

Metric entropy and its uses

Many statistical problems require manipulating and controlling collections of random variables indexed by sets with an infinite number of elements. There are many examples of such stochastic processes. For instance, a continuous-time random walk can be viewed as a collection of random variables indexed by the unit interval $[0, 1]$. Other stochastic processes, such as those involved in random matrix theory, are indexed by vectors that lie on the Euclidean sphere. Empirical process theory, an area that includes the Glivenko-Cantelli laws discussed in Chapter 4, is concerned with stochastic processes that are indexed by sets of functions.

3
4
5
6
7
8
9
10

Whereas any finite set can be measured in terms of its cardinality, measuring the “size” of a set with infinitely many elements requires more delicacy. The concept of metric entropy, which dates back to the seminal work of Kolmogorov, Tikhomirov and others in the Russian school, provides one way in which to address this difficulty. Though defined in a purely deterministic manner, in terms of packing and covering in a metric space, it plays a central role in understanding the behavior of stochastic processes. Accordingly, this chapter is devoted to an exploration of metric entropy, and its various uses in the context of stochastic processes.

11
12
13
14
15
16
17
18

■ 5.1 Covering and packing

We begin by defining the notions of packing and covering a set in a metric space. Recall that a metric space (\mathbb{T}, ρ) consists of a non-empty set \mathbb{T} , equipped with a mapping $\rho : \mathbb{T} \times \mathbb{T} \rightarrow \mathbb{R}$ that satisfies the following properties:

20
21
22

(a) it is non-negative: $\rho(\theta, \theta') \geq 0$ for all pairs (θ, θ') , with equality if and only if $\theta = \theta'$.

23
24

(b) it is symmetric: $\rho(\theta, \theta') = \rho(\theta', \theta)$ for all pairs (θ', θ) , and

25

(c) the triangle inequality holds: i.e., $\rho(\theta, \theta') \leq \rho(\theta, \tilde{\theta}) + \rho(\tilde{\theta}, \theta')$ for all triples $(\theta, \theta', \tilde{\theta})$.

26

Familiar examples of metric spaces include the real space \mathbb{R}^d with the *Euclidean metric*

$$\rho(\theta, \theta') = \|\theta - \theta'\|_2 := \sqrt{\sum_{j=1}^d (\theta_j - \theta'_j)^2}, \quad (5.1)$$

and the discrete cube $\{0, 1\}^d$ with the *rescaled Hamming metric*

$$\rho_H(\theta, \theta') := \frac{1}{d} \sum_{j=1}^d \mathbb{I}[\theta_j \neq \theta'_j]. \quad (5.2)$$

We will also be interested in various metric spaces of functions, among them the usual spaces $L^2(\mu, [0, 1])$ with its metric

$$\|f - g\|_2 := \left[\int_0^1 (f(x) - g(x))^2 d\mu(x) \right]^{1/2}, \quad (5.3)$$

as well as the space $C[0, 1]$ of all continuous functions on $[0, 1]$ equipped with the sup-norm metric

$$\|f\|_\infty = \sup_{x \in [0, 1]} |f(x)|. \quad (5.4)$$

Given a metric space (\mathbb{T}, ρ) , a natural way in which to measure its size is in terms of number of balls of a fixed radius δ required to cover it, a quantity known as the covering number.

Definition 5.1 (Covering number). A δ -cover of a set \mathbb{T} with respect to a metric ρ is a set $\{\theta^1, \dots, \theta^N\} \subset \mathbb{T}$ such that for each $\theta \in \mathbb{T}$, there exists some $i \in \{1, \dots, N\}$ such that $\rho(\theta, \theta^i) \leq \delta$. The δ -covering number $N(\delta; \mathbb{T}, \rho)$ is the cardinality of the smallest δ -cover.

As illustrated in Figure 5-1(a), a δ -covering can be visualized as a collection of balls of radius δ that cover the set \mathbb{T} . When discussing metric entropy, we restrict our attention to metric spaces (\mathbb{T}, ρ) that are *totally bounded*, meaning that $N(\delta) = N(\delta; \mathbb{T}, \rho)$ is finite for all $\delta > 0$. See Exercise 5.1 for an example of a metric space that is *not* totally bounded.

It is easy to see that the covering number is non-increasing in δ , meaning that $N(\delta) \geq N(\delta')$ for all $\delta \leq \delta'$. Typically, the covering number diverges as $\delta \rightarrow 0^+$, and of interest to us is this growth rate on a logarithmic scale. More specifically, the quantity $\log N(\delta; \mathbb{T}, \rho)$ is known as the *metric entropy* of the set \mathbb{T} with respect to ρ .

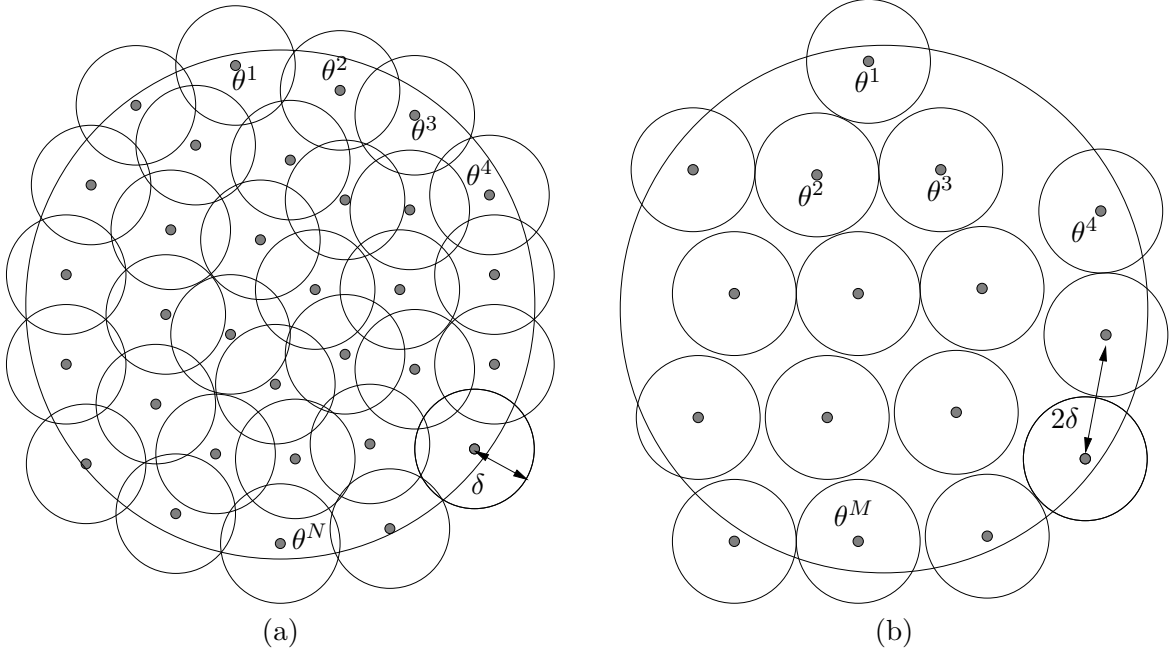


Figure 5-1. Illustration of packing and covering sets. (a) A δ -covering of T is a collection of elements $\{\theta^1, \dots, \theta^N\} \subset T$ such that for each $\theta \in T$, there is some element $j \in \{1, \dots, N\}$ such that $\rho(\theta, \theta^j) \leq \delta$. Geometrically, the union of the balls with centers θ^j and radius δ cover the set T . (b) A δ -packing of a set T is a collection of elements $\{\theta^1, \dots, \theta^M\} \subset T$ such that $\rho(\theta^j, \theta^k) > 2\delta$ for all $j \neq k$. Geometrically, it is a collection of balls of radius δ with centers contained in T such that no pair of balls have a non-empty intersection.

Example 5.1 (Covering numbers of unit cubes). Let us begin with a simple example of how covering numbers can be bounded. Consider the interval $[-1, 1]$ in \mathbb{R} , equipped with the metric $\rho(\theta, \theta') = |\theta - \theta'|$. Suppose that we divide the interval $[-1, 1]$ into $L := \lfloor \frac{1}{\delta} \rfloor + 1$ sub-intervals, centered at the points $\theta^i = -1 + 2(i-1)\delta$ for $i \in [L] = \{1, 2, \dots, L\}$, and each of length at most 2δ . By construction, for any point $\theta' \in [0, 1]$, there is some $j \in [L]$ such that $|\theta^j - \theta'| \leq \delta$, which shows that

$$N(\delta; [-1, 1], |\cdot|) \leq \frac{1}{\delta} + 1. \quad (5.5)$$

As an exercise, the reader should generalize this analysis, showing that for the d -dimensional cube $[-1, 1]^d$, we have $N(\delta; [-1, 1]^d, \|\cdot\|_\infty) \leq (1 + \frac{1}{\delta})^d$. ♣

Example 5.2 (Covering of the binary hypercube). Consider the binary hypercube $\mathbb{H}^d := \{0, 1\}^d$ equipped with the rescaled Hamming metric (5.2). First, let us upper bound its δ -covering number. Let $S = \{1, 2, \dots, \lceil (1 - \delta)d \rceil\}$, where $\lceil (1 - \delta)d \rceil$ denotes

the smallest integer larger than or equal to $(1 - \delta)d$. Consider the set of binary vectors

$$\mathbb{T}(\delta) := \{\theta \in \mathbb{H}^d \mid \theta_j = 0 \text{ for all } j \notin S\}.$$

By construction, for any binary vector $\theta' \in \mathbb{H}^d$, we can find a vector $\theta \in \mathbb{T}(\delta)$ such $\rho_H(\theta, \theta') \leq \delta$. (Indeed, we can match θ' exactly on all entries $j \in S$, and in the worst case, disagree on all the remaining $\lfloor \delta d \rfloor$ positions.) Since $\mathbb{T}(\delta)$ contains $2^{\lfloor (1-\delta)d \rfloor}$ vectors, we conclude that

$$\frac{\log N_H(\delta; \mathbb{H}^d)}{\log 2} \leq \lceil d(1 - \delta) \rceil.$$

- 1 This bound is useful but can be sharpened considerably by using a more refined argu-
2 ment (see Exercise 5.3).

