)$ norm on the sample points only. In many cases, this is a vast improvement—even when we make a uniform bound over the worst possible choice (X_i) of the sample. In our analyses, the only information about the sample we will use is the number n of sample points.

16.3 Empirical estimates for probabilities

Proposition 16.6 is a very powerful, general result that can be used to obtain bounds in a wide range of settings. To make the ideas as clear as possible, we will focus on the case where the class $\mathcal{F}$ consists of indicator functions. In other words, we are interested in making empirical estimates for the probability of a collection of events.

16.3.1 Indicator function classes

We will be working extensively with indicators, so it is worth a moment to recall the definitions. The indicator of an event $\mathsf{A} \subseteq \Omega$ is defined as

$$\mathbb{1}_{\mathsf{A}}(\omega) := \begin{cases} 1, & \omega \in \mathsf{A}; \\ 0, & \omega \notin \mathsf{A}. \end{cases}$$

To avoid extra notation, it is convenient to treat the event A and its indicator $\mathbb{1}_{\mathsf{A}}$ as the same object.

Let $\mathscr{C}$ be a collection of events from Ω. We can just as well think about $\mathscr{C}$ as a class of (indicator) functions. For each event $\mathsf{C} \in \mathscr{C}$, observe that

$$\mathbb{P}(\mathsf{C}) = \mu(\mathsf{C}) = \mathbb{E}_{(X_i)}\, \mu_n(\mathsf{C}),$$

because the empirical measure μ_n is an unbiased estimator for the population measure μ. Moreover,

$$\mu_n(\mathsf{C}) = \frac{1}{n} \sum\nolimits_{i=1}^n \delta_{X_i}(\mathsf{C}) = \frac{|\{i : X_i \in \mathsf{C}\}|}{n}.$$

These formulas suggest a simple protocol for estimating the probabilities of the events in the class $\mathscr{C}$. Generate and fix an iid sample $(X_i : i = 1, \dots, n)$ from the measure μ. To estimate the probability $\mu(\mathsf{C})$, we simply report the proportion of the sample points (X_i) that belong to the event C.

We can use the empirical process tools that we have developed to study the efficacy of this procedure. The uniform empirical error takes the form

$$\mathbb{E} \sup_{\mathsf{C} \in \mathscr{C}} |\mu_n(\mathsf{C}) - \mu(\mathsf{C})| = \mathbb{E} \sup_{\mathsf{C} \in \mathscr{C}} \left| \frac{|\{i : X_i \in \mathsf{C}\}|}{n} - \mathbb{P}(\mathsf{C}) \right|.$$

This is the "expected worst-case additive error in estimating the probabilities of all events in $\mathscr{C}$ from a shared sample (X_i)."

16.3.2 Uniform covering of sample points

We can use Proposition 16.6 to obtain an upper bound on the uniform empirical error in terms of the covering numbers of $\mathscr{C}$ by the $\mathsf{L}_2(\mu_n)$ norm. Today, to fix some ideas, we will pass to the stronger norm $\mathsf{L}_\infty(\mu_n)$. Recall that this norm is calculated as

$$\|f\|_{\mathsf{L}_\infty(\mu_n)} := \max_i |f(X_i)|.$$

The next exercise gives a description of covering numbers with respect to this norm.

Exercise 16.7 (Sample points: Uniform covering). Show that

$$\mathrm{N}(\mathscr{C}, \|\cdot\|_{\mathsf{L}_2(\mu_n)}; \varepsilon) \leq \mathrm{N}(\mathscr{C}, \|\cdot\|_{\mathsf{L}_\infty(\mu_n)}; \varepsilon).$$

(**Hint**: This is true for any function class.) Argue that, for any two events $\mathsf{C}, \mathsf{D} \in \mathscr{C}$,

This is the trivial metric with respect to an appropriate similarity relation.

$$\|\mathbb{1}_\mathsf{C} - \mathbb{1}_\mathsf{D}\|_{\mathsf{L}_\infty(\mu_n)} = \begin{cases} 0, & \text{if } \mathsf{C} \cap \{X_1, \dots, X_n\} = \mathsf{D} \cap \{X_1, \dots, X_n\}; \\ 1, & \text{otherwise.} \end{cases}$$

Finally, use this fact to prove

$$\mathrm{N}(\mathscr{C}, \|\cdot\|_{\mathsf{L}_\infty(\mu_n)}; \varepsilon) = \begin{cases} 1, & \text{if } \varepsilon \geq 1; \\ |\mathscr{C} \cap \{X_1, \dots, X_n\}|, & \text{if } 0 < \varepsilon < 1, \end{cases}$$

where $\mathscr{C} \cap \{X_1, \dots, X_n\} := \big\{\mathsf{C} \cap \{X_1, \dots, X_n\} : \mathsf{C} \in \mathscr{C}\big\}$.

The simplifications in the last problem come at the expense of weakening our bounds. Because we are working with indicator functions, this simplification is not as vulgar as in at the general case, but it is still not optimal. We will remedy this error in Lecture 17.

16.3.3 Combinatorial bounds

The passage to the $\mathsf{L}_\infty(\mu_n)$ norm allows us to obtain bounds for the empirical error that have a combinatorial flavor.

Proposition 16.8 (Empirical processes: Combinatorial bound). Let $\mathscr{C}$ be a class of indicator functions. Then

$$\mathbb{E} \sup_{\mathsf{C} \in \mathscr{C}} |\mu_n(\mathsf{C}) - \mu(\mathsf{C})| \leq \mathbb{E}_{(X_i)} \frac{\mathrm{Const}}{\sqrt{n}} \sqrt{\log |\mathscr{C} \cap \{X_1, \dots, X_n\}|}.$$

Proof. Applying Proposition 16.6 and the results of Exercise 16.7, we have

$$\begin{aligned} \mathbb{E} \sup_{\mathsf{C} \in \mathscr{C}} |\mu_n(\mathsf{C}) - \mu(\mathsf{C})| &\leq \mathbb{E}_{(X_i)} \frac{\mathrm{Const}}{\sqrt{n}} \int_0^\infty \mathrm{d}\varepsilon \sqrt{\log \mathrm{N}(\mathscr{C}, \mathsf{L}_2(\mu_n); \varepsilon)} \\ &\leq \mathbb{E}_{(X_i)} \frac{\mathrm{Const}}{\sqrt{n}} \int_0^1 \mathrm{d}\varepsilon \sqrt{\log \mathrm{N}(\mathscr{C}, \mathsf{L}_\infty(\mu_n); \varepsilon)} \\ &= \mathbb{E}_{(X_i)} \frac{\mathrm{const}}{\sqrt{n}} \sqrt{\log |\mathscr{C} \cap \{X_1, \dots, X_n\}|}. \end{aligned}$$

This calculation completes the argument. ∎

We have thus arrived at a purely combinatorial upper bound on the covering numbers, captured by the number of distinct subsets of $\mathsf{X} := \{X_1, \dots, X_n\}$ arising as an intersection of X with a member of $\mathscr{C}$.

Example 16.9 (Estimating distribution functions). Consider the domain $\Omega = \mathbb{R}$, and let μ be any probability distribution on $\mathbb{R}$. We would like to compute an empirical approximation of the distribution function

$$F(a) \coloneqq \mathbb{P}_{X\sim\mu}\{X \le a\} = \mu\{X \le a\} \quad \text{for } a \in \mathbb{R}.$$

Given a random sample $(X_1, \dots, X_n)$, we can form the empirical approximation

$$F_n(a) \coloneqq \frac{1}{n}|\{i : X_i \le a\}|.$$

Let us use our results to quantify the error in the approximation as a function of the number n of samples.

Consider the class of left half-lines:

$$\mathscr{C} \coloneqq \{(-\infty, a] : a \in \mathbb{R}\}.$$

The distribution function F packs up the probability $\mu(-\infty, a]$ for each event in this class, and the empirical estimate F_n packs up the empirical measures $\mu_n(-\infty, a]$ of the events. Therefore,

$$\sup_{a\in\mathbb{R}} |F_n(a) - F(a)| = \sup_{\mathsf{C}\in\mathscr{C}} |\mu_n(\mathsf{C}) - \mu(\mathsf{C})|.$$

We will use Proposition 16.8 to obtain a bound on the expectation of the uniform error.

Without loss of generality, relabel the points in the sample $X_1, \dots, X_n$ so they are listed in increasing order. We must determine the number of distinct sets of the form $\mathsf{C} \cap \{X_1, \dots, X_n\}$ where $\mathsf{C} \in \mathscr{C}$. We claim that there are at most $n + 1$ possibilities. Indeed, if a sample $X_i \in (-\infty, a]$, then the previous sample $X_{i-1} \in (-\infty, a]$. Therefore,

$$\mathsf{C} \cap \{X_1, \dots, X_n\} = \varnothing \quad \text{or} \quad \mathsf{C} \cap \{X_1, \dots, X_n\} = \{X_1, \dots, X_i\} \quad \text{for some index } i.$$

As a consequence,

$$\mathbb{E} \sup_{a\in\mathbb{R}} |F_n(a) - F(a)| = \mathbb{E} \sup_{\mathsf{C}\in\mathscr{C}} |\mu_n(\mathsf{C}) - \mu(\mathsf{C})| \le \frac{\text{Const}}{\sqrt{n}} \sqrt{\log(n+1)}.$$

In other words, for an arbitrary probability measure μ, if we want an estimate of the distribution function F with a uniform error of ε, then it suffices to take about $n = O(\varepsilon^{-2} \log(\varepsilon^{-2}))$ samples. This result is a suboptimal version of the Glivenko–Cantelli theorem; we will obtain the optimal result next time. ∎

16.4 Combinatorial dimension

Now, let us develop a far-reaching generalization of the combinatorial argument in Example 16.9.

16.4.1 Shattering and VC dimension

To employ the combinatorial bound (Proposition 16.8), we need to count the number of sets that arise when we intersect the class $\mathscr{C}$ with the finite sample $\{X_1, \dots, X_n\}$. The extreme case evidently occurs when *every subset* of the sample can arise from the intersection with an event C in the class.

We need a more refined method for counting the subsets that can occur. Instead, we want to understand when the class C can isolate *every subset of a subset* of the domain.

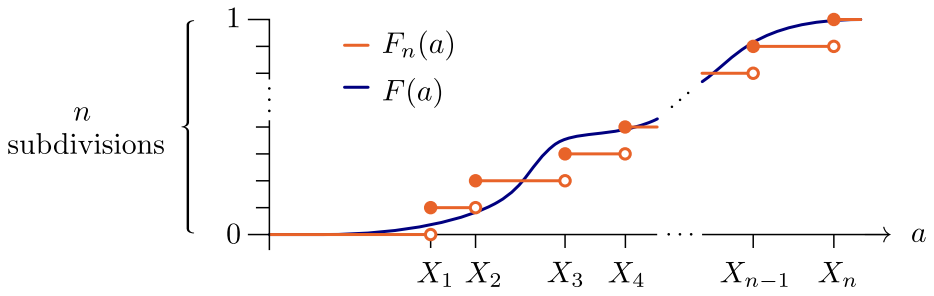


Figure 16.2 A distribution function F along with its empirical approximation F_n. Notice F_n can miss F entirely on a region.

Definition 16.10 (Shattering). A class $\mathscr{C}$ of subsets of the domain Ω *shatters* a set $I \subseteq \Omega$ when, for all $J \subseteq I$, there exists a set $C \in \mathscr{C}$ such that $J = C \cap I$. Equivalently, $\mathscr{C}$ shatters I when $\mathscr{C} \cap I = \mathscr{P}(I)$, the power set of I.

We include the case $J = \emptyset$.

We will define the combinatorial dimension of a class $\mathscr{C}$ to be the maximum cardinality of a shattered set.

Definition 16.11 (VC dimension). The *Vapnik–Chervonenkis (VC) dimension* of a class $\mathscr{C}$ of sets is the maximum cardinality $\mathrm{vc}(\mathscr{C})$ of a set I that is shattered by $\mathscr{C}$. Note that $\mathrm{vc}(\mathscr{C})$ can be infinite.

This definition is somewhat hard to appreciate at first sight. Nevertheless, there are several reasons that it is valuable. As we will see, there are many classes of sets for which we can bound the VC dimension. Second, we can obtain simple bounds on the sample complexity of statistical estimation problems in terms of the VC dimension. This will be the theme of the next few lectures.

16.4.2 VC dimension: Examples

Let us give some examples of classes of sets where we can compute the VC dimension using simple arguments. Keep in mind that it is easy to get turned around when establishing VC dimension bounds. In order to show that $\mathrm{vc}(\mathscr{C}) < n$, we must demonstrate that there is *no set with cardinality n* that is shattered by $\mathscr{C}$. Conversely, to show that $\mathrm{vc}(\mathscr{C}) \geq n$, we need only exhibit *one set with cardinality n* that is shattered by $\mathscr{C}$.

Example 16.12 (Left half-lines). Let $\Omega = \mathbb{R}$, and consider the class of left half-lines $\mathscr{C} = \{(-\infty, c] : c \in \mathbb{R}\}$. We claim that $\mathrm{vc}(\mathscr{C}) = 1$.

To verify this point, observe that every set of cardinality 1 is shattered. Now, consider any set $I = \{a, b\}$ with cardinality 2. By relabeling, we may assume that $a < b$. Intersecting I with the class $\mathscr{C}$ of half-lines, we can obtain the subsets $\emptyset$, $\{a\}$, and the set $\{a, b\}$ itself. We *cannot* recover the singleton $\{b\}$ because $b \in C$ implies that $a \in C$ for every set $C = (-\infty, c]$. ∎

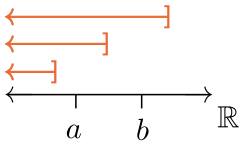


Figure 16.3 Any left half-line that covers b also covers a.

Exercise 16.13 (Half-planes). Suppose $\Omega = \mathbb{R}^2$ and $\mathscr{C}$ is the collection of all two-dimensional half-spaces; that is,

$$\mathscr{C} = \left\{\{\boldsymbol{x} : \langle \boldsymbol{a}, \boldsymbol{x} \rangle \leq b\} : \boldsymbol{a} \in \mathbb{R}^2, b \in \mathbb{R}\right\}.$$

Show that $\mathrm{vc}(\mathscr{C}) = 3$.

Exercise 16.14 (Half-spaces). Let $\Omega = \mathbb{R}^d$, and consider the class

$$\mathscr{C} = \big\{\{\boldsymbol{x} : \langle \boldsymbol{a}, \boldsymbol{x}\rangle \le b\} : \boldsymbol{a} \in \mathbb{R}^d, b \in \mathbb{R}\big\}.$$

Show that $\mathrm{vc}(\mathscr{C}) = d + 1$. **Hint:** Apply Radon's Theorem.

Exercise 16.15 (Axis-aligned rectangles). Let $\Omega = \mathbb{R}^2$, and consider

$$\mathscr{C} = \{[a, b] \times [c, d] : a < b, c < d \in \mathbb{R}\}.$$

Show that $\mathrm{vc}(\mathscr{C}) = 4$.

Exercise 16.16 (General rectangles). Let $\Omega = \mathbb{R}^2$, and let $\mathscr{C}$ be the set of all rectangles in $\mathbb{R}^2$. Show that $\mathrm{vc}(\mathscr{C}) = 7$.

Exercise 16.17 (Convex polygons). Let $\Omega = \mathbb{R}^2$, and let $\mathscr{C}$ be the class containing all convex polygons. Show that $\mathrm{vc}(\mathscr{C}) = +\infty$.

16.4.3 VC dimension: Counting intersections

The VC dimension is relevant to our study of empirical processes indexed by sets. The reason is that it serves as a bound on the cardinality $|\mathscr{C} \cap \{X_1, \dots, X_n\}|$ that appears in Proposition 16.8. To establish this fact, we need a classic result from extremal set theory.

Theorem 16.18 (Sauer–Shelah and others). Let $\mathscr{C}$ be a class of subsets of the domain Ω, and choose $n \ge \mathrm{vc}(\mathscr{C})$. For any collection $(X_1, \dots, X_n) \subset \Omega$,

$$|\mathscr{C} \cap \{X_1, \dots, X_n\}| \le \sum\nolimits_{i=0}^{\mathrm{vc}(\mathscr{C})} \binom{n}{i} \le \left(\frac{\mathrm{e}n}{\mathrm{vc}(\mathscr{C})}\right)^{\mathrm{vc}(\mathscr{C})}.$$

We will establish Theorem 16.18 in the next lecture.

16.4.4 Empirical processes: VC dimension bound

Using the Sauer–Shelah theorem, we can express the bound from Proposition 16.8 directly in terms of the VC dimension.

Corollary 16.19 (Empirical process: Preliminary VC bound). Let $\mathscr{C}$ be a class of subsets of a domain Ω, and choose a number $n \ge \mathrm{vc}(\mathscr{C})$. Let μ be the population measure, and let μ_n be the empirical measure associated with a sample of size n from the population measure. Then

$$\mathbb{E} \sup\nolimits_{\mathsf{C} \in \mathscr{C}} |\mu_n(\mathsf{C}) - \mu(\mathsf{C})| \le \mathrm{Const} \cdot \sqrt{\mathrm{vc}(\mathscr{C})} \cdot \sqrt{\frac{\log n}{n}}.$$

Proof. Combine Proposition 16.8 with Theorem 16.18 ∎

As a consequence of this result, we can simultaneously estimate the probability of every event in the class $\mathscr{C}$ up to an absolute error ε using a sample of size

$$n = O(\varepsilon^{-2} \mathrm{vc}(\mathscr{C}) \log(\varepsilon^{-2} \mathrm{vc}(\mathscr{C}))).$$

This statement remains true for every probability measure μ on the domain Ω.

As a particular example, Corollary 16.19 contains the result in Example 16.9. Using other VC dimension computations, we can obtain many similar results.

Nevertheless, Corollary 16.19 is suboptimal. We can trace the weakness to the fact that we passed from the $\mathsf{L}_2(\mu_n)$ norm to the $\mathsf{L}_\infty(\mu_n)$ norm in the proof of Proposition 16.8. In the next lecture, we will repair this defect.

17. VC Bounds for Empirical Processes

Date: 2 March 2021 **Scribe:** Joe Slote

In Lecture 16, we obtained bounds for the uniform error in estimating the probability of a class of events from an empirical sample. These bounds were expressed in terms of the VC dimension, a combinatorial notion of the complexity of the class of events. Unfortunately, these results include a parasitic logarithm that arises from making a coarse estimate in the proof. In this lecture, we will remove the parasitic factor. The resulting bound shows that the number of samples we need to estimate the probabilities of a class of events is proportional to the VC dimension.

Agenda:
1. L_2 covering numbers
2. Uniform Glivenko–Cantelli
3. Proof of Sauer–Shelah

17.1 Dudley's covering number bound

Corollary 16.19 from the last lecture gives a uniform bound on the error in estimating the probabilities of events from a class $\mathscr{C}$ using an empirical sample from the population measure. The proof of this result involves symmetrization and Dudley's chaining inequality, which leads to a bound in terms of the covering numbers of the class $\mathscr{C}$ with respect to the $L_2(\mu_n)$ norm, determined by the empirical measure μ_n. We passed to the stronger $L_\infty(\mu_n)$ norm, and we developed a combinatorial bound for the covering numbers in terms of the VC dimension. The weakness in this argument is the passage from $L_2(\mu_n)$ to $L_\infty(\mu_n)$, which results in an adverse dependency on the number n of sample points.