Let us lower bound its δ -covering number, where $\delta \in (0, 1/2)$. If $\{\theta^1, \dots, \theta^N\}$ is a δ -covering, then the (unrescaled) Hamming balls of radius $s = \delta d$ around each θ^ℓ must contain all 2^d vectors in the binary hypercube. Let $s = \lfloor \delta d \rfloor$ denote the largest integer less than or equal to δd . For each θ^ℓ , there are exactly $\sum_{j=0}^s \binom{d}{j}$ binary vectors lying within distance δd from it, and hence we must have $N \sum_{j=0}^s \binom{d}{j} \geq 2^d$. Now let $X_i \in \{0, 1\}$ be i.i.d. Bernoulli variables with parameter $1/2$. Re-arranging the previous inequality, we have

$$\frac{1}{N} \leq \sum_{j=0}^s \binom{d}{j} 2^{-d} = \mathbb{P}\left[\sum_{i=1}^d X_i \leq \delta d\right] \stackrel{(i)}{\leq} \exp\left(-\frac{d}{2}\left(\frac{1}{2} - \delta\right)^2\right),$$

where inequality (i) follows from Hoeffding's bound applied to the sum of i.i.d. Bernoulli variables. Following some algebra, we obtain the lower bound

$$\log N_H(\delta; \mathbb{H}^d) \geq \frac{d}{2} (1/2 - \delta)^2, \quad \text{valid for } \delta \in (0, 1/2).$$

This lower bound is qualitatively correct, but can be tightened by using a better upper bound on the binomial tail probability. For instance, from the result of Exercise 2.9, we have $\frac{1}{d} \log \mathbb{P}[\sum_{i=1}^d X_i \leq s] \leq -D(\delta \parallel 1/2)$, where $D(\delta \parallel 1/2)$ is the Kullback-Leibler divergence between the Bernoulli distributions with parameters δ and $1/2$ respectively. Using this tail bound within the same argument leads to the improved lower bound

$$\log N_H(\delta; \mathbb{H}^d) \geq d D(\delta \parallel 1/2), \quad \text{valid for } \delta \in (0, 1/2). \quad (5.6)$$

3



- 4 In the preceding examples, we used different techniques to upper and lower bound
5 the covering number. A complementary way in which to measure the massiveness of

sets, also useful for deriving bounds on the metric entropy, is known as the packing number.

Definition 5.2 (Packing number). A δ -packing of a set \mathbb{T} with respect to a metric ρ is a set $\{\theta^1, \dots, \theta^M\} \subset \mathbb{T}$ such that $\rho(\theta^i, \theta^j) > \delta$ for all distinct $i, j \in \{1, 2, \dots, M\}$. The δ -packing number $M(\delta; \mathbb{T}, \rho)$ is the cardinality of the largest δ -packing.

As illustrated in Figure 5-1(b), a δ -packing can be viewed as a collection of balls of radius δ , each centered at an element contained in \mathbb{T} , such that no two balls intersect. What is the relation between the covering number and packing numbers? Although not identical, they provide essentially the same measure of the massiveness of a set, as summarized in the following:

Lemma 5.1. For all $\delta > 0$, the packing and covering numbers are related as follows:

$$M(2\delta; \mathbb{T}, \rho) \stackrel{(a)}{\leq} N(\delta; \mathbb{T}, \rho) \stackrel{(b)}{\leq} M(\delta; \mathbb{T}, \rho). \quad (5.7)$$

We leave the proof of Lemma 5.1 for the reader (see Exercise 5.2). It shows that at least up to constant factors, the packing and covering numbers exhibit the same scaling behavior as $\delta \rightarrow 0$.

Example 5.3 (Packing of unit cubes). Returning to Example 5.1, we observe that the points $\{\theta^j, j = 1, \dots, L-1\}$ satisfy $|\theta^j - \theta^k| \geq 2\delta > \delta$ for all $j \neq k$, which implies that $M(2\delta; [-1, 1], |\cdot|) \geq \lfloor \frac{1}{\delta} \rfloor$. Combined with Lemma 5.1 and our previous upper bound (5.5), we conclude that $\log N(\delta; [-1, 1], |\cdot|) \asymp \log(1/\delta)$ for $\delta > 0$ sufficiently small. This argument extended to the d -dimensional cube with the sup-norm $\|\cdot\|_\infty$, showing that

$$\log N(\delta; [0, 1]^d, \|\cdot\|_\infty) \asymp d \log(1/\delta) \quad \text{for } \delta > 0 \text{ sufficiently small.} \quad (5.8)$$

Thus, we see how an explicit construction of a packing set can be used to lower bound the metric entropy. ♣

In Exercise 5.3, we show how a packing argument can be used to obtain a refined upper bound on the covering number of the Boolean hypercube from Example 5.2.

We now seek some more general understanding of what geometric properties govern metric entropy. Since covering is defined in terms of the number of balls—each with a fixed radius and hence volume—one would expect to see connections between covering numbers and volumes of these balls. The following lemma provides a precise statement of this connection in the case of norms on \mathbb{R}^d with open unit balls, for which the volume can be taken with respect to Lebesgue measure. Important examples are the usual

ℓ_q -balls, defined for $q \in [1, \infty]$ via

$$\mathbb{B}_q^d(1) := \{x \in \mathbb{R}^d \mid \|x\|_q \leq 1\}, \quad (5.9)$$

where for $q \in [1, \infty)$, the ℓ_q -norm is given by

$$\|x\|_q := \begin{cases} (\sum_{i=1}^d |x_i|^q)^{1/q} & \text{for } q \in [1, \infty), \text{ and} \\ \max_{i=1, \dots, d} |x_i| & \text{for } q = \infty. \end{cases} \quad (5.10)$$

1 The following lemma relates the metric entropy to the so-called volume ratio:

2

Lemma 5.2 (Volume ratios and metric entropy). Consider a pair of norms $\|\cdot\|$ and $\|\cdot\|'$ on \mathbb{R}^d , and let \mathbb{B} and \mathbb{B}' be their corresponding unit balls (i.e., $\mathbb{B} = \{\theta \in \mathbb{R}^d \mid \|\theta\| \leq 1\}$, with \mathbb{B}' similarly defined). Then the δ -covering number of \mathbb{B} in the $\|\cdot\|'$ norm obeys the bounds

3

$$\left(\frac{1}{\delta}\right)^d \frac{\text{vol}(\mathbb{B})}{\text{vol}(\mathbb{B}')} \stackrel{(a)}{\leq} N(\delta; \mathbb{B}, \|\cdot\|') \stackrel{(b)}{\leq} \frac{\text{vol}(\frac{2}{\delta}\mathbb{B} + \mathbb{B}')}{\text{vol}(\mathbb{B}')} \quad (5.11)$$

4

Whenever $\mathbb{B}' \subseteq \mathbb{B}$, the upper bound (b) may be simplified by observing that

$$\text{vol}(\frac{2}{\delta}\mathbb{B} + \mathbb{B}') \leq \text{vol}((\frac{2}{\delta} + 1)\mathbb{B}) = (\frac{2}{\delta} + 1)^d \text{vol}(\mathbb{B}),$$

5 which implies that $N(\delta; \mathbb{B}, \|\cdot\|') \leq (1 + \frac{2}{\delta})^d \frac{\text{vol}(\mathbb{B})}{\text{vol}(\mathbb{B}')}.$

Proof. On one hand, if $\{\theta^1, \dots, \theta^N\}$ is an δ -covering of \mathbb{B} , then we have

$$\mathbb{B} \subseteq \cup_{j=1}^N \{\theta^j + \delta\mathbb{B}'\},$$

6 which implies that $\text{vol}(\mathbb{B}) \leq N \text{vol}(\delta\mathbb{B}') = N \delta^d \text{vol}(\mathbb{B}')$, thus establishing the claim (5.11)(a).

To establish the upper bound (5.11)(b), let $\{\theta^1, \dots, \theta^M\}$ be a maximal δ -packing of \mathbb{B} in the $\|\cdot\|'$ norm; by maximality, this set must also be a δ -covering of \mathbb{B} under the $\|\cdot\|'$ norm. The balls $\{\theta^j + \frac{\delta}{2}\mathbb{B}', j = 1, \dots, M\}$ are all disjoint and contained within $\mathbb{B} + \frac{\delta}{2}\mathbb{B}'$. Taking volumes, we conclude that $\sum_{j=1}^M \text{vol}(\theta^j + \frac{\delta}{2}\mathbb{B}') \leq \text{vol}(\mathbb{B} + \frac{\delta}{2}\mathbb{B}')$, and hence

$$M \text{vol}(\frac{\delta}{2}\mathbb{B}') \leq \text{vol}(\mathbb{B} + \frac{\delta}{2}\mathbb{B}').$$

7 Finally, we have $\text{vol}(\frac{\delta}{2}\mathbb{B}') = (\frac{\delta}{2})^d \text{vol}(\mathbb{B}')$ and $\text{vol}(\mathbb{B} + \frac{\delta}{2}\mathbb{B}') = (\frac{\delta}{2})^d \text{vol}(\frac{2}{\delta}\mathbb{B} + \mathbb{B}')$, from
8 which the claim (5.11)(b) follows.

9

□

Let us illustrate Lemma 5.2 with an example. 1

Example 5.4 (Covering unit balls in their own metrics). As an important special case, if we take $\mathbb{B} = \mathbb{B}'$ in Lemma 5.2, then we obtain upper and lower bounds on the metric entropy of a given unit ball in terms of its own norm—namely, we have

$$d \log \left(\frac{1}{\delta} \right) \leq \log N(\delta; \mathbb{B}, \|\cdot\|) \leq d \log \left(1 + \frac{2}{\delta} \right). \quad (5.12)$$

When applied to the ℓ_∞ -norm, this result shows that the $\|\cdot\|_\infty$ -metric entropy of $\mathbb{B}_\infty^d = [-1, 1]^d$ scales as $d \log(1/\delta)$, so that we immediately recover the end result of our more direct analysis in Examples 5.1 and 5.3. As another special case, we also find that the Euclidean unit ball \mathbb{B}_2^d can be covered by at most $(1 + 2/\delta)^d$ balls with radius δ in the norm $\|\cdot\|_2$. In Example 5.8 to follow in the sequel, we use Lemma 5.2 to bound the metric entropy of certain ellipses in $\ell^2(\mathbb{N})$. ♣

Thus far, we have studied the metric entropy of various subsets of \mathbb{R}^d and $\ell^2(\mathbb{N})$. Now let us consider the metric entropy of some function classes. We begin with a simple parametric class of functions. 10

Example 5.5 (A parametric class of functions). For any fixed θ , define the real-valued function $f_\theta(x) := 1 - e^{-\theta|x|}$, and consider the function class 11

$$\mathcal{P} = \{f_\theta : [0, 1] \rightarrow \mathbb{R} \mid \theta \in [0, 1]\}. \quad 13$$

The set \mathcal{P} is a metric space under the uniform norm (also known as the sup-norm) given by $\|f - g\|_\infty := \sup_{x \in [0, 1]} |f(x) - g(x)|$. We claim that the covering in terms of the sup-norm is bounded above and below as¹

$$1 + \left\lfloor \frac{1 - 1/e}{2\delta} \right\rfloor \leq N_\infty(\delta; \mathcal{P}) \leq \frac{1}{2\delta} + 2.$$

Let us first establish the upper bound. For a given $\delta \in (0, 1)$, let us set $T = \lfloor \frac{1}{2\delta} \rfloor$, and define $\theta^i = 2\delta i$ for $i = 0, 1, \dots, T$. By also adding the point $\theta^{T+1} = 1$, we obtain a collection of points $\{\theta^0, \dots, \theta^T, \theta^{T+1}\}$ contained within $[0, 1]$. We claim that the associated functions $\{f_{\theta^0}, \dots, f_{\theta^{T+1}}\}$ form an δ -cover for \mathcal{P} . Indeed, for any $f_\theta \in \mathcal{P}$, we can find some θ^i in our cover such that $|\theta^i - \theta| \leq \delta$. We then have 14

$$\begin{aligned} \|f_{\theta^i} - f_\theta\|_\infty &= \max_{x \in [0, 1]} |e^{-\theta^i|x|} - e^{-\theta|x|}| \\ &\leq |\theta^i - \theta| \leq \delta. \end{aligned} \quad \begin{matrix} 19 \\ 20 \end{matrix}$$

Therefore, we can conclude that $N_\infty(\delta; \mathcal{P}) \leq T + 2 \leq \frac{1}{2\delta} + 2$. Turning to the

¹Note that for $a \in \mathbb{R}$, the notation $\lfloor a \rfloor$ denotes the greatest integer less than or equal to a .

lower bound, we establish it by lower bounding the packing number, and then applying Lemma 5.1. An explicit packing can be constructed as follows: first set $\theta^0 = 0$, and then define $\theta^i = -\log(1 - \delta i)$ for all i such that $\theta^i \leq 1$. We can define θ^i in this way until $1/e = 1 - T\delta$, or $T \geq \lfloor \frac{1-1/e}{\delta} \rfloor$. Moreover, note that for any $i \neq j$ in the resulting set of functions, we have $\|f_{\theta^i} - f_{\theta^j}\|_\infty \geq |f_{\theta^i}(1) - f_{\theta^j}(1)| \geq \delta$, by definition of θ^i . Therefore, we conclude that $M_\infty(\delta; \mathcal{P}) \geq \lfloor \frac{1-1/e}{\delta} \rfloor + 1$, and hence that

$$N_\infty(\delta; \mathcal{P}) \geq M_\infty(2\delta; \mathcal{P}) \geq \lfloor \frac{1-1/e}{2\delta} \rfloor + 1,$$

- 1 as claimed. We have thus established the scaling $\log N(\delta; \mathcal{P}, \|\cdot\|_\infty) \asymp \log(1/\delta)$ as
 2 $\delta \rightarrow 0$, which is the usual rate that one expects for a scalar parametric class. ♣

3 A function class with a metric entropy that scales as $\log(1/\delta)$ when $\delta \rightarrow 0$ is relatively
 4 small. Indeed, as shown in Example 5.1, the interval $[-1, 1]$ has metric entropy of this
 5 order, and the function class \mathcal{P} from Example 5.5 is not essentially different. Other
 6 function classes are much richer, and so their metric entropy exhibits a correspondingly
 7 faster growth, as shown by the following example.

Example 5.6 (Lipschitz functions on the unit interval). Now consider the class of Lipschitz functions

$$\mathcal{F}_L := \{g : [0, 1] \rightarrow \mathbb{R} \mid g(0) = 0, \text{ and } |g(x) - g(x')| \leq L|x - x'| \quad \forall x, x' \in [0, 1]\}. \quad (5.13)$$

Here $L > 0$ is a fixed constant, and all of the functions in the class obey the Lipschitz bound, uniformly over all of $[0, 1]$. Note that the function class \mathcal{P} from Example 5.5 is contained within the class \mathcal{F}_L . It is known that the metric entropy of the class \mathcal{F}_L with respect to the sup-norm scales as

$$\log N_\infty(\delta; \mathcal{F}_L) \asymp L/\delta, \quad \text{for suitably small } \delta > 0. \quad (5.14)$$

- 8 Consequently, the set of Lipschitz functions is a *much* larger class than the parametric
 9 function class from Example 5.5, since its metric entropy grows as $1/\delta$ as $\delta \rightarrow 0$, as
 10 compared to $\log(1/\delta)$.

Let us prove the lower bound in equation (5.14); via Lemma 5.1, it suffices to construct a sufficiently large packing of the set \mathcal{F}_L . For a given $\epsilon > 0$, define $M = \lfloor 1/\epsilon \rfloor$, and consider the points in $[0, 1]$ given by

$$x_i = (i - 1)\epsilon, \quad \text{for } i = 1, \dots, M, \text{ and } x_{M+1} = M\epsilon < 1.$$

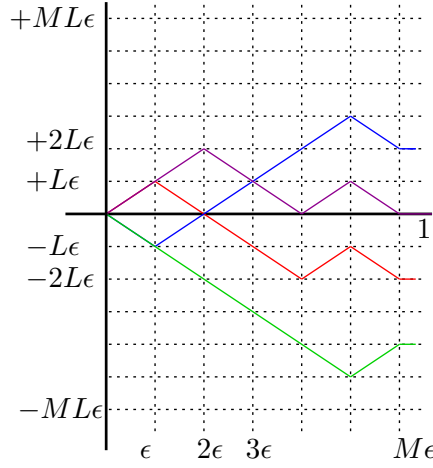


Figure 5-2. The function class $\{f_\beta, \beta \in \{-1, +1\}^M\}$ used to construct a packing of the Lipschitz class \mathcal{F}_L . Each function is piecewise linear over the intervals $[0, \epsilon]$, $[\epsilon, 2\epsilon]$, \dots , $[(M-1)\epsilon, M\epsilon]$ with slope either $+L$ or $-L$. There are 2^M functions in total, where $M = \lfloor 1/\epsilon \rfloor$.