The next result, due to Dudley, gives a bound on the $L_2(\mu)$ covering numbers of $\mathscr{C}$ in terms of the VC dimension. In contrast to the earlier result, this bound does not have any dependency on the size of the support of the measure μ.

Theorem 17.1 (Dudley). Let $\mathscr{C}$ be a class of events on a measurable space Ω. For $\varepsilon > 0$,

$$\sup_\mu N(\mathscr{C}, \|\cdot\|_{L_2(\mu)}; \varepsilon) \le \left(\frac{\text{Const}}{\varepsilon}\right)^{\text{Const}\cdot \text{vc}(\mathscr{C})}.$$

The supremum occurs over all probability measures on Ω.

Aside: A more difficult result, due to Haussler, shows that the sharp exponent in Theorem 17.1 is $2\,\text{vc}(\mathscr{C})$.

We will prove Theorem 17.1 after a few remarks. Let us emphasize that the covering number bound is uniform over all probability measures; it uses no information about the form of the measure. It is also fruitful to compare the new bound with our existing covering number bounds.

First, recall the bound for the covering numbers with respect to $L_\infty(\mu_n)$, where μ_n is an empirical measure with at most n atoms. Then

$$N(\mathscr{C}, L_2(\mu_n); \varepsilon) \le N(\mathscr{C}, L_\infty(\mu_n); \varepsilon) \le n^{\text{Const}\cdot \text{vc}(\mathscr{C})}.$$

The bound on the $L_\infty(\mu_n)$ covering numbers depends explicitly on the number n of atoms, although the scale ε no longer plays a role. In contrast, Theorem 17.1 removes this dimensional dependency.

Second, we may make a formal comparison between Theorem 17.1 and our volumetric covering bounds. For example, if B is the unit ball of a d-dimensional normed space $\|\cdot\|$, then

$$\mathrm{N}(\mathsf{B}, \|\cdot\|; \varepsilon) \le \left(\frac{\text{Const}}{\varepsilon}\right)^d.$$

The logarithm of the number of ε-balls we need to cover the unit ball is proportional to $\log(1/\varepsilon)$ and to the dimension d of the normed space. This is exactly the same scaling that we see in Theorem 17.1, except that the linear dimension is replaced by the combinatorial VC dimension.

17.1.1 The extraction lemma

The proof of Theorem 17.1 is based on a dimension reduction argument called probabilistic extraction. The idea is that we can discriminate a well-separated family of sets in $(\mathscr{C}, \mathsf{L}_2(\mu))$ by examining a very small number of points in the domain.

Lemma 17.2 (Extraction). Let $\{\mathsf{C}_1, \dots, \mathsf{C}_m\}$ be a collection of sets in Ω which are ε-separated with respect to the norm $\mathsf{L}_2(\mu)$. That is,

$$\|\mathbb{1}_{\mathsf{C}_i} - \mathbb{1}_{\mathsf{C}_j}\|_{\mathsf{L}_2(\mu)} > \varepsilon \quad \text{for all indices } i \ne j.$$

Then there exists a point set $\mathsf{X} = \{x_1, \dots, x_r\}$ with

$$r \le \text{Const} \cdot \varepsilon^{-2} \cdot \log m$$

that discriminates the sets:

$$\mathsf{C}_i \cap \mathsf{X} \ne \mathsf{C}_j \cap \mathsf{X} \quad \text{for all indices } i \ne j.$$

Proof. The proof relies on the probabilistic method. Draw and fix a random set $\mathsf{X} = \{X_1, \dots, X_r\}$ where $X_i \sim \mu$ iid. When r is sufficiently large, we will argue that X discriminates all pairs $(\mathsf{C}_i, \mathsf{C}_j)$ of the sets with positive probability. As a consequence, there must exist some set X consisting of r points that discriminates all the sets.

To that end, we first develop an upper bound on the probability that

$$\mathbb{P}\{\mathsf{C} \cap \mathsf{X} = \mathsf{C}' \cap \mathsf{X}\} \quad \text{for } \varepsilon\text{-separated sets } \mathsf{C}, \mathsf{C}'.$$

Observe that a sample X_i discriminates the two sets if and only if X_i belongs to exactly one of the sets: $X_i \in \mathsf{C} \triangle \mathsf{C}'$, where $\triangle$ is the symmetric difference. Furthermore, $\mathsf{C} \cap \mathsf{X} = \mathsf{C}' \cap \mathsf{X}$ precisely when X contains no distinguishing point. By the independence of the sample,

$$\begin{aligned}\mathbb{P}\{\mathsf{C} \cap \mathsf{X} = \mathsf{C}' \cap \mathsf{X}\} &= \prod\nolimits_{i=1}^{r} \mathbb{P}\{X_i \notin \mathsf{C} \triangle \mathsf{C}'\} = (1 - \mathbb{P}\{X_i \in \mathsf{C} \triangle \mathsf{C}'\})^r \\ &= (1 - \mu(\mathbb{1}_{\mathsf{C} \triangle \mathsf{C}'}))^r = \left(1 - \|\mathbb{1}_{\mathsf{C}} - \mathbb{1}_{\mathsf{C}'}\|_{\mathsf{L}_2(\mu)}^2\right)^r \le (1 - \varepsilon^2)^r.\end{aligned}$$

We have used the fact that $\mathbb{1}_{\mathsf{C} \triangle \mathsf{C}'} = (\mathbb{1}_{\mathsf{C}} - \mathbb{1}_{\mathsf{C}'})^2$, and the assumption that the two sets are ε-separated.

Thus, the probability that the point set X distinguishes each pair of sets $(\mathsf{C}_i, \mathsf{C}_j)$ is

$$\mathbb{P}\left\{\forall i \ne j : \mathsf{C}_i \cap \mathsf{X} \ne \mathsf{C}_j \cap \mathsf{X}\right\} \ge 1 - \binom{m}{2}\left(1 - \varepsilon^2\right)^r \ge 1 - m^2\left(1 - \varepsilon^2\right)^r > 0.$$

The final inequality holds when we choose $r = \text{Const} \cdot \varepsilon^{-2} \log m$. The result follows. ∎

Aside: The extraction lemma is similar in spirit to the Johnson–Lindenstrauss theorem on dimension reduction. Indeed, we have shown that an ε-separated family $\{\mathbb{1}_{\mathsf{C}_i}\}$ of indicator functions can be replaced by the restricted indicators $\{\mathbb{1}_{\mathsf{C}_i \cap \mathsf{X}}\}$ while preserving the fact that the sets are distinct, provided that X contains $r = \text{Const} \cdot \varepsilon^{-2} \log m$ points. If we allow a larger number $r = \text{Const} \cdot \varepsilon^{-4} \log m$ of sample points, then we can even ensure that the reduced sets remain well-separated:

$$\|\mathbb{1}_{\mathsf{C}_i \cap \mathsf{X}} - \mathbb{1}_{\mathsf{C}_j \cap \mathsf{X}}\|_{\mathsf{L}_2(\mu)} \geq \varepsilon/2 \quad \text{for all indices } i \neq j.$$

This improvement is not important for our purposes, but it plays a role when we generalize this theory from sets to functions.

17.1.2 Proof of Theorem 17.1

With the extraction lemma in hand, we easily complete the proof of Dudley's covering number bound.

Let $\{\mathsf{C}_1, \dots, \mathsf{C}_m\} \subseteq \mathscr{C}$ be an ε-packing of $(\mathscr{C}, \mathsf{L}_2(\mu))$ with maximum cardinality, so each pair of sets is ε-separated with respect to the $\mathsf{L}_2(\mu)$ norm. By extraction (Lemma 17.2), there is a set X consisting of $r = \text{Const} \cdot \varepsilon^{-2} \cdot \log m$ points with the property that each set $\mathsf{C}_i \cap \mathsf{X}$ is distinct from the rest. Thus, we may calculate that

$$\begin{aligned} m = |\{\mathsf{C}_1 \cap \mathsf{X}, \dots, \mathsf{C}_m \cap \mathsf{X}\}| &\leq |\mathscr{C} \cap \mathsf{X}| \leq \left(\frac{\mathrm{e} r}{\mathrm{vc}(\mathscr{C})} \right)^{\mathrm{vc}(\mathscr{C})} \\ &\leq \left(\frac{\log m}{\mathrm{vc}(\mathscr{C})} \cdot \frac{\text{Const}}{\varepsilon^2} \right)^{\mathrm{vc}(\mathscr{C})} \leq \left(m^{2\delta/\mathrm{vc}(\mathscr{C})} \cdot \frac{\text{Const}(\delta)}{\varepsilon^2} \right)^{\mathrm{vc}(\mathscr{C})}. \end{aligned}$$

The last inequality in the first line follows from the Sauer–Shelah theorem. In the last inequality, we have introduced a parameter $\delta \in (0, 1)$. Solving for m, we obtain

$$m \leq \left(\frac{\text{Const}(\delta)}{\varepsilon} \right)^{2\,\mathrm{vc}(\mathscr{C})/(1-\delta)}.$$

Finally, recall the duality between covering numbers and packing numbers:

$$\mathrm{N}(\mathscr{C}, \mathsf{L}_2(\mu); 2\varepsilon) \leq \mathrm{P}(\mathscr{C}, \mathsf{L}_2(\mu); \varepsilon) = m.$$

Introduce the bound for m to complete the argument.

17.2 VC bounds for empirical processes

We are now prepared to establish a better bound on the uniform empirical error in estimating the probabilities of a class of events.

Theorem 17.3 (Empirical processes: VC bound). Let μ be a probability measure on a measurable space Ω, and let $\mathscr{C}$ be a class of events with finite VC dimension. For $n \geq \mathrm{vc}(\mathscr{C})$, the empirical measure μ_n of an iid sample of n points from μ satisfies

$$\mathbb{E} \sup_{\mathsf{C} \in \mathscr{C}} |\mu_n(\mathsf{C}) - \mu(\mathsf{C})| \leq \text{Const} \cdot \sqrt{\frac{\mathrm{vc}(\mathscr{C})}{n}}.$$

In particular, to achieve uniform empirical error ε, it suffices to take $n = \text{Const} \cdot \varepsilon^{-2} \mathrm{vc}(\mathscr{C})$ iid samples.

As compared with Corollary 16.19, we have removed a parasitic logarithmic factor in the error bound.

Proof. We combine the symmetrized chaining bound (Proposition 16.6) on the empirical error with Dudley's covering number bound (Theorem 17.1) to obtain

$$\begin{aligned}
\mathbb{E} \sup_{\mathsf{C} \in \mathscr{C}} |\mu_n(\mathsf{C}) - \mu(\mathsf{C})| &\le \mathbb{E}_{(X_i)} \frac{\text{Const}}{\sqrt{n}} \int_0^\infty \mathrm{d}\varepsilon \sqrt{\log \mathrm{N}(\mathscr{C}, \mathsf{L}_2(\mu_n); \varepsilon)} \\
&\le \mathbb{E}_{(X_i)} \frac{\text{Const}}{\sqrt{n}} \int_0^1 \mathrm{d}\varepsilon \sqrt{\log \mathrm{N}(\mathscr{C}, \mathsf{L}_2(\mu_n); \varepsilon)} \\
&\le \frac{\text{Const}}{\sqrt{n}} \int_0^1 \mathrm{d}\varepsilon \sqrt{\log \left(\frac{\text{Const}}{\varepsilon}\right)^{\text{Const} \cdot \mathrm{vc}(\mathscr{C})}} \\
&= \text{Const} \cdot \sqrt{\frac{\mathrm{vc}(\mathscr{C})}{n}} \int_0^1 \mathrm{d}\varepsilon \sqrt{\log(\text{Const}/\varepsilon)}.
\end{aligned}$$

The integral is convergent, and the argument is complete. ∎

17.2.1 Examples

To understand the implications of Theorem 17.3, let us present a few examples.

Example 17.4 (Estimating distribution functions, reprise). Let $\Omega = \mathbb{R}$ equipped with a probability measure μ that has distribution function F. Draw an iid sample $(X_1, \dots, X_n)$ from μ with empirical measure μ_n. We can estimate the distribution function as

$$F_n(a) = \frac{|\{i : X_i \le a\}|}{n}.$$

To analyze the quality of this approximation, we define the class of left half-lines: $\mathscr{C} = \{(-\infty, a] : a \in \mathbb{R}\}$. Then

$$\sup_{a \in \mathbb{R}} |F_n(a) - F(a)| = \sup_{\mathsf{C} \in \mathscr{C}} |\mu_n(\mathsf{C}) - \mu(\mathsf{C})|.$$

We have seen that the $\mathrm{vc}(\mathscr{C}) = 1$. Theorem 17.3 implies that

$$\mathbb{E} \sup_{a \in \mathbb{R}} |F_n(a) - F(a)| \le \frac{\text{Const}}{\sqrt{n}}.$$

In other words, we can obtain a uniform empirical estimate of the distribution function F with error ε using $n = O(\varepsilon^{-2})$ samples. This is a logarithmic improvement over Example 16.9, and it gives the optimal scaling for a general measure. This result is called the Glivenko–Cantelli theorem. ▪

Example 17.5 (Half spaces). Let $\Omega = \mathbb{R}^d$ be equipped with a probability measure μ. Consider the class of half spaces

$$\mathscr{C} = \big\{\{\boldsymbol{x} : \langle \boldsymbol{a}, \boldsymbol{x} \rangle \le b\} : \boldsymbol{a} \in \mathbb{R}^d, b \in \mathbb{R}\big\}.$$

Exercise 16.14 reports that $\mathrm{vc}(\mathscr{C}) = d + 1$. Theorem 17.3 states that we can achieve a constant uniform error estimate for the measure of every halfspace in $\mathbb{R}^d$ with about $n \approx d$ samples. ▪

17.2.2 Uniform Glivenko–Cantelli classes

Here is a little more context. We say that a class $\mathscr{C}$ of sets is *uniform Glivenko–Cantelli* if

$$\sup_\mu \mathbb{E} \sup_{C \in \mathscr{C}} |\mu_n(C) - \mu(C)| \to 0 \quad \text{as } n \to \infty.$$

Here, μ ranges over probability measures on the domain, and μ_n is an empirical measure associated with an iid sample of n points from the measure μ. In other words, a class is uniform Glivenko–Cantelli if we can estimate the probability of every event to a uniform tolerance with a fixed number of samples, and we can make this tolerance as small as we like. Furthermore, we must be able to do so for any measure of probability.

The classic example of a uniform Glivenko–Cantelli class is the set of left half-lines in $\mathbb{R}$, which arises when we try to estimate distribution functions. The original Glivenko–Cantelli theorem asserts that this class is indeed a uniform Glivenko–Cantelli class.

Theorem 17.3 goes further. It states that every class $\mathscr{C}$ with finite VC dimension is uniform Glivenko–Cantelli. In fact, the converse is true as well.

Problem 17.6 (Finite VC dimension: Necessity). Prove that a class $\mathscr{C}$ of sets is uniform Glivenko–Cantelli if *and only if* $\mathscr{C}$ has finite VC dimension.

17.3 Sauer–Shelah: Proof

It remains to establish the Sauer–Shelah theorem, which allows us to count sets using VC dimension. This result has been independently discovered by many researchers, but we have given the most common nomenclature. For reference, we restate the result.

Theorem 17.7 (Sauer–Shelah and others). Let $\mathscr{C}$ be a class of subsets of the domain Ω, and choose $n \geq \operatorname{vc}(\mathscr{C})$. For any collection $(X_1, \dots, X_n) \subset \Omega$,

$$|\mathscr{C} \cap \{X_1, \dots, X_n\}| \leq \sum_{i=0}^{\operatorname{vc}(\mathscr{C})} \binom{n}{i} \leq \left(\frac{\mathrm{e} n}{\operatorname{vc}(\mathscr{C})} \right)^{\operatorname{vc}(\mathscr{C})}.$$

The proof of Theorem 17.7 relies on a lemma that relates the cardinality of a set family to the number of sets it shatters.

Lemma 17.8 (Pajor). Suppose that S is a finite set, and let $\mathscr{F}$ be a collection of subsets of S. Then

$$|\mathscr{F}| \leq |\{\mathsf{I} \subseteq \mathsf{S} : \mathscr{F} \text{ shatters } \mathsf{I}\}|$$

We will prove this lemma in a moment, but first let us use it to establish the main result.

Proof of Theorem 17.7 from Lemma 17.8. Consider the finite set $\mathsf{S} = \Omega \cap \{X_1, \dots, X_n\}$ with cardinality n. Apply Pajor's lemma to the class $\mathscr{F} = \mathscr{C} \cap \{X_1, \dots, X_n\}$ to obtain

$$\begin{aligned} |\mathscr{C} \cap \{X_1, \dots, X_n\}| &\leq |\{\mathsf{I} \subseteq \mathsf{S} : \mathscr{C} \text{ shatters } \mathsf{I}\}| \\ &\leq |\{\mathsf{I} \subseteq \mathsf{S} : |\mathsf{I}| \leq \operatorname{vc}(\mathscr{C})\}| = \sum_{i=0}^{\operatorname{vc}(\mathscr{C})} \binom{n}{i}. \end{aligned}$$

The second inequality holds because the VC dimension is the *maximum cardinality* of a shattered set, and we have included every subset of S with cardinality $\operatorname{vc}(\mathscr{C})$ or less. Finally, we count these subsets in the usual manner. ∎

17.3.1 Proof of Pajor's lemma

The proof of Lemma 17.8 uses induction on the cardinality $n = |\mathsf{S}|$ of the base set. For $n = 1$, then $\mathsf{S} = \{x\}$ for some element x. The only possibilities for the class $\mathscr{F}$ are $\{\varnothing\}$ or $\{\{x\}\}$ or $\{\varnothing, \{x\}\}$. By convention, every set class shatters $\varnothing$, so it is clear that Pajor's lemma holds for every collection $\mathscr{F}$ of subsets of $\mathsf{S} = \{x\}$.

Now assume the conclusion of the lemma holds for each base set containing exactly n points. Consider a base set S containing $n+1$ points and a class $\mathscr{F}$ of subsets of S. Partition the base set $\mathsf{S} = \mathsf{T} \dot\cup \{x_0\}$ where $x_0 \in \mathsf{S}$ is a fixed (but arbitrary) element. Partition the class $\mathscr{F} = \mathscr{F}_0 \dot\cup \mathscr{F}_1$, where

The notation $\dot\cup$ indicates a union of *disjoint* sets.

$$\mathscr{F}_0 = \{\mathsf{F} \in \mathscr{F} : x_0 \in \mathsf{F}\};$$
$$\mathscr{F}_1 = \{\mathsf{F} \in \mathscr{F} : x_0 \notin \mathsf{F}\}.$$

Apply the induction hypothesis to $\mathscr{F}_0$ and $\mathscr{F}_1$. To be concise, we introduce notation for the number of distinct subsets of S shattered by $\mathscr{F}$:

$$\mathrm{sh}(\mathsf{S} \,|\, \mathscr{F}) := |\{\mathsf{I} \subseteq \mathsf{S} : \mathscr{F} \text{ shatters } \mathsf{I}\}|.$$

We now claim that

$$|\mathscr{F}_0| = |\mathscr{F}_0 \cap \mathsf{T}| \le \mathrm{sh}(\mathsf{S} \,|\, \mathscr{F}_0 \cap \mathsf{T}) = \mathrm{sh}(\mathsf{S} \,|\, \mathscr{F}_0);$$
$$|\mathscr{F}_1| = |\mathscr{F}_1 \cap \mathsf{T}| \le \mathrm{sh}(\mathsf{S} \,|\, \mathscr{F}_1 \cap \mathsf{T}) = \mathrm{sh}(\mathsf{S} \,|\, \mathscr{F}_1).$$

Moving left to right, these (in)equalities follow from the definition of $\mathscr{F}_0$ and $\mathscr{F}_1$ and the induction hypothesis. Last, we use the fact that, for $x_0 \notin \mathsf{I}$, the class $\mathscr{F}$ shatters I if and only if $\mathscr{F} \cup \{x_0\}$ shatters I. We therefore have

$$|\mathscr{F}| \;=\; |\mathscr{F}_0| + |\mathscr{F}_1| \;\le\; \mathrm{sh}(\mathsf{S} \,|\, \mathscr{F}_0) + \mathrm{sh}(\mathsf{S} \,|\, \mathscr{F}_1).$$

We need to show that the right-hand side is bounded above by $\mathrm{sh}(\mathsf{S} \,|\, \mathscr{F})$.

Observe that

$$\begin{aligned}\mathrm{sh}(\mathsf{S} \,|\, \mathscr{F}_0) + \mathrm{sh}(\mathsf{S} \,|\, \mathscr{F}_1) = {} & |\{\mathsf{I} \text{ shattered by } \mathscr{F}_0 \text{ exclusively}\}| \\ & + |\{\mathsf{I} \text{ shattered by } \mathscr{F}_1 \text{ exclusively}\}| \\ & + 2 \cdot |\{\mathsf{I} \text{ shattered by } \mathscr{F}_0 \textit{ and } \mathscr{F}_1\}|.\end{aligned}$$

Sets that are shattered by both classes are double-counted, so we need to make sure that the $\mathscr{F}$ shatters another set in S that was not shattered by the subclasses. Now, observe that

$$\begin{aligned}\mathrm{sh}(\mathsf{S} \,|\, \mathscr{F}) = {} & |\{\mathsf{I} \text{ shattered by } \mathscr{F}_0 \text{ exclusively}\}| \\ & + |\{\mathsf{I} \text{ shattered by } \mathscr{F}_1 \text{ exclusively}\}| \\ & + |\{\mathsf{I} \text{ shattered by } \mathscr{F}_0 \text{ and } \mathscr{F}_1\}| \\ & + |\{\mathsf{I} \text{ shattered by neither } \mathscr{F}_0 \text{ nor } \mathscr{F}_1\}|.\end{aligned}$$

Therefore, we can finish the proof by showing that

$$|\{\mathsf{I} \text{ shattered by } \mathscr{F}_0 \text{ and } \mathscr{F}_1\}| \le |\{\mathsf{I} \text{ shattered by neither } \mathscr{F}_0 \text{ nor } \mathscr{F}_1\}|.$$

This is what we will do.

Suppose that a set I is shattered by both $\mathscr{F}_0$ and $\mathscr{F}_1$. Then the distinguished point $x_0 \notin \mathsf{I}$, because neither subclass can shatter a set containing x_0. Nevertheless, the augmented set $\mathsf{I} \cup \{x_0\}$ is shattered by the full class $\mathscr{F}$ because

$$\mathscr{F}_0 \cap (\mathsf{I} \cup \{x_0\}) = \{\mathsf{V} \cup \{x_0\} : \mathsf{V} \subseteq \mathsf{I}\};$$
$$\mathscr{F}_1 \cap (\mathsf{I} \cup \{x_0\}) = \mathscr{P}(\mathsf{I}).$$

Together, these two classes of subsets compose the power set $\mathcal{P}(\mathsf{I} \cup \{x_0\})$. Therefore, we have exhibited an injection $\mathsf{I} \mapsto \mathsf{I} \cup \{x_0\}$ for which

$$\{\mathsf{I} \text{ shattered by } \mathcal{F}_0 \text{ and } \mathcal{F}_1\} \hookrightarrow \{\mathsf{I} \text{ shattered by neither } \mathcal{F}_0 \text{ nor } \mathcal{F}_1\}.$$

The proof is complete.

18. Statistical Learning

Date: 4 March 2021 **Scribe:** Chi-Fang Chen

This lecture contains an introduction to the theory of statistical learning, which is an interesting application of empirical processes. We will see how the tools we have been developing in the recent lectures help us understand how much data is sufficient to solve certain idealized learning problems.

Agenda:
1. Set-up for statistical Learning
2. Risk + empirical risk
3. Classifiers and VC dimension
4. Approximation by Lipschitz functions

18.1 Supervised learning

We will study the problem using labeled data to make predictions about the labels of future observations. This is called a *supervised learning* problem.

18.1.1 Setup

Consider a measurable domain Ω, equipped with an (unknown) probability measure μ that describes the distribution of the population. Let $T : \Omega \to \mathbb{R}$ be an unknown *target function* that we would like to learn from observations.

Suppose that we collect an iid sample from the population measure μ, along with the observed values of the target function:

$$((X_i, T(X_i)) : i = 1, \dots, n) \quad \text{where } X_i \sim \mu \text{ iid.}$$

This is called *training data*. Given a new sample $X \sim \mu$ from the population, our task is to predict the value $T(X)$ of the target function.

A solution to the supervised learning problem is a function $f : \Omega \to \mathbb{R}$. Heuristically, we would like to ensure that

$$f(X) \approx T(X) \quad \text{for } X \sim \mu.$$

Note that this is a statistical question because we are assuming that future samples are drawn at random from the population; we are not trying to to obtain worst-case bounds. An instance of this problem is illustrated in Figure 18.1

> **Aside: (Ideal models).** The supervised learning model we are considering is a very simple one, and it is not necessarily realistic. In particular, we will assume that the training labels $T(X_i)$ are determined by a function T, and there is no statistical error in the observations. Beyond that, major challenges in contemporary machine learning include *transfer learning* and *data shift*. Roughly, these issues arise when we have to make predictions for sample points that are drawn from a different distribution from the training samples.

> **Aside: (Unsupervised learning).** The problems we are studying are *supervised* because the training samples are labeled with the values of the target function. In contrast, an *unsupervised learning* problem involves only the samples (X_i) from

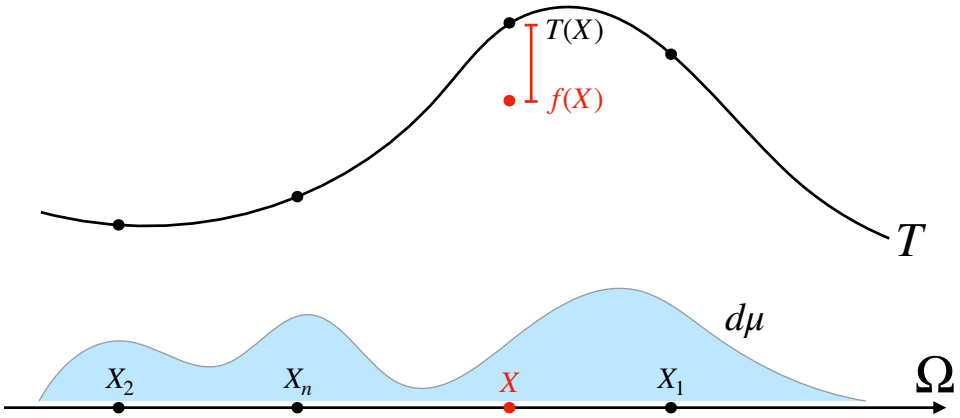


Figure 18.1 (Approximation). Given labeled training data $(X_i, T(X_i))$ for $i = 1, \ldots, n$, we want to find a hypothesis f so that $f(X) \approx T(X)$ for a random point $X \sim \mu$.

> the population, in which case our task is to approximate the distribution or find cross-cutting features that explain the variability in the sample.

18.1.2 Risk and empirical risk

To formulate the supervised learning problem mathematically, we will use the language of optimization. The first step is to quantify how well a given function $f : \Omega \to \mathbb{R}$ approximates the target function $T : \Omega \to \mathbb{R}$ on average over samples from the population.

Definition 18.1 (Risk). Given a hypothesis $f : \Omega \to \mathbb{R}$ and a target function $T : \Omega \to \mathbb{R}$, the *risk* $R(f)$ of the hypothesis is

$$R(f) := \mathbb{E}_{X \sim \mu}\left[(f(X) - T(X))^2\right] = \mu(f - T)^2. \tag{18.1}$$

The risk is also called the L_2 *loss*.

We cannot evaluate the risk directly because we do not have access to the population measure μ or the target function T. Instead, we would like to develop a proxy for the risk that we can compute from the observed training data. This quantity is called the *empirical risk*.

Definition 18.2 (Empirical risk). Given labeled data $((X_i, T(X_i)) : i = 1, \ldots, n)$ and a hypothesis function $f : \Omega \to \mathbb{R}$, the *empirical risk* is

$$R_n(f) := \frac{1}{n} \sum_i^n \left[(f(X_i) - T(X_i))^2\right]^2 = \mu_n(f - T)^2. \tag{18.2}$$

As usual, $\mu_n := \frac{1}{n}\delta_{X_i}$ is the empirical measure associated with the sample. The empirical risk is also called the *empirical* L_2 *loss*.

As in our study of empirical processes, the basic question is the extent to which the empirical measure μ_n can stand in for the population measure μ when we try to find a

hypothesis with low risk.

18.1.3 Hypothesis classes and risk minimization

We would like to quantify how many labeled samples $(X_i, T(X_i))$ we need to identify a hypothesis function f with low (empirical) risk. The answer to this question depends on the hypotheses that we are allowed to consider.

Instead of searching for an arbitrary hypothesis function $f : \Omega \to \mathbb{R}$ on the domain, we will restrict our attention to a restricted class $\mathcal{F}$ of hypothesis functions. From this class, we seek a hypothesis that minimizes the risk:

$$\begin{aligned} f^\star &= \arg\min_{f \in \mathcal{F}} R(f) \\ &= \arg\min_{f \in \mathcal{F}} \mathbb{E}_{X \sim \mu} \left[(f(X) - T(X))^2 \right]. \end{aligned} \tag{18.3}$$

This optimization problem is called *risk minimization*. If the target function T does not belong to the hypothesis class, then the minimum risk may be strictly positive. This is called the *misspecification error* that arises from the mismatch between the target and the models we are considering.

In practice, we cannot solve the risk minimization problem (18.3) because we cannot compute the risk from the observed training data. Instead, we will consider the associated empirical problem:

$$\begin{aligned} f_n^\star &= \arg\min_{f \in \mathcal{F}} R_n(f) \\ &= \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \left[(f(X_i) - T(X_i))^2 \right]. \end{aligned} \tag{18.4}$$

This optimization problem is called *empirical risk minimization* (ERM). The idea that (18.4) serves as a proxy for (18.3) is called the *empirical risk minimization principle*.

Of course, we are not interested in the empirical risk associated with a hypothesis but rather the true risk. Minimizing the empirical risk always yields a hypothesis with higher risk than the true optimizer $f^\star$.

Definition 18.3 (Excess risk). The *excess risk* associated with empirical risk minimization is the quantity

$$E_n := R(f_n^\star) - R(f^\star) \geq 0,$$

where $f_n^\star$ is a solution to (18.4) and $f^\star$ is a solution to (18.3).

The excess risk is the additional risk that we incur by learning a hypothesis from a finite number n of samples. The excess risk accrues above the misspecification error associated with the hypothesis class $\mathcal{F}$, which cannot be avoided.

Although we cannot evaluate the excess risk from observed data, we can still establish bounds on the excess risk that tell us how many samples n suffice to achieve a given level of excess risk. This problem is the focus of today's lecture.

18.1.4 Excess risk and empirical processes

We can control the excess risk by studying the supremum of an empirical process. This fact allows us to use the tools we have developed in the previous lectures.

Proposition 18.4 (Excess risk). Instate the prevailing notation. The excess risk of a

function $f_n^\star$ computed using the ERM principle satisfies

$$\begin{aligned} R(f_n^\star) - R(f^\star) &\le 2 \sup_{f \in \mathcal{F}} |R_n(f) - R(f)| \\ &= 2 \sup_{f \in \mathcal{F}} |\mu_n (f - T)^2 - \mu (f - T)^2|. \end{aligned}$$

Proof. Introduce the quantity $\Delta := \sup_{f \in \mathcal{F}} |R_n(f) - R(f)|$. By adding and subtracting $R_n(f_n^\star)$, we may calculate that

$$R(f_n^\star) \le R_n(f_n^\star) + \Delta \le R_n(f^\star) + \Delta \le R(f^\star) + 2\Delta.$$

The second bound relies on the fact that $f_n^\star$ minimizes the empirical risk R_n. To reach the last bound, we added and subtracted $R(f^\star)$. ∎

18.1.5 Discussion

Observe that the total risk associated with a learned hypothesis $f_n^\star$ can be decomposed into two terms:

$$R(f_n^\star) = R(f^\star) + E_n(f_n^\star).$$

The first term $R(f^\star)$ is the misspecification error associated with the hypothesis class $\mathcal{F}$, while the second term $E_n(f_n^\star)$ is the excess risk that we incur by learning from a finite amount of data.

This decomposition exposes a tradeoff. Although we can reduce the misspecification error by choosing a large hypothesis class $\mathcal{F}$, the excess risk typically increases when the hypothesis class is large. Indeed, for large hypothesis classes, we may need to collect a lot more data before the ERM problem (18.4) is a good proxy for the risk minimization problem (18.3). We can easily *overfit* the observed data if we are allowed to use a rich class of hypothesis, so we wind up modeling variability in the sample (X_i) rather than the target function T itself. This problem manifests in *poor generalization*, where the computed hypothesis $f_n^\star(X)$ is an ineffective model for the target $T(X)$ when X is a sample drawn from the population measure μ.

There may be further computational reasons for preferring particular hypothesis classes $\mathcal{F}$. In practice, we have to solve the ERM problem (18.4) numerically, and this optimization may be more tractable for some function classes than others. In this course, we focus only on the probabilistic aspects.

18.2 Classification

In this section, we discuss the problem of learning a binary classifier from labeled data.

18.2.1 Classification problems

As usual, suppose that Ω is a measurable domain equipped with a probability measure μ. Consider a Boolean function $T : \Omega \to \{0, 1\}$ on the domain. We can interpret the target function T as an assignment of each point in the domain into two exclusive categories (e.g., cat/dog).

A *classifier* is a Boolean function $f : \Omega \to \{0, 1\}$ that predicts the category of a point. The risk of a classifier equals the probability of misclassification:

$$R(f) = \mathbb{P}_{X \sim \mu} \{ f(X) \neq T(X) \}.$$

We specify the collection $\mathcal{F}$ of classifiers that we are willing to consider. Rich collections of classifiers may carve out complicated decision boundaries that delineate the two

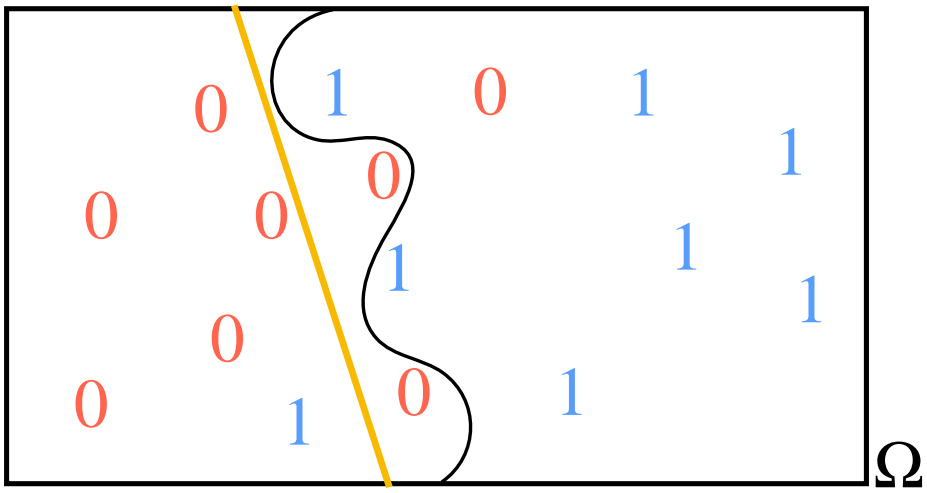


Figure 18.2 (Classification problem). Let $T : \Omega \to \{0, 1\}$ be a Boolean function. The task is to predict the value $T(X)$ at a new sample $X \sim \mu$. A classifier $f : \Omega \to \{0, 1\}$ is a rule for assigning a point to a category; the decision boundary (black squiggle) is the line between the two categories. It is common to use linear classifiers where the decision boundary is a straight line (yellow).

categories. On the other hand, we need more data to fit complicated decision boundaries, and we are more prone to overfitting and poor generalization.

Given a collection $\{(X_i, T(X_i)) : i = 1, \dots, n\}$ of labeled samples, we would like to learn a classifier. We can accomplish this goal by empirical risk minimization (18.4) over the family $\mathcal{F}$ of admissible classifiers. See Figure 18.2 for an illustration of the classification problem.

Example 18.5 (Linear classifier). Consider the domain $\Omega = \mathbb{R}^d$. A *linear classifier* is the indicator function of a half-space in $\mathbb{R}^d$. The family $\mathcal{F}$ of linear classifiers takes the form

$$\mathcal{F} = \left\{ \mathbb{1}\{\boldsymbol{x} \in \mathbb{R}^d : \langle \boldsymbol{a}, \boldsymbol{x} \rangle \le b\} : \boldsymbol{a} \in \mathbb{R}^d, b \in \mathbb{R} \right\}.$$

The decision boundaries are hyperplanes in $\mathbb{R}^d$. See Figure 18.2 for an example. ▪

18.2.2 Excess risk bound

We will show that the excess risk of a classifier is controlled by the VC dimension of the family $\mathcal{F}$ of admissible classifiers.

Theorem 18.6 (Excess risk of a classifier: VC dimension bound). Instate the prevailing notation. The excess risk of the classifier $f_n^\star$ computed using ERM (18.4) over a family $\mathcal{F}$ of Boolean hypothesis satisfies

$$\mathbb{E}[R(f_n^\star) - R(f^\star)] \le \text{Const} \sqrt{\frac{\text{vc}(\mathcal{F})}{n}}.$$

The expectation averages over random samples $(X_i : i = 1, \dots, n)$. In particular, $n \approx \varepsilon^{-2}\mathrm{vc}(\mathscr{F})$ samples typically suffice to achieve excess risk ε.

This result is uniform over all population measures μ and all target classifications T. It tells us that the amount of data we need to fit a classifier from a given family $\mathscr{F}$ is controlled by the combinatorial dimension of $\mathscr{F}$, which reflects the complexity of the decision boundaries that it can generate. This result, however, provides no information about the misspecification error associated with the class $\mathscr{F}$.

18.2.3 Example: Linear classifiers

Let us see how Theorem 18.6 applies to the case of a linear classifier on $\Omega = \mathbb{R}^d$. Recall that

$$\mathscr{F} = \left\{\mathbb{1}\{\boldsymbol{x} \in \mathbb{R}^d : \langle \boldsymbol{a}, \boldsymbol{x}\rangle \le b\} : \boldsymbol{a} \in \mathbb{R}^d, b \in \mathbb{R}\right\}.$$

We claim that $\mathrm{vc}(\mathscr{F}) = d + 1$. Therefore, it suffices to draw $n = O(\varepsilon^{-2}d)$ samples to obtain a linear classifier with excess risk at most ε.