Moreover, define the function $\phi : \mathbb{R} \rightarrow \mathbb{R}_+$ via

$$\phi(u) := \begin{cases} 0 & \text{for } u < 0 \\ u & \text{for } u \in [0, 1], \\ 1 & \text{otherwise.} \end{cases} \quad (5.15)$$

For each binary sequence $\beta \in \{-1, +1\}^M$, we may then define the function $f : [0, 1] \rightarrow [-L, +L]$ via

$$f_\beta(y) = \sum_{i=1}^M \beta_i L \epsilon \phi\left(\frac{y - x_i}{\epsilon}\right) \quad (5.16)$$

By construction, each function f_β is piecewise linear and continuous, with slope either $+L$ or $-L$ over each the intervals $[\epsilon(i-1), \epsilon i]$ for $i = 1 \dots, M$, and constant on the remaining interval $[M\epsilon, 1]$; see Figure 5-2 for an illustration. Moreover, it is straightforward to verify that $f_\beta(0) = 0$ and that f_β is Lipschitz with constant L , so that $f_\beta \in \mathcal{F}_L$.

Given a pair of distinct binary strings $\beta \neq \beta'$ and the two functions f_β and $f_{\beta'}$, there is at least one interval where the functions start at the same point, and have the opposite slope over an interval of length ϵ . Since the functions have slopes $+L$ and $-L$ respectively, we are guaranteed that $\|f_\beta - f_{\beta'}\|_\infty \geq 2L\epsilon$, showing that the set $\{f_\beta, \beta \in \{-1, +1\}^M\}$ forms a $2L\epsilon$ packing in the sup-norm. Since this set has cardinality $2^M = 2^{\lfloor 1/\epsilon \rfloor}$, after making the substitution $\epsilon = \delta/L$ and using Lemma 5.1,

we conclude that

$$\log N(\delta; \mathcal{F}_L, \|\cdot\|_\infty) \gtrsim L/\delta.$$

- 1 With a little more effort, it can also be shown that the set of functions $\{f_\beta, \beta \in \{-1, +1\}^M\}$
 2 define a suitable covering of the set \mathcal{F}_L , which establishes the overall claim (5.14). ♣

The preceding example can be extended to Lipschitz functions on the unit cube in higher dimensions, meaning real-valued functions on $[0, 1]^d$ such that

$$|f(x) - f(y)| \leq L \|x - y\|_\infty \quad \text{for all } x, y \in [0, 1]^d, \quad (5.17)$$

a class that we denote by $\mathcal{F}_L([0, 1]^d)$. An extension of our argument can then be used to show that

$$\log N_\infty(\delta; \mathcal{F}_L([0, 1]^d)) \asymp (L/\delta)^d.$$

- 3 It is worth contrasting the *exponential dependence* of this metric entropy on the di-
 4 mension d , as opposed to the linear dependence that we saw earlier for simpler sets
 5 (e.g., such as d -dimensional unit balls). This is a dramatic manifestation of the curse
 6 of dimensionality.

- 7
 8 Another direction in which Example 5.6 can be extended is to classes of functions that
 9 have higher-order derivatives.

Example 5.7 (Higher-order smoothness classes). We now consider an example of a function class based on controlling higher-order derivatives. For a suitably differentiable function f , let us adopt the notation $f^{(k)}$ to mean the k^{th} derivative. (Of course, $f^{(0)} = f$ in this notation.) For some integer α and parameter $\gamma \in (0, 1]$, consider the class of functions $f : [0, 1] \rightarrow \mathbb{R}$ such that

$$|f^{(j)}(x)| \leq C_j \quad \text{for all } x \in [0, 1], j = 0, 1, \dots, \alpha, \text{ and} \quad (5.18a)$$

$$|f^{(\alpha)}(x) - f^{(\alpha)}(y)| \leq L |x - y|^\gamma, \quad \text{for all } x, y \in [0, 1]. \quad (5.18b)$$

We claim that the metric entropy of this function class $\mathcal{F}_{\alpha, \gamma}$ scales as

$$\log N(\delta; \mathcal{F}_{\alpha, \gamma}, \|\cdot\|_\infty) \asymp (1/\delta)^{\frac{1}{\alpha+\gamma}}. \quad (5.19)$$

- 10 (Here we have absorbed the dependence on the constants C_j and L into the order
 11 notation.) Note that this claim is consistent with our calculation in Example 5.6, which
 12 is essentially the same as the class $\mathcal{F}_{0,1}$.

Let us prove the lower bound in the claim (5.19). As in the previous example, we do so by constructing a packing $\{f_\beta, \beta \in \{-1, +1\}^M\}$ for a suitably chosen integer M .

Define the function

$$\phi(y) := \begin{cases} c 2^{2(\alpha+\gamma)} y^{\alpha+\gamma} (1-y)^{\alpha+\gamma} & \text{for } y \in [0, 1], \\ 0 & \text{otherwise.} \end{cases} \quad (5.20)$$

If the pre-factor c is chosen small enough (as a function of the constants C_j and L), it can be seen that the function ϕ satisfies the conditions (5.18). Now for some $\epsilon > 0$, let us set $\delta = (\epsilon/c)^{1/(\alpha+\gamma)}$. This can be done such that $M := \lfloor 1/\delta \rfloor$ is not an integer, so that we consider the points in $[0, 1]$ given by

$$x_i = (i-1)\delta, \quad \text{for } i = 1, \dots, M, \text{ and } \quad x_{M+1} = M\delta < 1.$$

For each $\beta \in \{-1, +1\}^M$, let us define the function

$$f_\beta(x) := \sum_{i=1}^M \beta_i \delta^{1/(\alpha+\gamma)} \phi\left(\frac{x-x_i}{\delta}\right), \quad (5.21)$$

and note that it also satisfies the conditions (5.18). Finally, for two binary strings $\beta \neq \beta'$, there must exist some $i \in \{1, \dots, M\}$ and an associated interval $I_{i-1} = [x_{i-1}, x_i]$ such that

$$|f_\beta(x) - f_{\beta'}(x)| = 2^{1+2(\alpha+\gamma)} c \delta^{1/(\alpha+\gamma)} \phi\left(\frac{x-x_i}{\delta}\right) \quad \text{for all } x \in I_{i-1}.$$

By setting $x = x_i + \delta/2$, we see that

$$\|f_\beta - f_{\beta'}\|_\infty \geq 2c \delta^{\alpha+\gamma} = 2\epsilon,$$

so that the set $\{f_\beta, \beta \in \{-1, +1\}^M\}$ is a 2ϵ -packing. Thus, we conclude that

$$\log N(\delta; \mathcal{F}_{\alpha, \gamma}, \|\cdot\|_\infty) \gtrsim (1/\delta) \asymp (1/\epsilon)^{1/(\alpha+\gamma)},$$

as claimed. ♣ 1

Various types of function classes can be defined in terms of orthogonal expansions. Concretely, suppose that are given a sequence of functions $(\phi_j)_{j=1}^\infty$ belonging to $L^2[0, 1]$ and such that

$$\langle \phi_i, \phi_j \rangle = \int_0^1 \phi_i(x) \phi_j(x) dx = \begin{cases} 1 & \text{if } i = j, \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

For instance, the cosine basis is one such orthonormal basis, and there are many other interesting ones. Given such a basis, any function $f \in L^2[0, 1]$ can be expanded in the form $f = \sum_{j=1}^\infty \theta_j \phi_j$, where the expansion coefficients are given by the inner products

2Add details about cosine basis?

4

- 1 $\theta_j = \langle f, \phi_j \rangle$. By Parseval's theorem, we have $\|f\|_2^2 = \sum_{j=1}^{\infty} \theta_j^2$ so that $\|f\|_2 < \infty$ if and
 2 only $(\theta_j)_{j=1}^{\infty} \in \ell^2(\mathbb{N})$, the space of all square summable sequences. Various interesting
 3 classes of functions can be obtained by imposing additional constraints on the class of
 4 sequences, and one example is that of an ellipsoid constraint.

Example 5.8 (Function classes based on ellipsoids in $\ell^2(\mathbb{N})$). Given a sequence of non-negative real numbers $(\mu_j)_{j=1}^{\infty}$ such that $\sum_{j=1}^{\infty} \mu_j < \infty$, consider the ellipse

$$\mathcal{E} = \left\{ (\theta_j)_{j=1}^{\infty} \mid \sum_{j=1}^{\infty} \frac{\theta_j^2}{\mu_j} \leq 1 \right\} \subset \ell^2(\mathbb{N}). \quad (5.22)$$

- 5 Such ellipses play an important role in our discussion of reproducing kernel Hilbert
 6 spaces (see Chapter 12). More concretely, this example focuses on an ellipse specified
 7 by the sequence $\mu_j = j^{-2\alpha}$ for some parameter $\alpha > 1/2$. Ellipses of this type arise from
 8 certain classes of α -times differentiable functions; see Chapter 12 for details.

We claim that the metric entropy of the associated ellipse with respect to the norm $\|\cdot\|_2 = \|\cdot\|_{\ell^2(\mathbb{N})}$ scales as

$$\log N(\delta; \mathcal{E}, \|\cdot\|_2) \asymp \left(\frac{1}{\delta}\right)^\alpha \quad \text{for all suitably small } \delta. \quad (5.23)$$

Let us begin by proving the upper bound—in particular, for a given $\delta > 0$, let us upper bound $N(\sqrt{2}\delta)$, since the factor of $\sqrt{2}$ is irrelevant in establishing the claimed scaling. Let d be the smallest integer such that $\mu_d \leq \delta^2$, and consider the truncated ellipse

$$\tilde{\mathcal{E}} := \{\theta \in \mathcal{E} \mid \theta_j = 0 \text{ for all } j \geq d+1\}.$$

We claim that any δ -cover of this truncated ellipse, say $\{\theta^1, \dots, \theta^N\}$, forms a $\sqrt{2}\delta$ -cover of the full ellipse. Indeed, for any $\theta \in \mathcal{E}$, we have $\sum_{j=d+1}^{\infty} \theta_j^2 \leq \mu_d \sum_{j=d+1}^{\infty} \frac{\theta_j^2}{\mu_j} \leq \delta^2$, and hence

$$\min_{k \in [N]} \|\theta - \theta^k\|_2^2 = \min_{k \in [N]} \sum_{j=1}^d (\theta_j - \theta_j^k)^2 + \sum_{j=d+1}^{\infty} \theta_j^2 \leq 2\delta^2.$$

Consequently, it suffices to upper bound the cardinality N of this covering of $\tilde{\mathcal{E}}$. Since $\delta^2 \leq \mu_j$ for all $j \in \{1, \dots, d\}$, if we view $\tilde{\mathcal{E}}$ as a subset of \mathbb{R}^d , then it contains the ball $\mathbb{B}_2^d(\delta)$, and hence $\text{vol}(\tilde{\mathcal{E}} + \mathbb{B}_2^d(\delta/2)) \leq \text{vol}(2\tilde{\mathcal{E}})$. By standard formulae for the volume of ellipsoids, we have $\frac{\text{vol}(\tilde{\mathcal{E}})}{\text{vol}(\mathbb{B}_2^d(1))} = \prod_{j=1}^d \sqrt{\mu_j}$. Combined with the proof of Lemma 5.2,

we find that

$$\log N \leq d \log 2 + \frac{1}{2} \sum_{j=1}^d \log \mu_j - d \log(1/\delta) = d \log(2\delta) - \alpha \sum_{j=1}^d \log j,$$

using the fact that $\mu_j = j^{-2\alpha}$. Since $\sum_{j=1}^d \log j \geq d \log d - d$, we have

$$\log N \leq d(\log 2 + \alpha) + d\{\log(1/\delta) - \alpha \log d\} \leq d(\log 2 + \alpha).$$

where the final inequality follows since $\mu_d = d^{-2\alpha} \leq \delta^2$, or equivalently $\log(1/\delta) \leq \alpha \log d$.
Since $(d-1)^{-2\alpha} \geq \delta^2$, we have $d \leq (1/\delta)^{1/\alpha} + 1$, which completes the proof of the upper bound.

For the lower bound, we note that the ellipse \mathcal{E} contains the truncated ellipse $\tilde{\mathcal{E}}$, which (when viewed as a subset of \mathbb{R}^d) contains the ball $\mathbb{B}_2^d(\delta)$. Thus, we have

$$\log N\left(\frac{\delta}{2}; \mathcal{E}, \|\cdot\|_2\right) \geq \log N\left(\frac{\delta}{2}; \mathbb{B}_2^d(\delta), \|\cdot\|_2\right) \geq d \log 2,$$

where the final inequality uses the lower bound (5.12) from Example 5.4. Since $d \geq (1/\delta)^{1/\alpha}$, the lower bound follows. ♣ 5

■ 5.2 Gaussian and Rademacher complexity 6

Although metric entropy is a purely deterministic concept, it plays a fundamental role in understanding the behavior of stochastic processes. Given a collection of random variables $\{X_\theta, \theta \in \mathbb{T}\}$ indexed by \mathbb{T} , it is frequently of interest to analyze how the behavior of this stochastic process depends on the structure of the set \mathbb{T} . In the other direction, given knowledge of a stochastic process indexed by \mathbb{T} , it is often possible to infer certain properties of the set \mathbb{T} . In our treatment to follow, we will see instances of both directions of this interplay. 14

An important example of this interplay is provided by the stochastic processes that define the Gaussian and Rademacher complexities. Given a set $\mathbb{T} \subseteq \mathbb{R}^d$, the family of random variables $\{G_\theta, \theta \in \mathbb{T}\}$, where

$$G_\theta := \langle w, \theta \rangle = \sum_{i=1}^d w_i \theta_i, \quad \text{with } w_i \sim \mathcal{N}(0, 1), \text{ i.i.d.}, \quad (5.24)$$

defines a stochastic process is known as the *canonical Gaussian process* defined by \mathbb{T} .

As discussed earlier in Chapter 2, its expected supremum

$$\mathcal{G}(\mathbb{T}) := \mathbb{E} \left[\sup_{\theta \in \mathbb{T}} \langle \theta, w \rangle \right] \quad (5.25)$$

is known as the *Gaussian complexity* of \mathbb{T} . Like the metric entropy, the functional $\mathcal{G}(\mathbb{T})$ measures the size of the set \mathbb{T} in a certain sense. Replacing the standard Gaussian variables with random signs yields the *Rademacher process* $\{R_\theta, \theta \in \mathbb{T}\}$, where

$$R_\theta := \langle \varepsilon, \theta \rangle = \sum_{i=1}^d \varepsilon_i \theta_i, \quad \text{with } \varepsilon_i \text{ uniform over } \{-1, +1\}, \text{ i.i.d.} \quad (5.26)$$

- 1 Its expectation $\mathcal{R}(\mathbb{T}) := \mathbb{E} \left[\sup_{\theta \in \mathbb{T}} \langle \theta, \varepsilon \rangle \right]$ corresponds to the *Rademacher complexity* of \mathbb{T} .
- 2 As shown in Exercise 5.5, we have $\mathcal{R}(\mathbb{T}) \leq \sqrt{\frac{\pi}{2}} \mathcal{G}(\mathbb{T})$ for any set \mathbb{T} , but there are sets for
- 3 which the Gaussian complexity is substantially larger than the Rademacher complexity.
- 4

Example 5.9 (Rademacher/Gaussian complexity of Euclidean ball \mathbb{B}_2^d). Let us compute the Rademacher and Gaussian complexities of the Euclidean ball of unit norm—that is, $\mathbb{B}_2^d = \{\theta \in \mathbb{R}^d \mid \|\theta\|_2 \leq 1\}$. Computing the Rademacher complexity is straightforward: indeed, we have

$$\mathcal{R}(\mathbb{B}_2^d) = \mathbb{E} \left[\sup_{\|\theta\|_2 \leq 1} \langle \theta, \varepsilon \rangle \right] = \mathbb{E} \left[\left(\sum_{i=1}^d \varepsilon_i^2 \right)^{1/2} \right] = \sqrt{d}.$$

The same argument shows that $\mathcal{G}(\mathbb{B}_2^d) = \mathbb{E}[\|w\|_2]$ and by concavity of the square root function and Jensen's inequality, we have

$$\mathbb{E}\|w\|_2 \leq \sqrt{\mathbb{E}[\|w\|_2^2]} = \sqrt{d},$$

so that $\mathcal{G}(\mathbb{B}_2^d) \leq \sqrt{d}$. On the other hand, it can be shown that $\mathbb{E}\|w\|_2 \geq \sqrt{d}(1 - o(1))$. (This is a good exercise to work through, using concentration bounds for χ^2 variates from Chapter 2). Combining these upper and lower bounds, we conclude that

$$\mathcal{G}(\mathbb{B}_2^d)/\sqrt{d} = 1 - o(1), \quad (5.27)$$

- 5 so that the Rademacher and Gaussian complexities of \mathbb{B}_2^d are essentially equivalent. ♣


Example 5.10 (Rademacher/Gaussian complexity of \mathbb{B}_1^d). As a second example, let us consider the ℓ_1 -ball in d dimensions, denoted by \mathbb{B}_1^d . By the duality between the ℓ_1

and ℓ_∞ norms, we have

$$\mathcal{R}(\mathbb{B}_1^d) = \mathbb{E}[\sup_{\|\theta\|_1 \leq 1} \langle \theta, \varepsilon \rangle] = \mathbb{E}[\|\varepsilon\|_\infty] = 1.$$

Similarly, we have $\mathcal{G}(\mathbb{B}_1^d) = \mathbb{E}[\|w\|_\infty]$, and using the result of Exercise 2.11 on Gaussian maxima, we conclude that

$$\mathcal{G}(\mathbb{B}_1^d) / \sqrt{2 \log d} = 1 \pm o(1). \quad (5.28)$$

Thus, we see that the Rademacher and Gaussian complexities can differ by a factor of the order $\sqrt{\log d}$; this difference turns out to be the worst possible (see Exercise 5.5). But in either case, comparing with the Rademacher (or Gaussian) complexity of the Euclidean ball (5.27) shows that the ℓ_1 -ball is a much smaller set. 