To establish the claim, we first exhibit a set X of $d + 1$ points in $\mathbb{R}^d$ that we can shatter using the class $\mathscr{F}$. These points compose the vertices of a simplex:

$$\mathsf{X} = \{\mathbf{0}, \mathbf{e}_1, \cdots, \mathbf{e}_d\} \subset \mathbb{R}^d.$$

Second, we need to argue that it is impossible to shatter a set of $d + 2$ points using the family of half-spaces. This statement follows immediately from Radon's theorem.

Fact 18.7 (Radon's theorem). For any set X of $d + 2$ points in $\mathbb{R}^d$, there exist disjoint subsets R and B whose convex hulls have a nontrivial intersection:

$$\mathrm{conv}(\mathsf{R}) \cap \mathrm{conv}(\mathsf{B}) \neq \varnothing.$$

In particular, we cannot isolate R from B using a half-space. ∎

18.2.4 Covering numbers and squared loss

In view of Proposition 18.4, it is no surprise that the proof of Theorem 18.6 uses tools from empirical process theory. We require a lemma that allows us to control the covering numbers of the squared errors associated with a family $\mathscr{F}$ of classifiers in terms of the covering numbers of the family $\mathscr{F}$.

Lemma 18.8 Let $\mathscr{F}$ be a family of Boolean functions. For a Boolean target function T, consider the family of (Boolean) squared error functions:

$$\mathscr{L} = \{(f - T)^2 : f \in \mathscr{F}\}.$$

Then

$$\mathrm{N}(\mathscr{L}, \mathsf{L}_2(\mu_n); \varepsilon) \le \mathrm{N}(\mathscr{F}, \mathsf{L}_2(\mu_n); \varepsilon).$$

Proof. Consider a minimal ε-net $\{f_1, \dots, f_m\}$ for the metric space $(\mathscr{F}, \mathsf{L}_2(\mu_n))$. That is,

$$\min_i \|f - f_i\|_{\mathsf{L}_2(\mu_n)} \le \varepsilon \quad \text{for all } f \in \mathscr{F}.$$

Since the functions are Boolean, $(f - T)^2 = |f - T|$ for each $f \in \mathscr{F}$. The triangle inequality yields the (pointwise) bound

$$\left|(f - T)^2 - (f_i - T)^2\right| = \left||f - T| - |f_i - T|\right| \le |f - f_i|.$$

Thus, the shifted functions $\{f_i - T : i = 1, \dots, m\}$ provide an ε-net for $\mathscr{L}$. That is,

$$\min_i \|(f - T)^2 - (f_i - T)^2\|_{\mathsf{L}_2(\mu_n)} \leq \min_i \|f - f_i\|_{\mathsf{L}_2(\mu_n)}.$$

We have used the fact that the $\mathsf{L}_2(\mu_n)$ norm is monotone. ∎

18.2.5 Proof of Theorem 18.6

We are now prepared to establish the excess risk bound for a classifier computed via the ERM principle. Using Proposition 18.4, we may calculate that

$$\begin{aligned}\mathbb{E}[R(f_n^\star) - R(f^\star)] &\leq 2\,\mathbb{E} \sup_{f \in \mathscr{F}} |R_n(f) - R(f)| \\ &= 2\,\mathbb{E} \sup_{f \in \mathscr{F}} |\mu_n(f - T)^2 - \mu(f - T)^2|] \\ &= 2\,\mathbb{E} \sup_{g \in \mathscr{L}} |\mu_n(g) - \mu(g)|].\end{aligned}$$

We have defined the class $\mathscr{L}$ as in the statement of Lemma 18.8. To continue, we invoke the bound for an empirical process (Proposition 16.6):

$$\begin{aligned}\mathbb{E}[R(f_n^\star) - R(f^\star)] &\leq \frac{\text{Const}}{\sqrt{n}}\,\mathbb{E}_{(X_i)} \int_0^1 \mathrm{d}\varepsilon \sqrt{\log(\mathrm{N}(\mathscr{L}, \mathsf{L}_2(\mu_n); \varepsilon))} \\ &\leq \frac{\text{Const}}{\sqrt{n}}\,\mathbb{E}_{(X_i)} \int_0^1 \mathrm{d}\varepsilon \sqrt{\log(\mathrm{N}(\mathscr{F}, \mathsf{L}_2(\mu_n); \varepsilon))} \\ &\leq \frac{\text{Const}}{\sqrt{n}} \sqrt{\mathrm{vc}(\mathscr{F})}.\end{aligned}$$

The second inequality is Lemma 18.8. The last estimate follows from Dudley's covering number bound (Theorem 17.1).

18.3 A simple approximation problem

Another basic question in statistical learning is to approximate the target function. In this section, we will study a simple instance that we can treat with available results.

Consider the unit interval $\Omega = [0, 1]$ in the real line, equipped with a probability measure μ. Suppose that $T : [0, 1] \to [0, 1]$ is a target function that we wish to approximate. Our goal is to use a labeled sample $((X_i, T(X_i)) : i = 1, \dots, n)$ of training data to find a hypothesis function $f : [0, 1] \to [0, 1]$ that approximates the target function T well on average with respect to the measure μ. Here, the risk is simply the squared loss:

$$R(f) = \mathbb{E}_{X \sim \mu} \left[(f(X) - T(X))^2\right].$$

See Figure 18.1.

We restrict our attention to a small class of hypothesis functions:

$$\mathscr{F} = \{f : [0, 1] \to [0, 1] : f \text{ is } L\text{-Lipschitz}\}.$$

We can apply the empirical risk minimization principle to select a hypothesis $f_n^\star$ from the class $\mathscr{F}$. We will prove that a finite sample suffices to make the excess risk as small as we like. The misspecification error depends on how well we can approximate the target T by means of an L-Lipschitz function. We have the following result.

Theorem 18.9 (Excess risk for Lipschitz approximation). Instate the prevailing notation. The excess risk of the approximation $f_n^\star$ computed using ERM (18.4) over the family $\mathcal{F}$ of 1-Lipschitz functions on $[0,1]$ satisfies

$$\mathbb{E}[R(f_n^\star) - R(f^\star)] \le \text{Const}\sqrt{\frac{L}{n}}.$$

In particular, $n = O(\varepsilon^{-2}L)$ samples suffice to achieve an excess risk at most ε.

Notice the tradeoff here. As we dial up the Lipschitz constant L, the hypothesis class enlarges, and the misspecification error decreases. But we also need more samples to control the excess risk and produce a hypothesis that is competitive with the best L-Lipschitz model.

Proof. According to Proposition 18.4, we can control the excess risk using an empirical process:

$$\begin{aligned}\mathbb{E}[R(f_n^\star) - R(f^\star)] &\le 2\,\mathbb{E}\sup_{f\in\mathcal{F}} |R_n(f) - R(f)| \\ &= 2\,\mathbb{E}\sup_{f\in\mathcal{F}} |\mu_n(f-T)^2 - \mu(f-T)^2| \\ &\le \frac{4}{\sqrt{n}}\,\mathbb{E}\sup_{f\in\mathcal{F}} \left|\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^n \varepsilon_i (f(X_i) - T(X_i))^2\right|.\end{aligned}$$

The last inequality follows from symmetrization (Proposition 16.5).

For functions $f, g \in \mathcal{F}$, we have the pointwise bound

$$\left|(f-T)^2 - (g-T)^2\right| = (f+g-2T)(f-g) \le 2|f-g| \le 2\|f-g\|_\infty.$$

We have used the fact that f, g, T take values in $[0,1]$. Using Hoeffding's inequality, we see that the increments of the (conditioned) Rademacher process are subgaussian with variance proxy $4\|f-g\|_\infty^2$.

To continue, we invoke Dudley's chaining inequality (Theorem 12.1) to bound the supremum of the (conditioned) Rademacher process:

$$\begin{aligned}\mathbb{E}[R(f_n^\star) - R(f^\star)] &\le \frac{\text{Const}}{\sqrt{n}} \int_0^1 \mathrm{d}\varepsilon \sqrt{\log \mathrm{N}(\mathcal{F}, \|\cdot\|_\infty; \varepsilon)} \\ &\le \frac{\text{Const}}{\sqrt{n}} \int_0^1 \mathrm{d}\varepsilon \sqrt{\text{Const}\cdot L/\varepsilon} = \text{Const}\sqrt{\frac{L}{n}}.\end{aligned}$$

We have used the covering number bound for L-Lipschitz functions on $[0,1]$ with respect to $\|\cdot\|_\infty$; this result follows from the argument in Proposition 15.3. ∎

19. Positive Empirical Processes

Date: 9 March 2021 **Scribe:** Hsin-Yuan Huang

Agenda:
1. Definition and setup
2. Example: singular values
3. Small ball method
4. Minimum singular value

This section of the course has focused on empirical processes, a valuable formalism that allows us to study a variety of problems in statistical learning. We have used empirical processes to prove the following results:

- **Uniform law of large numbers.** This result shows that one can uniformly integrate an entire class of Lipschitz functions in a bounded interval using a fixed set of sampled points. See Lecture 15.
- **Uniform Glivenko–Cantelli.** A fixed set of sampled points can be used to approximate the probabilities for a large class of events. The number of the sampled points depends on the VC dimension of the class of events. See Lecture 17.
- **Empirical risk minimization.** Using observed data, one can learn a classifier by minimizing the error on the observed data without access the true population. See Lecture 18.

In this lecture, we will consider empirical processes indexed by a function class that contains only positive functions. We will develop a technique, called *the small ball method*, for bounding the infimum of a positive empirical process away from zero. As a particular example, we will apply the small ball method to study the minimum singular value in a random matrix under very mild conditions.

19.1 Setup

Let us recall the definition of empirical processes. Consider a measurable domain Ω equipped with a probability measure μ. Draw n independent samples $X_1, \dots, X_n$ from the probability measure μ, and construct the empirical measure

$$\mu_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

For a function $f : \Omega \to \mathbb{R}$, recall that

$$\mu(f) := \int_{X \in \Omega} f(X) \, \mathrm{d}\mu,$$

the expectation value of f with respect to the probability measure μ. For an event $\mathsf{C} \subseteq \Omega$, we also define the notation

$$\mu(\mathsf{C}) := \int_{X \in \Omega} \mathbb{1}\{X \in \mathsf{C}\} \, \mathrm{d}\mu,$$

the probability of the event C under the probability measure μ.

Now, consider a class $\mathcal{F}$ that contains functions from Ω to $\mathbb{R}$. Previous lectures have focused on the uniform empirical error

$$\mathbb{E} \sup_{f \in \mathcal{F}} |\mu_n(f) - \mu(f)|,$$

which reflects the deviation of the empirical measure μ_n from the true measure μ. One may also consider the extreme values of the empirical moments:

$$\mathbb{E}\inf_{f\in\mathcal{F}}\mu_n(f) \quad \text{and} \quad \mathbb{E}\sup_{f\in\mathcal{F}}\mu_n(f).$$

These quantities reflect the minimum and maximum expectation of a function f from the class $\mathcal{F}$ under the empirical measure. It is most common to seek lower bounds for the infimum and upper bounds for the supremum.

In this lecture, we will be interested in a particular problem that arises in a range of applications. Consider a class $\mathcal{F}$ that only contains *positive* functions $f : \Omega \to \mathbb{R}_+$. Then the empirical moments $\mu_n(f)$ are certainly positive as well. We will be interested in methods for proving that the infimum of a positive empirical process is *bounded away from zero*.

19.2 Example: Singular values of random matrices

To motivate the study of positive empirical processes, we first give an example that explains how this kind of problem arises. Let us see how to express the extreme singular values of a random matrix with independent rows using this formalism.

Consider a probability measure μ on the d-dimensional Euclidean space $\mathbb{R}^d$. Suppose we have sampled n independent random vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ according to the probability measure μ. We may form the random matrix

$$\boldsymbol{X} = \begin{bmatrix} \text{——} & \boldsymbol{x}_1^* & \text{——} \\ & \vdots & \\ \text{——} & \boldsymbol{x}_n^* & \text{——} \end{bmatrix} \in \mathbb{R}^{n\times d}.$$

The extremum singular values for the matrix $\boldsymbol{X}$ are given by the expressions

$$\sigma_{\min}^2(\boldsymbol{X}) := \inf_{\|\boldsymbol{u}\|_2=1} \|\boldsymbol{X}\boldsymbol{u}\|_2^2;$$
$$\sigma_{\max}^2(\boldsymbol{X}) := \sup_{\|\boldsymbol{u}\|_2=1} \|\boldsymbol{X}\boldsymbol{u}\|_2^2.$$

These quantities express how much the random matrix can contract or expand a unit vector.

It is not immediately clear how one could relate the extremum singular values to empirical processes. In order to do so, observe that

$$\|\boldsymbol{X}\boldsymbol{u}\|_2^2 = \sum_{i=1}^n \langle \boldsymbol{x}_i, \boldsymbol{u}\rangle^2. \tag{19.1}$$

We may define some positive functions, parameterized by unit vectors:

$$f_{\boldsymbol{u}}(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{u}\rangle^2 \quad \text{where } \|\boldsymbol{u}\|_2 = 1.$$

Using (19.1), we immediately realize that

$$\|\boldsymbol{X}\boldsymbol{u}\|_2^2 = \sum_{i=1}^n f_{\boldsymbol{u}}(\boldsymbol{x}_i).$$

Introduce the function class

$$\mathcal{F} = \{f_{\boldsymbol{u}} : \|\boldsymbol{u}\|_2 = 1\}.$$

Then

$$\sigma_{\min}^2(\boldsymbol{X}) = \inf_{f\in\mathcal{F}} \sum_{i=1}^n f(\boldsymbol{x}_i) = n \cdot \inf_{f\in\mathcal{F}} \mu_n(f);$$
$$\sigma_{\max}^2(\boldsymbol{X}) = \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(\boldsymbol{x}_i) = n \cdot \sup_{f\in\mathcal{F}} \mu_n(f).$$

Recall that the empirical measure μ_n is normalized by $1/n$.

In other words, we can control the extreme singular values of a random matrix with iid rows using bounds for positive empirical process.

In many applications, lower bounds for the minimum singular value play a more significant role than upper bounds for the maximum singular value. Indeed, the minimum singular value tells us how well the random matrix $\boldsymbol{X}$ preserves *separation* between vectors because

$$\|\boldsymbol{X}(\boldsymbol{u} - \boldsymbol{v})\|_2 \geq \sigma_{\min}(\boldsymbol{X}) \cdot \|\boldsymbol{u} - \boldsymbol{v}\|_2.$$

In particular, $\operatorname{null}(\boldsymbol{X}) = \{\boldsymbol{0}\}$ if and only if $\sigma_{\min}(\boldsymbol{X}) > 0$. So the minimum singular value signals whether the matrix $\boldsymbol{X}$ is an injection. Of course, $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ can be an injection only if $n \geq d$, so we will focus on this parameter regime.

In Section 19.6, we will sketch some applications in dimension reduction and signal processing.

19.3 Extrema via centering

Before we present the small ball method, let us describe some obvious approaches to bounding the extreme values of an empirical process so that we can see why they do not work.

Consider a general empirical process $(\mu_n(f) : f \in \mathcal{F})$, not necessarily positive. We can attempt to bound the supremum or the infimum by decomposing the process into its mean and its deviations:

$$\mathbb{E}\sup_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i) \leq \sup_{f\in\mathcal{F}} \sum_{i=1}^n \mathbb{E} f(X_i) + \mathbb{E}\sup_{f\in\mathcal{F}} \sum_{i=1}^n [f(X_i) - \mathbb{E} f(X_i)].$$

Likewise,

$$\mathbb{E}\inf_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i) \geq \inf_{f\in\mathcal{F}} \sum_{i=1}^n \mathbb{E} f(X_i) - \mathbb{E}\sup_{f\in\mathcal{F}} \sum_{i=1}^n [\mathbb{E} f(X_i) - f(X_i)].$$

It is often straightforward to compute $\mathbb{E} f(X_i)$. We can apply tools from the previous lectures to try to control the deviation term.

Unfortunately, our methods for controlling the uniform error in an empirical process only work well for very nice classes of functions. Indeed, we have focused exclusively on the case where $\mathcal{F}$ contains *bounded* functions. These arguments can be extended to address subgaussian functions, but it is already a difficult matter to obtain bounds for classes of subexponential functions, let alone functions without exponential moments.

19.4 The small ball method

In this section, we develop the simplest version of the small ball method for bounding the infimum of a positive empirical process away from zero.

19.4.1 Motivation

From now on, assume that $\mathcal{F}$ is a class of *positive* functions. We consider the positive empirical process

$$\sum_{i=1}^{n} f(X_i) \quad \text{for } f \in \mathcal{F}.$$

We would like to study the extreme values of this process, making as few assumptions as possible about the function class. For instance, we might consider functions that have only four finite moments: $\mu(f^4) < +\infty$.

First, consider the expected supremum:

$$\mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f(X_i).$$

Since the functions f are positive, the terms in the sum accumulate. To obtain a good upper bound on the supremum, we must ensure that *none* of the summands is large. For this reason, it may not be possible to obtain good upper bounds for functions with few moments.

In contrast, consider the expected infimum:

$$\mathbb{E} \inf_{f \in \mathcal{F}} \sum_{i=1}^{n} f(X_i).$$

The effects that make it hard to control the supremum now work to our benefit. If *a single term* in the sum is large, then the whole sum is also large. Therefore, it should be possible to bound the infimum away from zero under mild assumptions. The small ball method is a particular way of exploiting this intuition.

19.4.2 Main idea

A first version of the small ball method was proposed by Mendelson [Men14a] to address a problem in convex geometry. The approach has been refined and applied to problems in statistical learning [Men14b], random matrix theory [KM15], and signal processing [Tro15b].

For a positive empirical process, our goal is to establish that

$$\mathbb{E} \inf_{f \in \mathcal{F}} \sum_{i=1}^{n} f(X_i) \gg 0.$$

The main idea of this method is to *count* how many terms in the sum exceed a certain threshold τ:

$$|\{i : f(X_i) \geq \tau\}| = \sum_{i=1}^{n} \mathbb{1}\{f(X_i) \geq \tau\}.$$

In many cases, it is enough to show that there are a few large terms on average. Since indicator functions are bounded, we can easily apply methods we have developed for empirical processes to control this sum.

19.4.3 Step 1: Reduction to counting

Fix a level $\tau > 0$. By Markov's inequality,

$$\sum_{i=1}^{n} f(X_i) \geq \tau \cdot |\{i : f(X_i) \geq \tau\}|.$$

This bound holds for any *positive* function f and any realization (X_i) of the random sample. Thus, we can bound the expected infimum below by

$$\mathbb{E} \inf_{f \in \mathcal{F}} \sum_{i=1}^{n} f(X_i) \geq \tau \cdot \mathbb{E} \inf_{f \in \mathcal{F}} \sum_{i=1}^{n} \mathbb{1}\{f(X_i) \geq \tau\}.$$

This simple step is the key technical insight. Indeed, the indicators are bounded, regardless of the properties of the function f.