Example 5.11 (Gaussian complexity of ℓ_0 -balls). We now turn to the Gaussian complexity of a set defined in a combinatorial manner. As we explore at more length in later chapters, sparsity plays an important role in many classes of high-dimensional statistical models. The ℓ_1 -norm, as discussed in Example 5.10, is a convex constraint used to enforce sparsity. A more direct and combinatorial way is by limiting the number $\|\theta\|_0 := \sum_{j=1}^d \mathbb{I}[\theta_j \neq 0]$ of non-zero entries in θ . (Despite our notation, the ℓ_0 -“norm” is not actually a norm in the usual sense of the word.) For some integer $s \in \{1, 2, \dots, d\}$, the ℓ_0 -“ball” of radius s is given by

$$\mathbb{B}_0^d(s) := \{\theta \in \mathbb{R}^d \mid \|\theta\|_0 \leq s\}. \quad (5.29)$$

This set is non-convex, corresponding to the union of $\binom{d}{s}$ subspaces, one for each of the possible s -sized subsets of d co-ordinates. Since it contains these subspaces, it is also an unbounded set, so that in computing any type of complexity measure, it is natural to impose an additional constraint. For instance, let us consider the Gaussian complexity of the set

$$\mathbb{S}^d(s) := \mathbb{B}_0^d(s) \cap \mathbb{B}_2^d(1) = \{\theta \in \mathbb{R}^d \mid \|\theta\|_0 \leq s, \text{ and } \|\theta\|_2 \leq 1\}. \quad (5.30)$$

Exercise 5.7 leads the reader through the steps required to establish the upper bound

$$\mathcal{G}(\mathbb{S}^d(s)) \lesssim \sqrt{s \log \frac{e d}{s}}, \quad (5.31)$$

where $e \approx 2.7183$ is defined as usual. Moreover, we show in Exercise 5.8 that this bound is tight up to constant factors. 

The preceding examples focused on subsets of vectors in \mathbb{R}^d . Gaussian complexity also plays an important role in measuring the size of different classes of functions. For

a given class \mathcal{F} of real-valued functions with domain \mathcal{X} , let $x_1^n := \{x_1, \dots, x_n\}$ be a collection of n points within the domain, known as the *design points*. We can then define the set

$$\mathcal{F}(x_1^n) := \left\{ (f(x_1), f(x_2), \dots, f(x_n)) \mid f \in \mathcal{F} \right\} \subseteq \mathbb{R}^n. \quad (5.32)$$

- 1 Bounding the Gaussian complexity of this subset of \mathbb{R}^n yields a measure of the com-
 2 plexity of \mathcal{F} at scale n ; this measure plays an important role in our analysis of non-
 3 parametric least squares in Chapter 13.

It is most natural to analyze a version of the set $\mathcal{F}(x_1^n)$ that is rescaled, either by $n^{-1/2}$ or n^{-1} . It is useful to observe that the Euclidean metric on the rescaled set $\frac{\mathcal{F}(x_1^n)}{\sqrt{n}}$ corresponds to the *empirical $L^2(\mathbb{P}_n)$ norm* on the function space \mathcal{F} , defined as

$$\|f - g\|_n := \sqrt{\frac{1}{n} \sum_{i=1}^n (f(x_i) - g(x_i))^2}. \quad (5.33)$$

Note that if the function class \mathcal{F} is uniformly bounded (i.e., $\|f\|_\infty \leq b$ for all $f \in \mathcal{F}$), then we also have $\|f\|_n \leq b$ for all $f \in \mathcal{F}$. In this case, we always have the following (trivial) upper bound

$$\mathcal{G}\left(\frac{\mathcal{F}(x_1^n)}{n}\right) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \frac{w_i}{\sqrt{n}} \frac{f(x_i)}{\sqrt{n}}\right] \leq b \frac{\mathbb{E}[\|w\|_2]}{\sqrt{n}} \leq b,$$

- 4 where we have recalled our analysis of $\mathbb{E}[\|w\|_2]$ from Example 5.9. Thus, a bounded
 5 function class (evaluated at n points) has Gaussian complexity that is never larger than
 6 a (scaled) Euclidean ball in \mathbb{R}^n . A more refined analysis will show that the Gaussian
 7 complexity of $\frac{\mathcal{F}(x_1^n)}{n}$ is often substantially smaller, depending on the structure of \mathcal{F} .
 8 We will study many instances of such refined bounds in the sequel.

9 ■ 5.3 Metric entropy and sub-Gaussian processes

- 10 Both the canonical Gaussian process (5.24) and the Rademacher process (5.26) are par-
 11 ticular examples of sub-Gaussian processes, which we now define in more generality.

Definition 5.3. A collection of zero-mean random variables $\{X_\theta, \theta \in \mathbb{T}\}$ is a *sub-Gaussian process* with respect to a metric ρ_X on \mathbb{T} if

$$\mathbb{E}[e^{\lambda(X_\theta - X_{\theta'})}] \leq \exp e^{\frac{\lambda^2 \rho_X^2(\theta, \theta')}{2}} \quad \text{for all } \theta, \theta' \in \mathbb{T}, \text{ and } \lambda \in \mathbb{R}. \quad (5.34)$$

By the results of Chapter 2, the bound (5.34) implies the tail bound

$$\mathbb{P}[|X_\theta - X_{\theta'}| \geq t] \leq 2e^{-\frac{t^2}{2\rho_X^2(\theta, \theta')}}.$$

and imposing such a tail bound is an equivalent way in which to define a sub-Gaussian process. It is easy to see that the canonical Gaussian and Rademacher processes are both sub-Gaussian with respect to the Euclidean metric $\|\theta - \theta'\|_2$.

Given a sub-Gaussian process, we use the notation $N_X(\delta; \mathbb{T})$ to denote the δ -covering number of \mathbb{T} with respect to ρ_X , and $N_2(\delta; \mathbb{T})$ to denote the covering number with respect to the Euclidean metric $\|\cdot\|_2$, corresponding to the case of a canonical Gaussian process. As we now discuss, these metric entropies can be used to construct upper bounds on various expected suprema involving the process.

■ 5.3.1 Upper bound by naive discretization

We begin with a simple upper bound obtained via a discretization argument. The basic idea is natural: by approximating the set \mathbb{T} up to some accuracy δ , we may replace the supremum over \mathbb{T} by a finite maximum over the δ -covering set, plus an approximation error that scales proportionally with δ . We let $D := \sup_{\theta, \theta' \in \mathbb{T}} \rho_X(\theta, \theta')$ denote the diameter of \mathbb{T} , and let $N_X(\delta; \mathbb{T})$ denote the δ -covering number of \mathbb{T} in the ρ_X -metric.

Proposition 5.1 (1-step discretization bound). Let $\{X_\theta, \theta \in \mathbb{T}\}$ be a zero-mean sub-Gaussian process with respect to the metric ρ_X . Then for any $\delta \in [0, D]$ such that $N_X(\delta; \mathbb{T}) \geq 10$, we have

$$\mathbb{E}\left[\sup_{\theta, \theta' \in \mathbb{T}} (X_\theta - X_{\theta'})\right] \leq 2\mathbb{E}\left[\sup_{\substack{\gamma, \gamma' \in \mathbb{T} \\ \rho_X(\gamma, \gamma') \leq \delta}} (X_\gamma - X_{\gamma'})\right] + 6\sqrt{D^2 \log N_X(\delta; \mathbb{T})}. \quad (5.35)$$

Remarks: It is convenient to state the upper bound in terms of the increments $X_\theta - X_{\theta'}$ so as to avoid issues of considering where the set \mathbb{T} is centered. However, the claim (5.35) always implies an upper bound on $\mathbb{E}[\sup_{\theta \in \mathbb{T}} X_\theta]$, since the zero-mean condition means that

$$\mathbb{E}[\sup_{\theta \in \mathbb{T}} X_\theta] = \mathbb{E}[\sup_{\theta \in \mathbb{T}} (X_\theta - X_{\theta_0})] \leq \mathbb{E}\left[\sup_{\theta, \theta' \in \mathbb{T}} (X_\theta - X_{\theta'})\right].$$

For each $\delta \in [0, D]$, the upper bound (5.35) consists of two quantities, corresponding to approximation error and estimation error respectively. As $\delta \rightarrow 0^+$, the approximation error (involving the constraint $\rho_X(\gamma, \gamma') \leq \delta$) shrinks to zero, whereas the estimation error (involving the metric entropy) grows. In practice, one chooses δ so as to achieve the optimal trade-off between these two terms.

Proof. For a given $\delta \geq 0$ and associated covering number $N = N_X(\delta; \mathbb{T})$, let $\{\theta^1, \dots, \theta^N\}$ be an δ -cover of \mathbb{T} . For any $\theta \in \mathbb{T}$, we can write $\theta = \theta^i + \Delta$ where $\rho_X(\Delta, 0) \leq \delta$, and hence

$$X_\theta - X_{\theta^1} = (X_{\theta^i + \Delta} - X_{\theta^i}) + (X_{\theta^i} - X_{\theta^1}) \leq \sup_{\substack{\gamma, \gamma' \in \mathbb{T} \\ \rho_X(\gamma, \gamma') \leq \delta}} (X_\gamma - X_{\gamma'}) + \max_{i=1,2,\dots,N} |X_{\theta^i} - X_{\theta^1}|$$

Given some other arbitrary $\theta' \in \mathbb{T}$, the same upper bound holds for $X_{\theta^1} - X_{\theta'}$, so that adding together the bounds, we obtain

$$\sup_{\theta, \theta' \in \mathbb{T}} (X_\theta - X_{\theta'}) \leq 2 \sup_{\substack{\gamma, \gamma' \in \mathbb{T} \\ \rho_X(\gamma, \gamma') \leq \delta}} (X_\gamma - X_{\gamma'}) + 2 \max_{i=1,2,\dots,N} |X_{\theta^i} - X_{\theta^1}| \quad (5.36)$$

Now by assumption, for each $i = 1, 2, \dots, N$, the random variable $X_{\theta^i} - X_{\theta^1}$ is zero-mean and sub-Gaussian with parameter at most $\rho_X(\theta^i, \theta^1) \leq D$. Consequently, by the behavior of sub-Gaussian maxima (see Exercise 2.11(c)), we are guaranteed that

$$\mathbb{E} \left[\max_{i=1,\dots,N} |X_{\theta^i} - X_{\theta^1}| \right] \leq 3\sqrt{D^2 \log N},$$

1 which yields the claim. □

In order to gain intuition, it is worth considering the special case of the canonical Gaussian (or Rademacher) process, in which case the relevant metric is the Euclidean norm $\|\theta - \theta'\|_2$. In order to reduce to the essential aspects of the problem, consider a set \mathbb{T} that contains the origin. The arguments² leading to the bound (5.35) imply that the Gaussian complexity $\mathcal{G}(\mathbb{T})$ is upper bounded as

$$\mathcal{G}(\mathbb{T}) \leq \min_{\delta \in [0, D]} \left\{ \mathcal{G}(\tilde{\mathbb{T}}(\delta)) + 3\sqrt{D^2 \log N_2(\delta; \mathbb{T})} \right\}, \quad (5.37)$$

where $N_2(\delta; \mathbb{T})$ is the δ -covering number in the ℓ_2 -norm, and

$$\tilde{\mathbb{T}}(\delta) := \{\gamma - \gamma' \mid \gamma, \gamma' \in \mathbb{T}, \|\gamma - \gamma'\|_2 \leq \delta\}.$$

2 The quantity $\mathcal{G}(\tilde{\mathbb{T}}(\delta))$ is referred to as a *localized Gaussian complexity*, since it measures
 3 the complexity of the set \mathbb{T} within an ℓ_2 -ball of radius δ . This idea of localization plays
 4 an important role in obtaining optimal rates for statistical problems; see Chapters 13
 5 and 14 for further discussion. We note also that analogous upper bounds hold for the
 6 Rademacher complexity $\mathcal{R}(\mathbb{T})$ in terms of a localized Rademacher complexity.
 7

In order to obtain concrete results from the discretization bound (5.37), it remains

²In this case, the argument can be refined so as to remove a factor of two.

to upper bound the localized Gaussian complexity, and then optimize the choice of δ . When \mathbb{T} is a subset of \mathbb{R}^d , the Cauchy-Schwarz inequality yields

$$\mathcal{G}(\widetilde{\mathbb{T}}(\delta)) = \mathbb{E} \left[\sup_{\theta \in \widetilde{\mathbb{T}}(\delta)} \langle \theta, w \rangle \right] \leq \delta \mathbb{E}[\|w\|_2] \leq \delta \sqrt{d},$$


which leads to the *naive discretization bound*

$$\mathcal{G}(\mathbb{T}) \leq \min_{\delta \in [0, D]} \left\{ \delta \sqrt{d} + 3\sqrt{D^2 \log N_2(\delta; \mathbb{T})} \right\}. \quad (5.38)$$

For some sets, this simple bound can yield useful results, whereas for other sets, the local Gaussian (or Rademacher) complexity needs to be controlled with more care. Let us illustrate the use of the bounds (5.35), (5.37) and (5.38) with some examples.

Example 5.12 (Gaussian complexity of unit ball). Recall our discussion of the Gaussian complexity of the the Euclidean ball \mathbb{B}_2^d from Example 5.9: using direct methods, we proved the scaling $\mathcal{G}(\mathbb{B}_2^d) = \sqrt{d}(1 - o(1))$. The purpose of this example is to show that Proposition 5.1 yields an upper bound with this type of scaling (albeit with poor control of the pre-factor). In particular, recall from Example 5.4 that the metric entropy number of the Euclidean ball is upper bounded as $\log N_2(\delta; \mathbb{B}_2^d) \leq d \log(1 + \frac{2}{\delta})$. Thus, setting $\delta = 1/2$ in the naive discretization bound (5.38), we obtain

$$\mathcal{G}(\mathbb{B}_2^d) \leq \sqrt{d} \left\{ \frac{1}{2} + 3\sqrt{2 \log 5} \right\}.$$

Relative to the exact result, the constant in this result is poor, but it does have the correct scaling as a function of d . 

Example 5.13 (Maximum singular value of sub-Gaussian random matrix). As a more substantive demonstration of Proposition 5.1, let us show how it can be used to control the expected ℓ_2 -operator norm of a sub-Gaussian random matrix. Let $\mathbf{W} \in \mathbb{R}^{n \times d}$ be a random matrix with zero-mean i.i.d. entries W_{ij} , each sub-Gaussian with parameter $\sigma = 1$. Examples include the standard Gaussian ensemble $W_{ij} \sim \mathcal{N}(0, 1)$, and the Rademacher ensemble $W_{ij} \in \{-1, +1\}$ equiprobably. The ℓ_2 -operator norm (or spectral norm) of the matrix \mathbf{W} is given corresponds to its maximum singular value; equivalently, it is defined as $\|\mathbf{W}\|_{\text{op}} := \sup_{v \in \mathbb{S}^{d-1}} \|\mathbf{W}v\|_2$, where $\mathbb{S}^{d-1} = \{v \in \mathbb{R}^d \mid \|v\|_2 = 1\}$ is the Euclidean unit sphere in \mathbb{R}^d . Here we sketch out an approach for proving the bound $\mathbb{E}[\|\mathbf{W}\|_{\text{op}}/\sqrt{n}] \lesssim 1 + \sqrt{\frac{d}{n}}$, leaving certain details for the reader in Exercise 5.11.

Let us define the class of matrices

$$\mathbb{M}^{n,d}(1) := \left\{ \Theta \in \mathbb{R}^{n \times d} \mid \text{rank}(\Theta) = 1, \|\Theta\|_F = 1 \right\}, \quad (5.39)$$

corresponding to the set of $n \times d$ matrices of rank one with unit Frobenius norm $\|\Theta\|_F^2 = \sum_{i=1}^n \sum_{j=1}^d \Theta_{ij}^2$. As verified in Exercise 5.11(a), we then have the variational representation

$$\|\mathbf{W}\|_{\text{op}} = \sup_{\Theta \in \mathbb{M}^{n,d}(1)} X_{\Theta}, \quad \text{where} \quad X_{\Theta} := \langle \mathbf{W}, \Theta \rangle = \sum_{i=1}^n \sum_{j=1}^d \mathbf{W}_{ij} \Theta_{ij}. \quad (5.40)$$

In the Gaussian case, this representation shows that $\mathbb{E}[\|\mathbf{W}\|_{\text{op}}]$ is equal to the Gaussian complexity $\mathcal{G}(\mathbb{M}^{n,d}(1))$. For any sub-Gaussian random matrix, we show in part (b) of Exercise 5.11 that the stochastic process $\{X_{\Theta}, \Theta \in \mathbb{M}^{n,d}(1)\}$ is zero-mean, and sub-Gaussian with respect to the Frobenius norm $\|\Theta - \Theta'\|_F$. Consequently, Proposition 5.1 implies that for all $\delta \in [0, 1]$, we have the upper bound

$$\mathbb{E}[\|\mathbf{W}\|_{\text{op}}] \leq 2 \mathbb{E} \left[\sup_{\substack{\text{rank}(\Gamma) = \text{rank}(\Gamma') = 1 \\ \|\Gamma - \Gamma'\|_F \leq \delta}} \langle \Gamma - \Gamma', \mathbf{W} \rangle \right] + 6 \sqrt{\log N_F(\delta; \mathbb{M}^{n,d}(1))}, \quad (5.41)$$

where $N_F(\delta; \mathbb{M}^{n,d}(1))$ denotes the δ -covering number in Frobenius norm. In part (c) of Exercise 5.11, we prove the upper bound

$$\mathbb{E} \left[\sup_{\substack{\text{rank}(\Gamma) = \text{rank}(\Gamma') = 1 \\ \|\Gamma - \Gamma'\|_F \leq \delta}} \langle \Gamma - \Gamma', \mathbf{W} \rangle \right] \leq \sqrt{2} \delta \mathbb{E}[\|\mathbf{W}\|_{\text{op}}], \quad (5.42)$$

and in part (d), we upper bound the metric entropy as

$$\log N_F(\delta; \mathbb{M}^{n,d}(1)) \leq (n + d) \log \left(1 + \frac{2}{\delta} \right). \quad (5.43)$$

Substituting these upper bounds into equation (5.41), we obtain

$$\mathbb{E}[\|\mathbf{W}\|_{\text{op}}] \leq \min_{\delta \in [0,1]} \left\{ 2\sqrt{2}\delta \mathbb{E}[\|\mathbf{W}\|_{\text{op}}] + 6 \sqrt{(n + d) \log(1 + 2/\delta)} \right\}.$$

Fixing $\delta = \frac{1}{4\sqrt{2}}$ (as one particular choice) and re-arranging terms yields the upper bound

$$\frac{1}{\sqrt{n}} \mathbb{E}[\|\mathbf{W}\|_{\text{op}}] \leq c_1 \left(1 + \sqrt{\frac{d}{n}} \right),$$