19.4.4 Step 2: Centering

Next, we invoke the standard method to control the sum of indicators by decomposing it into an expectation and a deviation term:

$$\begin{aligned}
&\mathbb{E}\inf_{f\in\mathcal{F}}\sum_{i=1}^{n}\mathbb{1}\{f(X_i)\geq\tau\}\\
&\quad\geq\inf_{f\in\mathcal{F}}\sum_{i=1}^{n}\mathbb{E}\,\mathbb{1}\{f(X_i)\geq\tau\}\\
&\qquad+\mathbb{E}\inf_{f\in\mathcal{F}}\sum_{i=1}^{n}\big(\mathbb{1}\{f(X_i)\geq\tau\}-\mathbb{E}\,\mathbb{1}\{f(X_i)\geq\tau\}\big)\\
&\quad\geq\inf_{f\in\mathcal{F}}n\cdot\mathbb{P}_{X\sim\mu}\{f(X)\geq\tau\}\\
&\qquad-\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\big(\mathbb{E}\,\mathbb{1}\{f(X_i)\geq\tau\}-\mathbb{1}\{f(X_i)\geq\tau\}\big)\right|.
\end{aligned}$$

The first inequality is the triangle inequality. In the last step, we have simply used the fact that $X_i\sim\mu$ iid.

Now, the first term

$$\inf_{f\in\mathcal{F}}\mathbb{P}_{X\sim\mu}\{f(X)\geq\tau\}$$

controls the probability that any function $f\in\mathcal{F}$ is likely to take a small value on a typical sample. This term tends to be large when $f(X)$ is not too "spiky." The probability of interest is the *complement* of the *small ball probability* $\mathbb{P}\{0\leq f(X)<\tau\}$, hence the nomenclature.

The second term equals

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\big(\mathbb{1}\{f(X_i)\geq\tau\}-\mathbb{E}\,\mathbb{1}\{f(X_i)\geq\tau\}\big)\right|$$

is the deviation of a sum of bounded functions of iid random variables, which we can handle using our toolkit for empirical processes.

19.4.5 Step 3: Deviation term

There are several methods for controlling the deviation term. For continuity with recent lectures, let us develop a general bound using VC theory. Alternative approaches include the direct application of the chaining inequality, perhaps combined with tools like the Rademacher comparison theorem.

Consider the class $\mathcal{C}_\tau$ that contains that super-level sets of the functions f at the level τ:

$$\mathcal{C}_\tau:=\big\{\{y\in\Omega:f(y)\geq\tau\}:f\in\mathcal{F}\big\}. \tag{19.2}$$

According to Theorem 17.3,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\big(\mathbb{1}\{f(X_i)\geq\tau\}-\mathbb{E}\,\mathbb{1}\{f(X_i)\geq\tau\}\big)\right|\leq\mathrm{Const}\sqrt{n\cdot\mathrm{vc}(\mathcal{C}_\tau)}.$$

The VC dimension of the class $\mathcal{C}_\tau$ can be computed for many types of functions.

19.4.6 The small ball bound

We may combine the results from the last three sections to reach the following theorem.

Theorem 19.1 (Small ball method: VC bound). Let (X_i) be an iid sample from a

probability measure μ, and consider a class $\mathscr{F}$ of positive functions. For each $\tau > 0$,

$$\mathbb{E} \inf_{f \in \mathscr{F}} \sum_{i=1}^{n} f(X_i) \geq \tau \left[n \cdot \inf_{f \in \mathscr{F}} \mathbb{P}\{f(X) \geq \tau\} - \sqrt{n} \cdot \mathrm{Const} \sqrt{\mathrm{vc}(\mathscr{C}_\tau)} \right],$$

where $\mathscr{C}_\tau$ is the class (19.2) of super-level sets.

You can see that there is a discrepancy in the scaling of the small ball term (which is linear in the sample size n) and the scaling of the deviation term (which is proportional to $\sqrt{n}$). Therefore, we can obtain a nontrivial bound provided that the small ball term is strictly positive for some τ and the VC dimension of the class $\mathscr{C}_\tau$ is finite.

19.5 Example: Minimum singular value of a heavy-tailed matrix

As a particular example, let us show how Theorem 19.1 leads to a lower bound on the minimum singular value of a random matrix whose rows are independent but may have heavy tails.

Let μ be a probability measure on $\mathbb{R}^d$, not necessarily centered at the origin. For simplicity, we assume that the measure is isotropic: $\mathbb{E}_\mu[\boldsymbol{x}\boldsymbol{x}^\mathsf{T}] = \mathbf{I}$. We also assume that the measure has uniformly bounded fourth moments:

$$\mathbb{E}_{\boldsymbol{x} \sim \mu} \langle \boldsymbol{x}, \boldsymbol{u} \rangle^4 \leq h \left(\mathbb{E}_{\boldsymbol{x} \sim \mu} \langle \boldsymbol{x}, \boldsymbol{u} \rangle^2 \right)^2 = h \|\boldsymbol{u}\|_2^4 \quad \text{for all } \boldsymbol{u} \in \mathbb{R}^d.$$

This hypothesis permits the case where the rows have tails that decay fairly slowly, but the decay must be uniform in every direction.

Draw independent samples $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ from the probability measure μ, where $n \geq d$. Form the (tall) random matrix

$$\boldsymbol{X} = \begin{bmatrix} \text{——} & \boldsymbol{x}_1^\mathsf{T} & \text{——} \\ & \vdots & \\ \text{——} & \boldsymbol{x}_n^\mathsf{T} & \text{——} \end{bmatrix} \in \mathbb{R}^{n \times d}.$$

For a vector $\boldsymbol{u} \in \mathbb{R}^d$, define the positive functions $f_{\boldsymbol{u}}(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{u} \rangle^2$ on $\mathbb{R}^d$. Then we can express

$$\sigma_{\min}^2(\boldsymbol{X}) = \inf_{\|\boldsymbol{u}\|_2 = 1} \sum_{i=1}^{n} f_{\boldsymbol{u}}(\boldsymbol{x}_i).$$

It is also helpful to define the class $\mathscr{F} = \{f_{\boldsymbol{u}} : \|\boldsymbol{u}\|_2 = 1\}$.

According to Theorem 19.1, for all $\tau > 0$,

$$\mathbb{E}\, \sigma_{\min}^2(\boldsymbol{X}) \geq \tau \left[n \cdot \inf_{f_{\boldsymbol{u}} \in \mathscr{F}} \mathbb{P}_{\boldsymbol{x} \sim \mu}\{f_{\boldsymbol{u}}(\boldsymbol{x}) \geq \tau\} - \sqrt{n} \cdot \mathrm{Const} \sqrt{\mathrm{vc}(\mathscr{C}_\tau)} \right]. \tag{19.3}$$

The class $\mathscr{C}_\tau$ is defined from $\mathscr{F}$ via (19.2). Let us develop bounds for the two terms in the bracket.

19.5.1 The small ball probability

The first term in (19.3) is a small ball probability. We can obtain a lower bound using the second-moment method. For a fixed vector with $\|\boldsymbol{u}\|_2 = 1$,

$$\begin{aligned} \mathbb{P}\{f_{\boldsymbol{u}}(\boldsymbol{x}) \geq \tau\} &= \mathbb{P}\left\{ \langle \boldsymbol{x}, \boldsymbol{u} \rangle^2 \geq \tau\, \mathbb{E} \langle \boldsymbol{x}, \boldsymbol{u} \rangle^2 \right\} \\ &\geq (1 - \tau)^2 \frac{\left(\mathbb{E} \langle \boldsymbol{x}, \boldsymbol{u} \rangle^2 \right)^2}{\mathbb{E} \langle \boldsymbol{x}, \boldsymbol{u} \rangle^4} \geq (1 - \tau)^2 \cdot h^{-1}. \end{aligned}$$

The first relation holds by the definition of $f_{\boldsymbol{u}}$ and the assumption that the measure μ is isotropic. The second inequality is Paley–Zygmund, and the last relation follows from the assumption that μ has uniform fourth moments. Taking the infimum over unit vectors,

$$\inf_{f_{\boldsymbol{u}} \in \mathcal{F}} \mathbb{P}\{f_{\boldsymbol{u}}(\boldsymbol{x}) \geq \tau\} \geq (1-\tau)^2 \cdot h^{-1}. \tag{19.4}$$

This lower bound is always constant under the assumptions we have posed.

19.5.2 The VC dimension

Recall that $\mathcal{C}_\tau$ is a class of super-level sets of the functions $f_{\boldsymbol{u}}$. The elements of the class take the form

$$\mathsf{C}_{\boldsymbol{u}} = \{\boldsymbol{y} \in \mathbb{R}^d : f_{\boldsymbol{u}}(y) \geq \tau\} = \{\langle \boldsymbol{y}, \boldsymbol{u}\rangle \geq \sqrt{\tau}\} \cup \{\langle \boldsymbol{y}, \boldsymbol{u}\rangle \leq -\sqrt{\tau}\}.$$

In other words, each set $\mathsf{C}_{\boldsymbol{u}}$ is a union of two half-spaces in $\mathbb{R}^d$. Recall that the class of all halfspaces in $\mathbb{R}^d$ has a VC dimension of $d+1$. The next problem implies that

$$\operatorname{vc}(\mathcal{C}_\tau) \leq \text{Const} \cdot (d+1). \tag{19.5}$$

Therefore, regardless of the level τ, the class $\mathcal{C}_\tau$ of super-level sets has controlled VC dimension.

Problem 19.2 (VC dimension of a union). For any class $\mathcal{H}$ of sets, prove that

$$\operatorname{vc}(\mathcal{H} \cup \mathcal{H}) \leq \text{Const} \cdot \operatorname{vc}(\mathcal{H}).$$

The class $\mathcal{H} \cup \mathcal{H}$ contains all unions of two sets, each drawn from $\mathcal{H}$. **Hint:** To obtain this result, it is easiest to employ the shattering function:

$$\pi_{\mathcal{H}}(m) := \sup_{|\mathsf{X}|=m} |\{\mathsf{H} \cap \mathsf{X} : \mathsf{H} \in \mathcal{H}\}|.$$

This function counts the number of distinct sets we can obtain by intersecting elements from the class $\mathcal{H}$ with a fixed set $\mathsf{X} \subseteq \Omega$ of cardinality m. Verify that $\pi_{\mathcal{H}\cup\mathcal{H}}(m) \leq \pi_{\mathcal{H}}(m)^2$, and use the definition of VC dimension.

19.5.3 Lower bound on the minimum singular value

Combine (19.3), (19.4), and (19.5) to obtain

$$\mathbb{E}\,\sigma_{\min}^2(\boldsymbol{X}) \geq \tau \left[n \cdot (1-\tau)^2 \cdot h^{-1} - \sqrt{n} \cdot \text{Const}\sqrt{d} \right].$$

Choose $\tau = 1/2$ and rearrange to arrive at the bound.

$$\mathbb{E}\,\sigma_{\min}^2(\boldsymbol{X}) \geq \text{const} \cdot \sqrt{n} \cdot \left(h^{-1}\sqrt{n} - \text{Const}\sqrt{d} \right). \tag{19.6}$$

Therefore, we obtain a nontrivial bound on the expectation $\mathbb{E}\,\sigma_{\min}^2(\boldsymbol{X})$ of the squared minimum singular value when the number n of rows satisfies $n \geq \text{Const}\sqrt{d}$.

For comparison, recall that Gordon's minimax theorem implies a precise bound for a standard normal matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{n\times d}$:

$$\mathbb{E}\,\sigma_{\min}(\boldsymbol{\Gamma}) \geq \sqrt{n} \cdot \left(\sqrt{n} - \sqrt{d} \right).$$

The new result (19.6) has some unspecified constants, but it gives a qualitatively similar bound under more general hypotheses. We only assume that the random matrix $\boldsymbol{X}$ has independent, isotropic rows that have uniform fourth moments. A key advantage of the small ball method is that it applies under rather weak moment assumptions.

19.6 Extensions and applications

The small ball method admits many extensions, and it has a wide range of applications. For examples, see the papers [Men14a; Men14b; KM15; Tro15b].

19.6.1 Restricted minimum singular values

For mathematical data science, one of the most valuable improvements is that the small ball method allows us to control *restricted* minimum singular values. That is, for a set $\mathsf{T} \subseteq \mathbb{S}^{d-1} \subset \mathbb{R}^d$ of unit vectors, we can obtain a lower bound for

$$\sigma_{\min}^2(\boldsymbol{X}; \mathsf{T}) := \inf_{\boldsymbol{u} \in \mathsf{T}} \|\boldsymbol{X}\boldsymbol{u}\|_2^2.$$

Roughly speaking, the lower bounds for the restricted singular value can be expressed in terms of the Gaussian width $w(\mathsf{T})$ of the index set:

$$w(\mathsf{T}) := \mathbb{E} \sup_{\boldsymbol{u} \in \mathsf{T}} \langle \boldsymbol{g}, \boldsymbol{u} \rangle \quad \text{where } \boldsymbol{g} \sim \textsc{normal}(\mathbf{0}, \mathbf{I}_d).$$

Under modest assumptions on $\boldsymbol{X}$ and T, we can obtain results like

$$n \geq \text{Const} \cdot w(\mathsf{T})^2 \quad \text{implies that} \quad \sigma_{\min}(\boldsymbol{X}; \mathsf{T}) \gg 0. \tag{19.7}$$

See [Tro15b, Prop. 5.1] for a detailed statement.

19.6.2 Dimension reduction

The small ball method can be applied to study randomized dimension reduction. For example, we can prove a partial version of the Johnson–Lindenstrauss result for a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ with independent, isotropic rows that satisfy some uniform moment bounds.

Consider a discrete set $\mathsf{A} = \{\boldsymbol{a}_i : i = 1, \dots, N\}$ of points in a Euclidean space $\mathbb{R}^d$. The random matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ can be used to embed the set A into $\mathbb{R}^n$:

$$\{\boldsymbol{X}\boldsymbol{a}_i : i = 1, \dots, N\} \subset \mathbb{R}^n.$$

To serve this function, we want to make sure that each well-separated pair of points in A remains separated after the embedding. We will argue that this outcome occurs when the embedding dimension $n \geq \text{Const} \cdot \log N$.

To analyze when this happens, construct the set of normalized secants:

$$\mathsf{T} = \left\{ \frac{\boldsymbol{a}_i - \boldsymbol{a}_j}{\|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2} : i \neq j \right\}.$$

Suppose that

$$\sigma_{\min}^2(\boldsymbol{X}; \mathsf{T}) = \inf_{\boldsymbol{u} \in \mathsf{T}} \|\boldsymbol{X}\boldsymbol{u}\|_2^2 = c \gg 0.$$

Then, the embedded point set satisfies the bounds

$$\|\boldsymbol{X}\boldsymbol{a}_i - \boldsymbol{X}\boldsymbol{a}_j\|_2^2 \geq c \|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2^2.$$

Using the heuristic result (19.7), we see that the number $c \gg 0$ when the embedding dimension n satisfies

$$n \geq \text{Const} \cdot w(\mathsf{T})^2 \geq \text{Const} \cdot \log N.$$

Indeed, the Gaussian width of a set of $\binom{N}{2}$ points in the unit sphere does not exceed $\text{Const} \cdot \sqrt{\log N}$. In fact, the width can be even smaller when the points are clustered in certain ways. Furthermore, the same argument yields embedding results for infinite sets, provided that the set of normalized secants has controlled Gaussian width.

19.6.3 Signal processing

Another application of the small ball method arises in signal processing [Tro15b]. Let $\boldsymbol{z}_\natural \in \mathbb{R}^d$ be an unknown signal. Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ be a measurement matrix that is known to us and does not depend on the signal. Suppose that we acquire a noisy linear observation $\boldsymbol{y}$ of the signal:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{z}_\natural + \boldsymbol{e} \in \mathbb{R}^n$$

The error $\boldsymbol{e} \in \mathbb{R}^n$ is unknown, but we suppose that its magnitude $\|\boldsymbol{e}\|_2$ is available to us. This formula models a measurement process or a communication channel.

In many settings, we have knowledge about the structure of the signal $\boldsymbol{z}_\natural$. For example, the signal may be a sparse vector or a low-rank matrix. We can exploit this prior information to design a (convex) optimization method for recovering the signal $\boldsymbol{z}_\natural$:

$$\text{minimize}_{\boldsymbol{z} \in \mathbb{R}^d} \quad \text{complexity}(\boldsymbol{z}) \quad \text{subject to} \quad \|\boldsymbol{X}\boldsymbol{z} - \boldsymbol{y}\|_2 \le \|\boldsymbol{e}\|_2.$$

The objective function depends on the type of signal; for instance, we could use the ℓ_1 norm to search for a sparse signal $\boldsymbol{z}_\natural$. The signal estimate $\widehat{\boldsymbol{z}} \in \mathbb{R}^d$ that results from the optimization problem satisfies a (deterministic) error bound of the form

$$\|\widehat{\boldsymbol{z}} - \boldsymbol{z}_\natural\| \le \frac{2\|\boldsymbol{e}\|_2}{\sigma_{\min}(\boldsymbol{X}; \mathsf{T})}.$$

In this expression, the set $\mathsf{T} \subseteq \mathbb{S}^{d-1} \subset \mathbb{R}^d$ depends on the descent cone of the objective function at the ground truth $\boldsymbol{z}_\natural$. See [Tro15b, Prop. 2.6] for a complete statement.

As a consequence of this machinery, if we can show that the restricted minimum singular value $\sigma_{\min}(\boldsymbol{X}; \mathsf{T}) \gg 0$, then we can guarantee the reconstructed vector $\widehat{\boldsymbol{z}}$ is a good approximation to the true signal $\boldsymbol{z}_\natural$. When the measurement matrix $\boldsymbol{X}$ is random, then we may be able to use the small ball method to obtain bounds for the restricted singular value. This methodology allows us to treat some *idealized* signal processing problems, such as compressed sensing problems or phase retrieval from random measurements.

Lecture bibliography

[KM15] V. Koltchinskii and S. Mendelson. "Bounding the smallest singular value of a random matrix without concentration". In: *International Mathematics Research Notices* 2015.23 (2015), pages 12991–13008.

[Men14a] S. Mendelson. "A remark on the diameter of random sections of convex bodies". In: *Geometric aspects of functional analysis*. Springer, 2014, pages 395–404.

[Men14b] S. Mendelson. "Learning without concentration". In: *Conference on Learning Theory*. PMLR. 2014, pages 25–39.

[Tro15b] J. A. Tropp. "Convex recovery of a structured signal from independent random linear measurements". In: *Sampling Theory, a Renaissance*. Springer, 2015, pages 67–101.

IV. problem sets

Problem Set 1

This assignment covers Chebyshev's inequality, linear and nonlinear variance bounds, Poincaré inequalities, the Laplace transform method, linear and nonlinear cgf bounds, and (modified) log-Sobolev inequalities.

Exercises

1 **(Variance mix).** Let Z be a square-integrable, real random variable: $\mathbb{E}\, Z^2 < +\infty$.

(a) Show that $\operatorname{Var}[Z] = \inf_{a \in \mathbb{R}} \mathbb{E}(Z - a)^2$.

(b) Let Z' be an independent copy of Z. Establish the identities

$$\operatorname{Var}[Z] = \tfrac{1}{2}\, \mathbb{E}(Z - Z')^2 = \mathbb{E}(Z - Z')_+^2 = \mathbb{E}(Z - Z')_-^2.$$

The functions $(a)_+ := \max\{a, 0\}$ and $(a)_- := \max\{-a, 0\}$ bind before the square and expectation.

(c) Assume that Z takes values in $[a, b]$. Show that $\operatorname{Var}[Z] \le \frac{1}{4}|b - a|^2$. **Hint:** Use (a).

2 **(Martinet).** Let $(X_1, \dots, X_n)$ be a family of random variables, not necessarily independent. Consider a square-integrable random variable of the form $Z := f(X_1, \dots, X_n)$. Define $Y_0 = \mathbb{E}\, Z$ and

$$Y_i := \mathbb{E}[Z \mid X_1, \dots, X_i] \quad \text{for } i = 1, \dots, n.$$

Construct the *difference sequence* $\Delta_i := Y_i - Y_{i-1}$ for $i = 1, \dots, n$.

(a) Show that $(Y_i : i = 0, \dots, n)$ is a martingale sequence (called a *Doob martingale*). That is, $\mathbb{E}[Y_{i+1} \mid X_1, \dots, X_i] = Y_i$ almost surely.

(b) Confirm that the martingale differences are orthogonal: $\mathbb{E}[\Delta_i \Delta_j] = 0$ for $i \neq j$.

3 **(Subgaussian omnibus).** Consider a real random variable Z that is centered: $\mathbb{E}\, Z = 0$. Define the *cumulant generating function* $\xi_Z(\theta) := \log \mathbb{E}\, \mathrm{e}^{\theta Z}$ for $\theta \in \mathbb{R}$. The cgf may take the value $+\infty$.

(a) Compute (or look up) the cgf of a real, centered normal random variable.

(b) A centered, real random variable Z is called *subgaussian* if it satisfies (i) below. Show that the other two statements are equivalent to (i) up to scaling v by a constant factor.

(i) $\xi_Z(\theta) \le v\theta^2/2$ for all $\theta \in \mathbb{R}$, where $v \ge 0$ is called the *variance proxy*.

(ii) $\mathbb{P}\{|Z| \ge t\} \le C\, \mathrm{e}^{-ct^2/v}$ for all $t \ge 0$ and constants $c, C > 0$.

(iii) $(\mathbb{E}\,|Z|^p)^{1/p} \le C' \sqrt{pv}$ for all $p \ge 1$ and a constant $C' > 0$.

Hint: Prove that (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) $\Rightarrow$ (i). In sequence, you'll use the Laplace transform method, integration by parts, and a Taylor series expansion.

(c) (*) We say that a centered, real random variable Z is *subexponential* if $\xi_{|Z|}(\theta) \le R\theta^2$ for all $|\theta| < \theta_0$, where $R \ge 0$ and $\theta_0 > 0$. Formulate and prove an equivalence result analogous to (a).

4 **(Mad maximal).** This exercise develops some important bounds on maxima.

(a) Consider centered, subgaussian random variables $(X_1, \dots, X_n)$, with variance proxies bounded above by v, not necessarily independent. Prove that

$$\mathbb{E} \max_i X_i \le \sqrt{2v \log n}.$$

Hint: Use Jensen to check that $\mathbb{E} Z \le \theta^{-1} \xi_Z(\theta)$ for $\theta > 0$. Bound the maximum by a sum.

(b) (*) Considered centered, subexponential random variables $(X_1, \dots, X_n)$, not necessarily independent. What is the analog of the result in (a)?

Problems

1 **(Rad).** A *Rademacher* random variable ε takes values ± 1 with equal probability: $\varepsilon \sim \text{UNIFORM}\{\pm 1\}$. A (real-valued) *Rademacher series* is a random variable of the form

$$\sum_{i=1}^n \varepsilon_i a_i \quad \text{where } \varepsilon_i \text{ are iid Rademacher and } \boldsymbol{a} = (a_1, \dots, a_n) \in \mathbb{R}^n.$$

A *Rademacher process* is a family of Rademacher series, all involving the same Rademacher variables:

$$\left(\sum_{i=1}^n \varepsilon_i a_i : \boldsymbol{a} \in \mathsf{T} \right), \quad \text{where } \varepsilon_i \text{ are iid Rademacher and } \mathsf{T} \subset \mathbb{R}^n.$$

(a) Assume $\|\boldsymbol{a}\|_{\ell_1} = 1$. When are the minimum and maximum variance of $\sum_{i=1}^n \varepsilon_i a_i$ attained?

(b) Apply Chebyshev's inequality and Hoeffding's inequality to the Rademacher series $\sum_{i=1}^n \varepsilon_i a_i$.

(c) For $p \ge 1$, Khintchine's inequalities state that

$$c\mu_2 \le \mu_p \le C\sqrt{p}\mu_2 \quad \text{where} \quad \mu_p := \left(\mathbb{E} \left| \sum_{i=1}^n \varepsilon_i a_i \right|^p \right)^{1/p}.$$

Derive the upper bound in Khintchine's inequality from Hoeffding's inequality and Exercise (3)(a). (*) Prove the lower inequality. **Hint:** Use Hölder to bound μ_2 in terms of μ_1.

(d) Use Efron–Stein–Steele (ESS) to bound the variance of the supremum of a Rademacher process:

$$\text{Var} \left[\sup_{\boldsymbol{a} \in \mathsf{T}} \sum_{i=1}^n \varepsilon_i a_i \right] \le 2 \sup_{\boldsymbol{a} \in \mathsf{T}} \sum_{i=1}^n a_i^2 =: 2 \,\text{radius}^2_{\ell_2}(\mathsf{T}).$$

Hint: Use the formulation of ESS with the plus function.

2 **(Positive concentration).** Positive random variables enjoy special concentration properties which ensure that they are bounded away from zero. Assume that $Z \ge 0$ and that Z is square integrable.

(a) (*) Prove the Paley–Zygmund inequality (aka *second moment method* bound):

$$\mathbb{P}\{Z \ge (1-t)(\mathbb{E} Z)\} \ge \frac{t^2 (\mathbb{E} Z)^2}{\mathbb{E} Z^2} \quad \text{for } t \in (0, 1].$$

Hint: Write

$$\mathbb{E}[Z] = \mathbb{E}[Z\mathbb{1}\{Z < (1-t)(\mathbb{E}\,Z)\}] + \mathbb{E}[Z\mathbb{1}\{Z \geq (1-t)(\mathbb{E}\,Z)\}].$$

Use Cauchy–Schwarz.

(b) Show that the cgf $\xi_{Z-\mathbb{E}\,Z}(-\theta) \leq \theta^2(\mathbb{E}\,Z^2)/2$ for $\theta \geq 0$. Establish the subgaussian lower tail bound

$$\mathbb{P}\{Z \leq (1-t)(\mathbb{E}\,Z)\} \leq \exp\left(\frac{-t^2(\mathbb{E}\,Z)^2}{2\,\mathbb{E}\,Z^2}\right) \quad \text{for } t \in (0,1].$$

Hint: Develop a bound for e^{-a} when $a \geq 0$.

(c) Consider an independent family $(X_i : i = 1, \dots, n)$ of positive random variables, and define $Z = \sum_{i=1}^n X_i$. Derive an upper bound for $\mathbb{P}\{Z \leq (1-t)(\mathbb{E}\,Z)\}$.

(d) (*) Assume that $Z \geq 0$ and Z is absolutely continuous, with density that is bounded above by one. Give two examples of random variables that have this property. Prove that $\xi_Z(-\theta) \leq -\log\theta$ for $\theta > 0$, and use this inequality to derive an upper bound on $\mathbb{P}\{Z \leq t\}$. What is the analogue of (c)?

3 **(Convex Poincaré).** This problem develops a somewhat general Poincaré inequality.

(a) Consider a bounded, real random variable: $a \leq X \leq b$. For all convex $f : [a,b] \to \mathbb{R}$, prove that

$$\mathrm{Var}[f(X)] \leq (b-a)^2\,\mathbb{E}[f'(X)^2].$$

Hint: Use the exchangeable pairs representation for the variance and the identity

$$f(x) - f(y) = (x-y)\int_0^1 f'((1-s)x + sy)\,\mathrm{d}s,$$

valid for all $x, y \in [a,b]$.

(b) A *separately convex* function $f : \mathbb{R}^n \to \mathbb{R}$ is convex when restricted to each coordinate. Confirm that a convex function is separately convex. Tensorize the convex Poincaré inequality to obtain a variance bound for a separately convex function of independent, bounded random variables.

(c) (*) Let $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ be a random matrix with independent entries that take values in $[-1,1]$. Give a dimension-free bound for $\mathrm{Var}[\|\boldsymbol{X}\|]$. **Hint:** The spectral norm $\|\cdot\|$ is convex and 1-Lipschitz. (**) What is the scale of $\mathbb{E}\,\|\boldsymbol{X}\|$?

Applications

1 **(Robust mean estimation).** By the LLN, sample averages converge to the population mean. Nevertheless, for a small sample, the sample average is unreliable. One remedy is to use a more robust method, such as the median-of-means estimator. Consider a real random variable X that is square-integrable.

(a) Let $\bar{X}_k$ be the average of k independent copies of X. Use Chebyshev's inequality to obtain a concentration inequality for $\bar{X}_k$.

(b) Let Y_n be the median of n independent copies of $\bar{X}_k$. For a level $t > 0$, use Chernoff's inequality to obtain a subexponential bound for $\mathbb{P}\{|Y_n - \mathbb{E}\,X| \geq t\}$. **Hint:** Consider the probability that more than $n/2$ independent realizations of $\bar{X}_k$ lie outside the interval $\mathbb{E}\,X \pm t$.

(c) (*) Suppose that we sample m independent copies of X. For a given level $t > 0$, what is the best setting of the parameters k and n to minimize the bound on the failure probability?

2 **(Jack the Knife).** The ESS inequality was developed to study the *jackknife*, a statistical methodology based on subsampling. For $k \in \mathbb{N}$, let $f_k : \mathbb{R}^k \to \mathbb{R}$ be a family of estimators for a parameter θ of a distribution. Assume each f_k is symmetric (invariant under permutation of arguments). Consider iid real random variables $(X_1, \dots, X_n)$. The full-data estimate $Z := f_n(X_1, \dots, X_n)$. Define jackknife replicates

$$Z^{(i)} := f_{n-1}(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) \quad \text{for } i = 1, \dots, n.$$

Construct the average $\bar{Z}_n := n^{-1} \sum_{i=1}^n Z^{(i)}$ of the jackknife replicates.

(a) A (simplified) jackknife estimate for the variance of the estimator f_{n-1} is

$$\widehat{\mathrm{Var}} := \sum\nolimits_{i=1}^n \big(Z^{(i)} - \bar{Z}_n\big)^2 = \frac{1}{2n} \sum\nolimits_{i,j=1}^n \big(Z^{(i)} - Z^{(j)}\big)^2.$$

Show that $\mathrm{Var}[f_{n-1}] \leq \mathbb{E}[\widehat{\mathrm{Var}}]$. That is, the jackknife overestimates variance on average.

(b) (*) Shockingly, the jackknife can (sometimes) evaluate the *bias* of a parameter estimate. Consider

$$\widehat{\mathrm{Bias}} := (n-1)\big(\bar{Z}_n - Z\big).$$

Suppose that the estimators f_k are also quadratic (can be written as a degree-two polynomial). The sample variance is one such functional. Show that $\mathbb{E}[\widehat{\mathrm{Bias}}] = \mathbb{E}\,Z - \theta$. The jackknife is an unbiased estimator of the bias! (*) Can you give more examples of exchangeable, quadratic statistics?

3 **(Bin packing).** Bounded differences can be used to study a classic combinatorial optimization problem called stochastic bin packing. Consider a real random variable $0 \leq W \leq 1$ that describes the length of a randomly chosen suitcase. Each overhead bin on an aircraft can hold suitcases with total length one. Let Z_n be the minimum number of bins sufficient to hold n suitcases whose lengths are iid copies of W.

(a) Show that $\mathbb{E}[Z_n] \geq n\,\mathbb{E}\,W$.

(b) Show the $\mathrm{Var}[Z_n] \leq n/4$. Explain the significance.

4 **(Second-order Gaussian chaos).** Fix a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ with zero diagonal. Let $\boldsymbol{z} = (z_1, \dots, z_n) \sim \textsc{normal}(\mathbf{0}, \mathbf{I}_n)$ be a standard normal random vector. The quadratic form $X = \boldsymbol{z}^* \boldsymbol{A} \boldsymbol{z}$ is called a *second-order Gaussian chaos*. It is a simple model for interactions among entities. (Why?)

(a) Compute $\mathbb{E}\,X$ and $\mathrm{Var}[X]$.

(b) Bound $\mathrm{Var}[X]$ using the Gaussian Poincaré inequality. Compare with the exact variance.

(c) Explain why $X \sim \sum_{i=1}^n \lambda_i (z_i^2 - 1)$, where $(\lambda_1, \dots, \lambda_n)$ are the eigenvalues of $\boldsymbol{A}$.

(d) Confirm that $\xi_{z_i^2-1}(\theta) \leq \theta^2/(1-(2\theta)_+)$ for all $\theta < 1/2$. Compare with the Bernstein cgf.

(e) Develop the upper tail bound

$$\mathbb{P}\{X > t\} \leq \exp\left(\frac{-t^2/4}{\|\boldsymbol{A}\|_{\mathrm{F}}^2 + t\|\boldsymbol{A}\|}\right),$$

where $\|\boldsymbol{A}\|_{\mathrm{F}}$ is the Frobenius norm and $\|\boldsymbol{A}\|$ is the spectral norm.

Problem Set 2

This assignment covers linear and nonlinear cgf bounds, entropy, (modified) log-Sobolev inequalities, Herbst's argument, symmetrization, moment inequalities, matrix concentration, and applications.

Problems

1 **(Nonlinear Bernstein inequality).** Suppose that $\boldsymbol{x} = (X_1, \dots, X_n)$ is a random vector that satisfies a modified log-Sobolev inequality (MLSI):

$$\operatorname{ent}(\mathrm{e}^{f(\boldsymbol{x})}) \le C\ \mathbb{E}[\|\nabla f(\boldsymbol{x})\|_2^2\, \mathrm{e}^{f(\boldsymbol{x})}] \quad \text{for nice } f : \mathbb{R}^n \to \mathbb{R}.$$

We have used a uniform bound on the norm of the gradient to derive normal concentration. In this problem, we will see how to obtain bounds that reflect the typical size of the gradient.

(a) **(*MLSI implies Poincaré).** Show that the MLSI implies the Poincaré inequality with constant C:

$$\operatorname{Var}[f(\boldsymbol{x})] \le C\ \mathbb{E}[\|\nabla f(\boldsymbol{x})\|_2^2].$$

Hint: Apply the MLSI to the function $\log(1 + \eta f)$, and take the limit as $\eta \to 0$.

(b) **(*Young's inequality).** Suppose that Y is a positive random variable with $\mathbb{E}\, Y = 1$, and let W be another random variable. Prove Young's inequality for entropy:

$$\mathbb{E}[WY] \le \log \mathbb{E}\, \mathrm{e}^W + \mathbb{E}[Y \log Y].$$

In other words, entropy is the Fenchel dual of the exponential mean (or cgf). (*) What is the equality condition? **Hint:** Consider random variables Y and $Z = \mathrm{e}^W / \mathbb{E}\, \mathrm{e}^W$, and use the fact that relative entropy is positive.

(c) **(Nonlinear Bernstein).** In this part, we will develop a nonlinear analog of the Bernstein inequality.

(i) Using (c), deduce that $\operatorname{Var}[f] \le C\psi^{-1}\xi_{\|\nabla f\|_2^2}(\psi)$ for all $\psi > 0$. In other words, the exponential mean of the energy is a plausible variance proxy.

(ii) Apply Young's inequality to decouple the expression $\mathbb{E}[\|\nabla f\|_2^2\, \mathrm{e}^f]$. **Hint:** Normalize by $\psi\, \mathbb{E}\, \mathrm{e}^f$.

(iii) Use the MLSI and the last result to obtain a bound for the entropy:

$$\frac{\operatorname{ent}(\mathrm{e}^{\theta f})}{\theta^2\, \mathbb{E}\, \mathrm{e}^{\theta f}} \le \frac{C\psi^{-1}\xi_{\|\nabla f\|_2^2}(\psi)}{1 - C\theta^2/\psi} \quad \text{when } \psi > C\theta^2.$$

(iv) Apply the Herbst argument to obtain a Bernstein-type bound for the cgf of $f - \mathbb{E} f$:

$$\xi_{f-\mathbb{E} f}(\theta) \le \frac{C(\theta^2/\psi)\xi_{\|\nabla f\|_2^2}(\psi)}{1 - C\theta^2/\psi} \le \frac{C(\theta^2/\psi)\xi_{\|\nabla f\|_2^2}(\psi)}{1 - \sqrt{C\theta^2/\psi}}.$$

(v) Deduce a Bernstein-type concentration inequality for f. You may leave ψ as a free parameter.

(d) **(Self-bounded functions).** We say that a centered random variable f is *self-bounded* when $\|\nabla f\|_2^2 \le af + b$ for $a \ge 0$ and $b \in \mathbb{R}$.

1. Using (c)(iv), produce a bound for $\xi_f(\theta)$. **Hint:** Choose $\psi = \theta/a$.
2. Deduce an upper tail bound for f.

(e) **(Psd quadratic forms).** For simplicity, assume that $\boldsymbol{x}$ is an isotropic column vector: $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^*] = \mathbf{I}$. Let $\boldsymbol{A} \in \mathbb{H}_n$ be a psd matrix, and consider the centered random variable $f(\boldsymbol{x}) = \boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} - \operatorname{tr}(\boldsymbol{A})$.

(i) Give examples of isotropic random vectors that satisfy a (convex) MLSI.

(ii) Show that f is self-bounded, and use (d) to obtain an upper tail bound. Compare with the Gaussian chaos bound on PS1.

(iii) Observe that the result from (iii) holds under a convex MLSI.

(f) **(*Indefinite quadratic forms).** As an example, consider the random variable $q(\boldsymbol{x}) = \boldsymbol{x}^*\boldsymbol{B}\boldsymbol{x}$ where $\boldsymbol{B} \in \mathbb{H}_n$ is symmetric but not necessarily psd. Reduce concentration for the indefinite quadratic form q to concentration for the psd quadratic form q. Deduce a complete concentration inequality for q.

2 **(Matrix Moment Inequalities).** In this problem, you will develop a complete proof of the matrix moment inequalities. These are polynomial moment analogs of the matrix Bernstein inequality.

(a) **(Gaussian symmetrization).** First, we show that we can symmetrize an independent sum using Gaussians instead of Rademachers. Consider an independent family $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ of integrable random variables taking values in a (finite-dimensional) normed linear space X. Consider an independent family $(\varepsilon_1, \dots, \varepsilon_n)$ of iid Rademachers and an independent family $(g_1, \dots, g_n)$ of iid standard normal variables.

(i) Establish the Gaussian symmetrization principle:

$$\mathbb{E}\left\|\sum\nolimits_{i=1}^n (\boldsymbol{x}_i - \mathbb{E}\,\boldsymbol{x}_i)\right\|_{\mathsf{X}} \le \sqrt{2\pi}\ \mathbb{E}\left\|\sum\nolimits_{i=1}^n g_i(\boldsymbol{x}_i - \mathbb{E}\,\boldsymbol{x}_i)\right\|_{\mathsf{X}}.$$

Hint: We have $g_i \sim \varepsilon_i |g_i|$. Extend to a positive convex function of the norm.

(ii) (*) By imitating the proof of the symmetrization principle, show that

$$\frac{1}{2}\,\mathbb{E}\left\|\sum\nolimits_{i=1}^n \varepsilon_i(\boldsymbol{x}_i - \mathbb{E}\,\boldsymbol{x}_i)\right\|_{\mathsf{X}} \le \mathbb{E}\left\|\sum\nolimits_{i=1}^n (\boldsymbol{x}_i - \mathbb{E}\,\boldsymbol{x}_i)\right\|_{\mathsf{X}}$$

We have already obtained the matching upper bound. (*) Extend to the case of a positive convex function of the norm.

(iii) **(*Contraction principle).** Consider $(a_1, \dots, a_n) \in \mathbb{R}^n$ with $|a_i| \le 1$ for each i. Show that

$$\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\sum\nolimits_{i=1}^n a_i\varepsilon_i\boldsymbol{x}_i\right\|_{\mathsf{X}} \le \mathbb{E}_{\boldsymbol{\epsilon}}\left\|\sum\nolimits_{i=1}^n \varepsilon_i\boldsymbol{x}_i\right\|_{\mathsf{X}}.$$

Hint: A convex function on a compact set achieves its maximum at an extreme point.

(iv) (*) Use the contraction principle to derive a lower bound in the Gaussian symmetrization principle. Is it dimension-free?

(b) **(Standard Normal Moments).** Next, we perform an important calculation that forms the pattern for the next argument. Let $g \sim \text{NORMAL}(0, 1)$.

(i) Prove the Gaussian integration by parts identity. For every differentiable function $f : \mathbb{R} \to \mathbb{R}$ where the expectations are defined,

$$\mathbb{E}[gf(g)] = \mathbb{E}[f'(g)].$$

(ii) For $p \in \mathbb{N}$, use Gaussian integration by parts to verify that $\mathbb{E}\, g^{2p} = (2p-1)!!$.

(iii) (*) Establish the inequality $[(2p-1)!!]^{1/(2p)} \leq \sqrt{(2p+1)/\mathrm{e}}$.

(c) **(Matrix Khintchine).** Let $\boldsymbol{A}_1, \dots, \boldsymbol{A}_n \in \mathbb{H}_d$ be fixed matrices, and let $(g_1, \dots, g_n)$ be iid standard normal variables. Define the random matrix $\boldsymbol{X} := \sum_{i=1}^n g_i \boldsymbol{A}_i$. We will establish the matrix Khintchine inequality. For each integer $p \geq \lceil \log d \rceil$,

$$\left(\mathbb{E}\, \|\boldsymbol{X}\|^{2p}\right)^{1/(2p)} \leq \sqrt{2p+1}\, \|\mathbb{E}\, \boldsymbol{X}^2\|^{1/2} = \sqrt{2p+1}\, \left\|\sum\nolimits_{i=1}^n \boldsymbol{A}_i^2\right\|^{1/2}.$$

What is important is that the leading constant has order $\sqrt{p}$, so you do not need to get hung up on obtaining the precise value $\sqrt{2p+1}$.

(i) Verify that $\|\boldsymbol{A}\| \leq (\operatorname{tr} \boldsymbol{A}^p)^{1/p} \leq d^{1/p} \|\boldsymbol{A}\|$ for all $\boldsymbol{A} \in \mathbb{H}_d$ and $p \in \mathbb{N}$.

(ii) Apply Gaussian integration by parts to obtain the identity

$$\begin{aligned} \mathbb{E} \operatorname{tr} \boldsymbol{X}^{2p} &= \sum\nolimits_{i=1}^n \mathbb{E} \operatorname{tr}[g_i \boldsymbol{A}_i \boldsymbol{X}^{2p-1}] \\ &= \sum\nolimits_{i=1}^n \sum\nolimits_{q=0}^{2p-2} \mathbb{E} \operatorname{tr}[\boldsymbol{A}_i \boldsymbol{X}^q \boldsymbol{A}_i \boldsymbol{X}^{2p-2-q}]. \end{aligned}$$

Hint: For a matrix-valued function $f : t \mapsto \boldsymbol{A}(t)^{q+1}$ and positive integer q, the derivative is $f'(t) = \sum_{r=0}^q \boldsymbol{A}(t)^r \boldsymbol{A}'(t) \boldsymbol{A}(t)^{q-r}$.

(iii) (*) For $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{H}_d$ and integers $0 \leq q \leq 2r$, establish the inequality

$$\operatorname{tr}[\boldsymbol{A}\boldsymbol{B}^q \boldsymbol{A}\boldsymbol{B}^{2r-q}] \leq \operatorname{tr}[\boldsymbol{A}^2 \boldsymbol{B}^{2r}].$$

Hint: Use eigenvalue decompositions and the generalized AM–GM inequality.

(iv) Complete the proof of the matrix Khintchine inequality. In spirit, the argument that we gave in class to obtain the scalar Khintchine inequality is modeled on the same ideas.

(v) (*) For psd $\boldsymbol{A}$, show that $\operatorname{intdim}(\boldsymbol{A}) := \operatorname{tr} \boldsymbol{A} / \|\boldsymbol{A}\| \leq \operatorname{rank}(\boldsymbol{A})$. Explain the term "intrinsic dimension."

(vi) (*) Prove matrix Khintchine holds with $\operatorname{intdim}(\sum_{i=1}^n \boldsymbol{A}_i^2)$ in place of d, where $p = \lceil \log d \rceil$.

(d) **(*Matrix Rosenthal I).** Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \in \mathbb{H}_d$ be statistically independent, random psd matrices. For each integer p with $p \geq \lceil \log d \rceil \geq 1$, show that

$$\begin{aligned} &\left(\mathbb{E} \left\|\sum\nolimits_{i=1}^n \boldsymbol{X}_i\right\|^{2p}\right)^{1/2p} \\ &\qquad \leq \left[\left\|\sum\nolimits_{i=1}^n \mathbb{E}\, \boldsymbol{X}_i\right\|^{1/2} + \text{Const} \cdot \sqrt{p} \cdot \left(\mathbb{E} \max\nolimits_i \|\boldsymbol{X}_i\|^{2p}\right)^{1/4p}\right]^2. \end{aligned}$$

Hint: The argument is the same as in class, but it requires a modest amount of matrix analysis.

(e) **(Matrix Rosenthal II).** Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \in \mathbb{H}_d$ be statistically independent, zero-mean random matrices. For each integer p with $p \geq \lceil \log d \rceil \geq 1$, show that

$$\left(\mathbb{E} \left\| \sum\nolimits_{i=1}^{n} \boldsymbol{X}_i \right\|^{4p} \right)^{1/4p} \leq \text{Const} \cdot \sqrt{p} \cdot \left\| \sum\nolimits_{i=1}^{n} \mathbb{E}\, \boldsymbol{X}_i^2 \right\|^{1/2} + \text{Const} \cdot p \cdot \left(\mathbb{E} \max\nolimits_i \|\boldsymbol{X}_i\|^{4p} \right)^{1/4p}.$$

Applications

1 **(Johnson–Lindenstrauss).** In this application, we will look at a famous theorem of Bill Johnson & Yoram Lindenstrauss on dimension reduction.

> **Theorem 19.3 (Johnson–Lindenstrauss).** Let $\boldsymbol{a}_1, \dots, \boldsymbol{a}_N \in \mathbb{R}^d$ be an arbitrary set of fixed points, and choose a parameter $\varepsilon > 0$. For each $m \geq \text{const} \cdot \varepsilon^{-2} \log N$, there exists a (linear) dimension reduction map $\boldsymbol{\Phi} : \mathbb{R}^d \to \mathbb{R}^m$ with the property that
>
> $$(1 - \varepsilon)\|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2^2 \leq \|\boldsymbol{\Phi}\boldsymbol{a}_i - \boldsymbol{\Phi}\boldsymbol{a}_j\|_2^2 \leq (1 + \varepsilon)\|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2^2 \qquad (19.8)$$
>
> for all $i, j = 1, \dots, N$.

This result states that all pairwise distances are preserved even though the embedding dimension m is *logarithmic* in the size N of the point set! The JL Theorem has many (theoretical) applications in modern computer science, because it (theoretically) allows us to replace a high-dimensional problem by a lower-dimensional problem that may be easier to solve.

(a) Consider a matrix $\boldsymbol{\Phi} \in \mathbb{R}^{m \times d}$ with iid $\textsc{normal}(0, m^{-1})$ entries. Prove the Johnson–Lindenstrauss theorem for this choice of the dimension reduction map.

(b) Explain why a similar result is valid for any random matrix with iid centered, bounded entries.

(c) (*) Let's see how well (that is, badly) this result works in practice. Download the MATLAB file `myisolet.mat` from the course website. It contains a 617×1559 matrix whose *columns* are data points. Perform the following experiment 100 times. For embedding dimensions $m \in \{2^i : 0 \leq i \leq 8\}$, construct a random embedding matrix $\boldsymbol{\Phi} \in \mathbb{R}^{m \times 617}$. Apply the random embedding $\boldsymbol{\Phi}$ to the data, and compute the empirical distortion $\varepsilon(m)$. That is, calculate the smallest number ε where the $2\binom{1559}{2}$ distinct inequalities in (19.8) hold simultaneously. For two choices of m, plot a histogram of the empirical distortion $\varepsilon(m)$ over the 100 trials. Plot the empirical average distortion $\varepsilon(m)$ as a function of m. What do you conclude from these experiments?

2 **(Spectral Clustering).** The *stochastic block model* (SBM) is a simple (i.e., ridiculous) random model for community structure. Consider $2n$ individuals, partitioned into two communities I and I^c of equal cardinality. Fix probabilities $p, q \in [0, 1]$

with $p > q$. We construct a random graph on these $2n$ vertices, where the presence of an edge means that two individuals are acquainted. For each set $\{u, v\}$ of distinct vertices, we introduce an edge independently at random (a) with probability p when u, v belong to the same community or (b) with probability q when u, v belong to opposite communities. Let $\boldsymbol{A}$ be the random adjacency matrix.

One basic question is whether we can identify the communities from a single observation of the graph. A simple but useful approach is spectral clustering. That is, we compute the unit-ℓ_2-norm eigenvector $\boldsymbol{u}_2(\boldsymbol{A})$ associated with the *second* largest eigenvalue of $\boldsymbol{A}$. Define the random set $\hat{I} := \{i : (\boldsymbol{u}_2)_i > 0\}$. We will develop conditions under which $\hat{I}$ aligns with one of the two communities I or I^c.

(a) Write the random adjacency matrix $\boldsymbol{A}$ as a sum of independent random matrices. Apply the matrix Bernstein inequality or (easier) the matrix Rosenthal inequality to see that

$$\left(\mathbb{E}\,\|\boldsymbol{A} - \mathbb{E}\,\boldsymbol{A}\|^2\right)^{1/2} \leq \text{const} \cdot \left[\sqrt{\varrho \log n} + \log n\right],$$

where $\varrho := 0.5(p+q)n$ is the expectation of the average degree of a vertex in the graph.

(b) Compute $\mathbb{E}\,\boldsymbol{A}$ and its eigenvalues. What is the unit-norm eigenvector associated with the largest eigenvalue? What is the unit-norm eigenvector associated with the second largest eigenvalue?

(c) Write a paragraph to explain why the spectral clustering procedure is a natural mechanism for community detection.

(d) **(**Davis–Kahan).** Let $\boldsymbol{S}, \boldsymbol{T}$ be symmetric matrices of the same size. Suppose that the ith eigenvalue of $\boldsymbol{S}$ is separated from the rest of spectrum:

$$\text{gap}_i(\boldsymbol{S}) := \min_{j \neq i} |\lambda_i(\boldsymbol{S}) - \lambda_j(\boldsymbol{S})| =: \varepsilon.$$

Then the acute angle θ_i between the unit-norm eigenvectors $\boldsymbol{u}_i(\boldsymbol{S})$ and $\boldsymbol{u}_i(\boldsymbol{T})$ associated with the ith eigenvalues satisfies the inequality

$$\sin\theta_i \leq 2\varepsilon^{-1}\|\boldsymbol{S} - \boldsymbol{T}\|.$$

In particular, $\min_{\pm} \|\boldsymbol{u}_i(\boldsymbol{S}) \pm \boldsymbol{u}_i(\boldsymbol{T})\|_2 \leq 2^{3/2}\varepsilon^{-1}\|\boldsymbol{S} - \boldsymbol{T}\|$.

(e) Compute $\text{gap}_2(\mathbb{E}\,\boldsymbol{A})$, and define $\mu n := \text{gap}_2$.

(f) Suppose that the average expected degree $\varrho \gg \log n$. Show that

$$\mathbb{E} \min_{\pm} \|\boldsymbol{u}_2(\mathbb{E}\,\boldsymbol{A}) \pm \boldsymbol{u}_2(\boldsymbol{A})\|_2^2 \leq \text{const} \cdot \frac{\log n}{\mu^2 n}.$$

(g) Conclude that the estimated community $\hat{I}$ almost coincides with a true community I or I^c:

$$\mathbb{E} \max\left\{\#(\hat{I} \cap I), \#(\hat{I} \cap I^c)\right\} \geq n - \text{const} \cdot \frac{\log n}{\mu^2}$$

Why can't we tell which community $\hat{I}$ will line up with?

(h) Write a paragraph to explain what assumptions on p, q, n are needed to identify communities in the SBM via spectral clustering. Explain what your conclusions mean intuitively.

Problem Set 3

This assignment covers Gaussian processes, random processes, comparison theorems, and phase transition phenomena.

Exercises

1 **(Milk Duds).** For $n \in \mathbb{N}$, consider the sets

$$\mathsf{T}_n := \left\{ (1 + \log i)^{-1/2} \mathbf{e}_i : i = 1, \dots, n \right\} \subset \mathbb{R}^n.$$

Let $(X_t : t \in \mathsf{T}_n)$ be the centered canonical Gaussian process on T_n.

(a) By direct argument, show that $\mathbb{E} \sup_{t \in \mathsf{T}_n} X_t \leq \text{Const}$.

(b) Compute the covering numbers $\mathrm{N}(\mathsf{T}_n, \ell_2; \varepsilon)$ for small-ish ε.

(c) Instantiate Sudakov's minoration to confirm that $\sup_{t \in \mathsf{T}_n} X_t \geq \text{const}$.

(d) Instantiate Dudley's integral inequality to see that $\sup_{t \in \mathsf{T}_n} X_t \leq \text{const} \cdot \log \log n$.

(e) Show that the generic chaining bound $\gamma_2(\mathsf{T}_n, \ell_2) \leq \text{Const}$.

(f) Write a paragraph to explain the import of these facts.

2 **(Chain Gang).** In this problem, we investigate the relationships among our lower and upper bounds for random processes.

(a) Let $(\mathsf{T}, \mathrm{dist})$ be a metric space. Show that the Dudley sum and integral are equivalent:

$$\begin{aligned} \int_0^\infty \sqrt{\log \mathrm{N}(\mathsf{T}, \mathrm{dist}; \varepsilon)}\, \mathrm{d}\varepsilon &\leq \text{Const} \cdot \sum\nolimits_{k \in \mathbb{Z}} 2^{-k} \sqrt{\log \mathrm{N}(\mathsf{T}, \mathrm{dist}; 2^{-k})} \\ &\leq \text{Const} \cdot \int_0^\infty \sqrt{\log \mathrm{N}(\mathsf{T}, \mathrm{dist}; \varepsilon)}\, \mathrm{d}\varepsilon. \end{aligned}$$

(b) **(Hard Labor).** Let $\mathsf{T} \subset \mathbb{R}^n$. Prove that

$$\begin{aligned} \sup_{\varepsilon > 0} \varepsilon \sqrt{\log \mathrm{N}(\mathsf{T}, \ell_2; \varepsilon)} &\leq \text{Const} \cdot \int_0^\infty \sqrt{\log \mathrm{N}(\mathsf{T}, \ell_2; \varepsilon)}\, \mathrm{d}\varepsilon \\ &\leq \text{Const} \cdot (\log n) \cdot \sup_{\varepsilon > 0} \varepsilon \sqrt{\log \mathrm{N}(\mathsf{T}, \ell_2; \varepsilon)}. \end{aligned}$$

For a canonical Gaussian process, the gap between Sudakov's lower bound and Dudley's lower bound is at worst logarithmic in the dimension. **Hint:**

Use the volumetric estimate to control the growth of the covering numbers as $\varepsilon \to 0$. **(*)** Is the second inequality sharp?

(c) Let $(\mathsf{T}, \mathrm{dist})$ be a metric space. Prove that

$$\gamma_2(\mathsf{T}, \mathrm{dist}) \le \mathrm{Const} \cdot \int_0^\infty \sqrt{\mathrm{N}(\mathsf{T}, \mathrm{dist}; \varepsilon)}\, \mathrm{d}\varepsilon.$$

Conclude that Talagrand's bound is never worse than Dudley's, modulo constants.

Problems

1. **(Férnique–Sudakov–Vitale Comparison).** We used a version of the following result to establish Sudakov's minoration:

> **Theorem 19.4 (Vitale).** Let $(X_t : t \in \mathsf{T})$ and $(Y_t : t \in \mathsf{T})$ be Gaussian processes with the following properties:
>
> $$\begin{aligned} \mathbb{E}\, X_t &= \mathbb{E}\, Y_t && \text{for all } t \in \mathsf{T}; \\ \mathbb{E}(X_s - X_t)^2 &\le \mathbb{E}(Y_s - Y_t)^2 && \text{for all } s, t \in \mathsf{T}. \end{aligned}$$
>
> Then
>
> $$\mathbb{E} \sup_{t \in \mathsf{T}} X_t \le \mathbb{E} \sup_{t \in \mathsf{T}} Y_t.$$

(a) For $\beta > 0$, define the soft-max function

$$f_\beta(\boldsymbol{a}) := \beta^{-1} \log \sum\nolimits_{i=1}^n \mathrm{e}^{\beta a_i} \quad \text{for } \boldsymbol{a} \in \mathbb{R}^n.$$

Check that

$$\max_{1 \le i \le n} a_i \le f_\beta(\boldsymbol{a}) \le \max_{1 \le i \le n} a_i + \beta^{-1} \log n.$$

(b) Compute the first and second partial derivatives of the soft-max function.

(c) Establish Vitale's comparison. **Hint:** Use the Gaussian interpolation result. The proof is easier if you assume that the processes are centered, but a similar argument works in general.

(d) Let $\mathsf{T} \subset \mathbb{R}^n$. Let $\varphi_i : \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz functions. For a standard normal vector $\boldsymbol{g} \in \mathbb{R}^n$, use Vitale's comparison to verify that

$$\mathbb{E} \sup_{\boldsymbol{t} \in \mathsf{T}} \sum\nolimits_{i=1}^n g_i\, \varphi_i(t_i) \le \mathbb{E} \sup_{\boldsymbol{t} \in \mathsf{T}} \sum\nolimits_{i=1}^n g_i t_i.$$

This comparison principle is useful when studying empirical processes.

(e) **(*)** Talagrand established that the same result holds if we replace the Gaussian random variables by iid Rademacher variables (ε_i):

$$\mathbb{E} \sup_{\boldsymbol{t} \in \mathsf{T}} \sum\nolimits_{i=1}^n \varepsilon_i\, \varphi_i(t_i) \le \mathbb{E} \sup_{\boldsymbol{t} \in \mathsf{T}} \sum\nolimits_{i=1}^n \varepsilon_i t_i.$$

Prove it. **Hint:** The argument involves a somewhat tedious case analysis, but you only need to treat a single summand.

(f) (*) Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a matrix. Let $\boldsymbol{D} = \operatorname{diag}(\delta_1, \dots, \delta_n)$, where $\delta_i \sim \textsc{bern}(\delta)$ iid. Use centering, symmetrization, and the Rademacher comparison from (e) to show that

$$\mathbb{E}\,\|\boldsymbol{DA}\|_{2\to 1} \le \delta\|\boldsymbol{A}\|_{2\to 1} + \sqrt{2\delta(1-\delta)}\|\boldsymbol{A}\|_{\mathrm{F}}.$$

Recall that $\|\boldsymbol{A}\|_{2\to 1} \coloneqq \sup_{\|\boldsymbol{u}\|_2=1} \|\boldsymbol{Au}\|_1$. In other words, a random row submatrix inherits "its share" of the $2 \to 1$ operator norm.

Applications

1. **(Duck and Cover).** The *Hamming cube* is the set $\mathsf{H}^n := \{0,1\}^n$ of bit strings of length n, equipped with the metric

$$\operatorname{dist}_{\mathsf{H}}(\boldsymbol{x}, \boldsymbol{y}) := \#\{i : x_i \neq y_i\} \quad \text{for } \boldsymbol{x}, \boldsymbol{y} \in \mathsf{H}^n.$$

A *Hamming ball* is a set of the form $\mathsf{B}_r(\boldsymbol{x}) := \{\boldsymbol{y} \in \mathsf{H}^n : \operatorname{dist}_{\mathsf{H}}(\boldsymbol{x}, \boldsymbol{y}) \le r\}$, where the point $\boldsymbol{x} \in \mathsf{H}^n$ and the radius $r > 0$.

(a) Verify that $\operatorname{dist}_{\mathsf{H}}$ is a metric.

(b) For $r \in \mathbb{N}$ and $\boldsymbol{x} \in \mathsf{H}^n$, establish bounds for the cardinality of a Hamming ball:

$$\left(\frac{n}{r}\right)^r \le \#\mathsf{B}_r(\boldsymbol{x}) \le \left(\frac{\mathrm{e}n}{r}\right)^r.$$

Hint: Recall that $\mathrm{e}^r \ge \sum_{i=0}^k r^i/i!$ for $r \ge 0$.

(c) Use a volumetric argument to prove that

$$2^n \left(\frac{r}{\mathrm{e}n}\right)^r \le \mathscr{N}(\mathsf{H}^n, \operatorname{dist}_{\mathsf{H}}; r) \le \mathscr{P}(\mathsf{H}^n, \operatorname{dist}_{\mathsf{H}}; r/2) \le 2^n \left(\frac{r}{2n}\right)^{r/2}.$$

Fix natural numbers r, k, n with $k \le n$. An *error correcting code* consists of an *encoding map* $E : \mathsf{H}^k \to \mathsf{H}^n$ that assigns an n-bit codeword to a k-bit message, along with a *decoding map* $D : \mathsf{H}^n \to \mathsf{H}^k$ that maps an n-bit string back to a k-bit message. The code is resilient to arbitrary errors in r bits of a codeword if and only if

$$D(\boldsymbol{y}) = \boldsymbol{x} \quad \text{for each } \boldsymbol{x} \in \mathsf{H}^k \text{ and each } \boldsymbol{y} \in \mathsf{B}_r(E(\boldsymbol{x})).$$

We say that such a code has parameters (r, k, n).

(d) Assume that $\log_2 \mathscr{P}(\mathsf{H}^n, \operatorname{dist}_{\mathsf{H}}; r) \ge k$. Prove that there exists an error correcting code with the parameters (r, k, n).

(e) Assume that $n \ge k + 2r\log_2(\mathrm{e}n/(2r))$. Show that there exists a code with parameters (r, k, n).

(f) State and prove converses to the last two results.

(g) Express these results in terms of the *rate* $R := k/n$ and the fraction of errors $\delta := r/n$.

2. **(Phase Transitions).** In this problem, we will establish another important Gaussian comparison inequality due to Yehoram Gordon, as well as a modern variant obtained by graduate students at Caltech. As an application, we will investigate a remarkable phase transition phenomenon that holds in geometric probability.

Theorem 19.5 (Gordon). Let $\{X_{uv} : u \in \mathsf{U}, v \in \mathsf{V}\}$ and $\{Y_{uv} : u \in \mathsf{U}, v \in \mathsf{V}\}$ be centered Gaussian processes on a compact metric space $(\mathsf{U} \times \mathsf{V}, \mathrm{dist})$. Assume that

$$\begin{aligned}
\mathbb{E}(X_{uv} - X_{uv'})^2 &\le \mathbb{E}(Y_{uv} - Y_{uv'})^2 && \text{for all } u \in \mathsf{U} \text{ and } v, v' \in \mathsf{V};\\
\mathbb{E}(X_{uv} - X_{u'v'})^2 &\ge \mathbb{E}(Y_{uv} - Y_{u'v'})^2 && \text{for all } u \ne u' \text{ in } \mathsf{U} \text{ and all } v, v' \in \mathsf{V};\\
\mathbb{E}\, X_{uv}^2 &= \mathbb{E}\, Y_{uv}^2 && \text{for all } u \in \mathsf{U} \text{ and } v \in \mathsf{V}.
\end{aligned}$$

For all $\tau \in \mathbb{R}$,

$$\mathbb{P}\left\{\min_{u\in\mathsf{U}} \max_{v\in\mathsf{V}} X_{uv} \ge \tau\right\} \le \mathbb{P}\left\{\min_{u\in\mathsf{U}} \max_{v\in\mathsf{V}} Y_{uv} \ge \tau\right\}.$$

In particular,

$$\mathbb{E} \min_{u\in\mathsf{U}} \max_{v\in\mathsf{V}} X_{uv} \le \mathbb{E} \min_{u\in\mathsf{U}} \max_{v\in\mathsf{V}} Y_{uv}.$$

(a) Use Kahane's theorem to prove Gordon's theorem.

(b) Let $\mathsf{U} \subset \mathbb{R}^m$ and $\mathsf{V} \subset \mathbb{R}^n$. Let $\boldsymbol{\Gamma} \in \mathbb{R}^{n\times m}$ be a standard normal matrix. Let $\boldsymbol{g} \in \mathbb{R}^m$ and $\boldsymbol{h} \in \mathbb{R}^n$ be independent standard normal. Mimic the proof of Chevet's theorem to get

$$\mathbb{P}\left\{\min_{\boldsymbol{u}\in U} \max_{\boldsymbol{v}\in V} \langle \boldsymbol{\Gamma u},\ \boldsymbol{v}\rangle < \tau\right\} \le 2\mathbb{P}\left\{\min_{\boldsymbol{u}\in U} \max_{\boldsymbol{v}\in V} \left[\|\boldsymbol{v}\|_2 \langle \boldsymbol{g},\ \boldsymbol{u}\rangle + \|\boldsymbol{u}\|_2 \langle \boldsymbol{h},\ \boldsymbol{v}\rangle\right] \le \tau\right\}.$$

Furthermore,

$$\mathbb{E} \min_{\boldsymbol{u}\in\mathsf{U}} \max_{\boldsymbol{v}\in\mathsf{V}} \langle \boldsymbol{\Gamma u},\ \boldsymbol{v}\rangle \ge \mathbb{E} \min_{\boldsymbol{u}\in\mathsf{U}} \max_{\boldsymbol{v}\in\mathsf{V}} \left[\|\boldsymbol{v}\|_2 \langle \boldsymbol{g},\ \boldsymbol{u}\rangle + \|\boldsymbol{u}\|_2 \langle \boldsymbol{h},\ \boldsymbol{v}\rangle\right].$$

(c) Show that the mth largest singular value of the standard normal matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{n\times m}$ satisfies

$$\mathbb{E}\, \sigma_m(\boldsymbol{\Gamma}) \ge \sqrt{n-1} - \sqrt{m}.$$

This inequality is numerically sharp. **(**)** Argue that we can replace $\sqrt{n-1}$ by $\sqrt{n}$.

(d) **(*)** Assume that U and V are both convex. Establish the reversed inequality

$$\mathbb{P}\left\{\min_{\boldsymbol{u}\in\mathsf{U}} \max_{\boldsymbol{v}\in\mathsf{V}} \langle \boldsymbol{\Gamma u},\ \boldsymbol{v}\rangle > \tau\right\} \le 2\mathbb{P}\left\{\min_{\boldsymbol{u}\in\mathsf{U}} \max_{\boldsymbol{v}\in\mathsf{V}} \left[\|\boldsymbol{v}\|_2 \langle \boldsymbol{g},\ \boldsymbol{u}\rangle + \|\boldsymbol{u}\|_2 \langle \boldsymbol{h},\ \boldsymbol{v}\rangle\right] \ge \tau\right\}.$$

Hint: Consider the negation of the random processes, invoke Sion's theorem, repeat the proof of (b), and then use the inf–sup inequality to wrestle the

result into the stated form. In this form, the result is due to Thrampoulidis, Oymak, & Hassibi.

The Gaussian minimax theorem has remarkable applications in mathematical signal processing. We will focus on a relatively simple geometric example.

(e) Let $\mathsf{T} \subset \mathbb{R}^n$ be a compact, convex subset of the Euclidean unit ball. Let $\mathsf{L}_d \subset \mathbb{R}^n$ be a uniformly random subspace with *codimension* d. Consider the probability that the random subspace misses the set:

$$p(d) := \mathbb{P}\{\mathsf{T} \cap \mathsf{L}_d = \emptyset\}.$$

Explain how to express this event as the minimax of a bilinear form in a standard normal matrix. **Hint:** We can realize L_d as the null space of a standard normal matrix $\mathbf{\Gamma} \in \mathbb{R}^{n \times d}$ (a.s.).

(f) We define the excess width functional as

$$\mathscr{E}_d(\mathsf{T}) := \mathbb{E} \min_{\boldsymbol{t} \in \mathsf{T}} \left(\sqrt{d} \|\boldsymbol{t}\|_2 - \langle \boldsymbol{g}, \boldsymbol{t} \rangle \right).$$

As usual, $\boldsymbol{g} \in \mathbb{R}^n$ is a standard normal vector. Show that the excess width is a monotone increasing function of d.

(g) Develop and prove a suitable form of the following statement:

$$p(d) \approx \begin{cases} 0, & \mathscr{E}_d(\mathsf{T}) - \text{const} < 0; \\ 1, & \mathscr{E}_d(\mathsf{T}) + \text{const} > 0. \end{cases}$$

In other words, there is a *phase transition* in the probability that a d-codimensional subspace misses the set T. **Hint:** The excess width is the expectation of a Lipschitz function of a Gaussian vector, so it must concentrate.

(h) **(*)** The probability simplex $\Delta_n := \{\boldsymbol{t} \in \mathbb{R}^n : t_i \geq 0 \text{ and } \sum_i t_i = 1\}$. Design and execute a computer experiment to witness the existence of the phase transition phenomenon for the probability simplex. For a range of dimensions n, plot the empirical miss probability as a function of d/n.

(i) **(**)** Calculate the asymptotic excess width of the probability simplex as $d, n \to \infty$ and $d/n \to \text{const}$. Superimpose this curve on your plots.

back matter

Bibliography

[AW02] R. Ahlswede and A. Winter. "Strong converse for identification via quantum channels". In: *IEEE Transactions on Information Theory* 48.3 (2002), pages 569–579.

[Bar05] A. Barvinok. "Measure Concentration". Math 710, University of Michigan. 2005. URL: http://www.math.lsa.umich.edu/~barvinok/total710.pdf.

[BL14] W. Bednorz and R. Latała. "On the boundedness of Bernoulli processes". In: *Annals of Mathematics* 180.3 (2014), pages 1167–1203.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. A nonasymptotic theory of independence, With a foreword by Michel Ledoux. Oxford University Press, Oxford, 2013. DOI: 10.1093/acprof:oso/9780199535255.001.0001.

[COS20] Y. Chen, H. Owhadi, and A. M. Stuart. "Consistency of Empirical Bayes And Kernel Flow For Hierarchical Parameter Estimation". In: *arXiv preprint arXiv:2005.11375* (2020).

[Che77] S. Chevet. "Séries de variables aléatoires Gaussiennes à valeurs dans $E\hat{\otimes}_\varepsilon F$. Application aux produits d'espaces de Wiener abstraits". In: *Séminaire sur ia Géométrie des Espaces de Banach (1977-1978), Exp. No. 19, 15, École Polytechnique, Palaiseau* (1977).

[Fer75] X. Fernique. "Regularité des trajectoires des fonctions aléatoires gaussiennes". In: *Ecole d'Eté de Probabilités de Saint-Flour IV—1974*. Springer Berlin Heidelberg, 1975, pages 1–96.

[FR13] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*. Birkhäuser/Springer, New York, 2013. DOI: 10.1007/978-0-8176-4948-7.

[Gor88] Y. Gordon. "On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$". In: *Geometric Aspects of Functional Analysis*. Springer Berlin Heidelberg, 1988, pages 84–106.

[Gor85] Y. Gordon. "Some Inequalities for Gaussian Processes and Applications". In: *Israel Journal of Mathematics* 50.4 (1985), pages 265–289.

[Haa81] U. Haagerup. "The best constants in the Khintchine inequality". eng. In: *Studia Mathematica* 70.3 (1981), pages 231–283. URL: http://eudml.org/doc/218383.

[HMT11] N. Halko, P. G. Martinsson, and J. A. Tropp. "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions". In: *SIAM Review* 53.2 (Jan. 2011), pages 217–288. DOI: 10.1137/090771806.

[Han18a] R. van Handel. "Chaining, interpolation, and convexity". In: *J. Eur. Math. Soc. (JEMS)* 20.10 (2018), pages 2413–2435. DOI: 10.4171/JEMS/815.

[Han18b] R. van Handel. "Chaining, interpolation and convexity II: The contraction principle". In: *Ann. Probab.* 46.3 (2018), pages 1764–1805. DOI: 10.1214/17-AOP1214.

[Kah86] J.-P. Kahane. "Une inegalité du type de Slepian et Gordon sur les processus gaussiens". In: *Israel Journal of Mathematics* 55.1 (Feb. 1986), pages 109–110. DOI: 10.1007/BF02772698.

[KM15] V. Koltchinskii and S. Mendelson. "Bounding the smallest singular value of a random matrix without concentration". In: *International Mathematics Research Notices* 2015.23 (2015), pages 12991–13008.

[LO94] R. Latała and K. Oleszkiewicz. "On the best constant in the Khinchin–Kahane inequality". In: *Studia Math.* 109.1 (1994), pages 101–104.

[Led01] M. Ledoux. *The concentration of measure phenomenon*. American Mathematical Society, Providence, RI, 2001. DOI: 10.1090/surv/089.

[LT11] M. Ledoux and M. Talagrand. *Probability in Banach spaces*. Isoperimetry and processes, Reprint of the 1991 edition. Springer-Verlag, Berlin, 2011.

[Lie73] E. H. Lieb. "Convex trace functions and the Wigner–Yanase–Dyson conjecture". In: *Les rencontres physiciens-mathématiciens de Strasbourg-RCP25* 19 (1973), pages 0–35.

[LP86] F. Lust-Piquard. "Inégalités de Khintchine dans C_p $(1 < p < \infty)$". In: *CR Acad. Sci. Paris* 303 (1986), pages 289–292.

[Men14a] S. Mendelson. "A remark on the diameter of random sections of convex bodies". In: *Geometric aspects of functional analysis*. Springer, 2014, pages 395–404.

[Men14b] S. Mendelson. "Learning without concentration". In: *Conference on Learning Theory*. PMLR. 2014, pages 25–39.

[Oli10] R. Oliveira. "Sums of random Hermitian matrices and an inequality by Rudelson". In: *Electronic Communications in Probability* 15 (2010), pages 203–212.

[PS08] G. Pavliotis and A. Stuart. *Multiscale methods: averaging and homogenization*. Springer Science & Business Media, 2008.

[Paz12] A. Pazy. *Semigroups of linear operators and applications to partial differential equations*. Springer Science & Business Media, 2012.

[Rob20] J. C. Robinson. *An Introduction to Functional Analysis*. Cambridge University Press, 2020.

[Ros11] N. Ross. "Fundamentals of Stein's method". In: *Probab. Surv.* 8 (2011), pages 210–293. DOI: 10.1214/11-PS182.

[Sle62] D. Slepian. "The one-sided barrier problem for Gaussian noise". In: *The Bell System Technical Journal* 41.2 (1962), pages 463–501. DOI: 10.1002/j.1538-7305.1962.tb02419.x.

[Tal96] M. Talagrand. "A new look at independence". In: *Ann. Probab.* 24.1 (1996), pages 1–34.

[Tal14] M. Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*. Springer Science & Business Media, 2014.

[Tal+96] M. Talagrand et al. "Majorizing measures: the generic chaining". In: *The Annals of Probability* 24.3 (1996), pages 1049–1103.

[TOH14] C. Thrampoulidis, S. Oymak, and B. Hassibi. "A Tight Version of the Gaussian min-max theorem in the Presence of Convexity". In: *CoRR* abs/1408.4837 (2014). URL: http://arxiv.org/abs/1408.4837. arXiv: 1408.4837.

[Tro15a] J. A. Tropp. "An introduction to matrix concentration inequalities". In: *Found. Trends Mach. Learn.* 8.1–2 (2015), pages 1–230.

[Tro17] J. A. Tropp. "ACM 217: Lecture notes on concentration inequalities". Available on request. 2017.

[Tro19] J. A. Tropp. "Matrix concentration and computational linear algebra". Caltech CMS Lecture Notes 2019-01. 2019.

[Tro15b] J. A. Tropp. "Convex recovery of a structured signal from independent random linear measurements". In: *Sampling Theory, a Renaissance*. Springer, 2015, pages 67–101.

[van16] R. van Handel. "Probability in High Dimensions". APC 550 Lecture Notes, Princeton Univ. 2016. URL: https://web.math.princeton.edu/~rvan/APC550.pdf.

[van17] R. van Handel. "Structured random matrices". In: *Convexity and concentration*. Volume 161. IMA Vol. Math. Appl. Springer, New York, 2017, pages 107–156.

[Ver12] R. Vershynin. "Introduction to the non-asymptotic analysis of random matrices". In: *Compressed sensing*. Cambridge Univ. Press, Cambridge, 2012, pages 210–268.

[Ver18] R. Vershynin. *High-dimensional probability*. An introduction with applications in data science, With a foreword by Sara van de Geer. Cambridge University Press, Cambridge, 2018. DOI: 10.1017/9781108231596.

[Wai19] M. J. Wainwright. *High-dimensional statistics*. A non-asymptotic viewpoint. Cambridge University Press, Cambridge, 2019. DOI: 10.1017/9781108627771.

[Wei12] H. F. Weinberger. *A first course in partial differential equations: with complex variables and transform methods*. Courier Corporation, 2012.