- 1 for some universal constant $c_1 > 1$. Again, this yields the correct scaling of $\mathbb{E}[\|\mathbf{W}\|_{\text{op}}]$
- 2 as a function of (n, d) . As we explore in Exercise 5.14, for Gaussian random matrices,
- 3 a more refined argument using the Sudakov-Fernique comparison can be used to prove
- 4 the upper bound with $c_1 = 1$, which is the best possible. In Example 5.21 to follow, we
- 5 establish a matching lower bound of the same order. ♣

Let us now turn to some examples of Gaussian complexity involving function spaces. Recall the definition (5.32) of the set $\mathcal{F}(x_1^n)$ as well as the empirical L^2 -norm (5.33). As a consequence of the inequalities

$$\|f - g\|_n \leq \max_{i=1, \dots, n} |f(x_i) - g(x_i)| \leq \|f - g\|_\infty,$$

we have the following relations among metric entropies,

$$\log N_2(\delta; \mathcal{F}(x_1^n)/\sqrt{n}) \leq \log N_\infty(\delta; \mathcal{F}(x_1^n)) \leq \log N(\delta; \mathcal{F}, \|\cdot\|_\infty), \quad (5.44)$$

which will be useful in our development. 1

Example 5.14 (Empirical Gaussian complexity for a parametric function class). Let us bound the Gaussian complexity of the set $\mathcal{P}(x_1^n)/n$ generated by the simple parametric function class \mathcal{P} from Example 5.5. Using the bound (5.44), it suffices to control the ℓ_∞ -covering number of \mathcal{P} . From our previous calculations, it can be seen that as long as $\delta \leq 1/4$, we have $\log N_\infty(\delta; \mathcal{P}) \leq \log(1/\delta)$. Moreover, since the function class is uniformly bounded (i.e., $\|f\|_\infty \leq 1$ for all $f \in \mathcal{P}$), the diameter in empirical L^2 -norm is also well-controlled—in particular, we have $D^2 = \sup_{f \in \mathcal{P}} \frac{1}{n} \sum_{i=1}^n f^2(x_i) \leq 1$. Consequently, the discretization bound (5.35) implies that

$$\mathcal{G}(\mathcal{P}(x_1^n)/n) \leq \frac{1}{\sqrt{n}} \inf_{\delta \in (0, 1/4]} \left\{ \delta \sqrt{n} + 3\sqrt{\log(1/\delta)} \right\}.$$

In order to optimize the scaling of the bound, we set $\delta = 1/(4\sqrt{n})$, and thereby obtain the upper bound

$$\mathcal{G}(\mathcal{P}(x_1^n)/n) \lesssim \sqrt{\frac{\log n}{n}}. \quad (5.45)$$


As we will see later, the Gaussian complexity for this function class is actually upper bounded by $1/\sqrt{n}$, so that the crude bound from Proposition 5.1 captures the correct behavior only up to a logarithmic factor. We will later develop more refined techniques that remove this logarithmic factor. ♣

Example 5.15 (Gaussian complexity for smoothness classes). Now recall the class \mathcal{F}_L of Lipschitz functions from Example 5.6. From the bounds on metric entropy given there, as long as $\delta \in (0, \delta_0)$ for a sufficiently small $\delta_0 > 0$, we have $\log N_\infty(\delta; \mathcal{F}_L) \leq \frac{cL}{\delta}$ for some constant c . Since the functions in \mathcal{F}_L are uniformly bounded by one, the discretization bound implies that

$$\mathcal{G}(\mathcal{F}_L(x_1^n)/n) \leq \frac{1}{\sqrt{n}} \inf_{\delta \in (0, \delta_0)} \left\{ \delta \sqrt{n} + 3\sqrt{\frac{cL}{\delta}} \right\}.$$

To obtain the tightest possible upper bound (up to constant factors), we set $\delta = n^{-1/3}$, and hence find that

$$\frac{1}{n} \mathcal{G}(\mathcal{F}_L(x_1^n)) \lesssim n^{-1/3}. \quad (5.46)$$

- 1 By comparison to the parametric scaling (5.45), this upper bound decays much more
2 slowly. 

3 ■ 5.4 Chaining and Dudley's entropy integral

4 In this section, we introduce an important method known as chaining, and show how it
5 can be used to obtain tighter bounds on the expected suprema of sub-Gaussian processes.
6 Recall the discretization bound from Proposition 5.1: it was based on a simple one-step
7 discretization in which we replaced the supremum over a large set with a finite maximum
8 over a δ -cover plus an approximation error. We then bounded the finite maximum by
9 combining the union bound with a Gaussian tail bound. In this section, we describe a
10 substantial refinement of this procedure, in which we decompose the supremum into a
11 sum of finite maxima over sets that are successively refined. The resulting procedure is
12 known as the *chaining method*.

In this chapter, we show how chaining can be used to derive a classical upper bound, originally due to Dudley [Dud67], on the expected supremum of a sub-Gaussian process. In Section 5.7, we show how related arguments can be used to control the probability of a deviation above this expectation. Let $\{X_\theta, \theta \in \mathbb{T}\}$ be a zero-mean sub-Gaussian process with respect to the (pseudo)-metric ρ_X (see Definition 5.3). Define $D = \sup_{\theta, \theta' \in \mathbb{T}} \rho_X(\theta, \theta')$, and the δ -truncated *Dudley's entropy integral*

$$\mathcal{J}(\delta; D) := \int_\delta^D \sqrt{\log N_X(u; \mathbb{T})} du, \quad (5.47)$$

13 where we recall that $N_X(u; \mathbb{T})$ denotes the δ -covering number of \mathbb{T} with respect to ρ_X .

14 **Theorem 5.1** (Dudley's entropy integral bound). Let $\{X_\theta, \theta \in \mathbb{T}\}$ be a zero-mean sub-Gaussian process with respect to the induced pseudometric ρ_X . Then for any $\delta \in [0, D]$, we have

$$\mathbb{E} \left[\sup_{\theta, \theta' \in \mathbb{T}} (X_\theta - X_{\theta'}) \right] \leq 2 \mathbb{E} \left[\sup_{\gamma, \gamma' \in \mathbb{T}, \rho_X(\gamma, \gamma') \leq \delta} (X_\gamma - X_{\gamma'}) \right] + 24 \mathcal{J}(\delta/2; D). \quad (5.48)$$

16
17
18 **Remarks:** There is no particular significance to the constant 24, which could be im-
19 proved with a more careful analysis. We have stated the bound in terms of the increment
20 $X_\theta - X_{\theta'}$, but it can easily be translated into an upper bound on $\mathbb{E}[\sup_{\theta \in \mathbb{T}} X_\theta]$. (See the

discussion following Proposition 5.1. The usual form of Dudley's bound corresponds to the case $\delta = 0$, and so is in terms of the entropy integral $\mathcal{J}(0; D)$. The additional flexibility to choose $\delta \in [0, D]$ can be useful in certain problems.

Proof. We begin with the inequality (5.36) previously established in the proof of Proposition 5.1—namely,

$$\sup_{\theta, \theta' \in \mathbb{T}} (X_\theta - X_{\theta'}) \leq 2 \sup_{\substack{\gamma, \gamma' \in \mathbb{T} \\ \rho_X(\gamma, \gamma') \leq \delta}} (X_\gamma - X_{\gamma'}) + 2 \max_{i=1,2,\dots,N} |X_{\theta^i} - X_{\theta^1}|.$$

In the proof of Proposition 5.1, we simply upper bounded the maximum over $i = 1, \dots, N$ using union bound. In this proof, we pursue a more refined chaining argument. Define $\mathbb{T} = \{\theta^1, \dots, \theta^N\}$, and for each integer $m = 1, 2, \dots, L$, let $\mathbb{T}_m = \{\beta^1, \beta^2, \dots, \beta^{N_m}\}$ be a minimal $\epsilon_m = D2^{-m}$ covering set of \mathbb{T} in the metric ρ_X , so that each set has $|\mathbb{T}_m| = N_m = N_X(\epsilon_m; \mathbb{T})$ elements. Since \mathbb{T} is finite, there is some finite integer³ L for which $\mathbb{T} = \cup_{m=1}^L \mathbb{T}_m$. For each $m = 1, \dots, L$, define the mapping $\pi_m : \mathbb{T} \rightarrow \mathbb{T}_m$ via

$$\pi_m(\theta) = \arg \min_{\beta \in \mathbb{T}_m} \rho_X(\theta, \beta),$$

so that $\pi_m(\theta)$ is the best approximation of $\theta \in \mathbb{T}$ from the set \mathbb{T}_m . Using this notation, we can decompose the random variable X_θ into a sum of increments in terms of an associated sequence $(\gamma^1, \dots, \gamma^L)$, where we define $\gamma^L = \theta$, and $\gamma^{m-1} := \pi_{m-1}(\gamma^m)$ recursively for $m = L, L-1, \dots, 2$. By construction, we then have the *chaining relation*

$$X_\theta - X_{\gamma^1} = \sum_{m=2}^L (X_{\gamma^m} - X_{\gamma^{m-1}}), \quad (5.49)$$

and hence $|X_\theta - X_{\gamma^1}| \leq \sum_{m=2}^L \max_{\beta \in \mathbb{T}_m} |X_\beta - X_{\pi_{m-1}(\beta)}|$. See Figure 5-3 for an illustration of this set-up.

Thus, we have decomposed the difference between X_θ and the final element X_{γ^1} in its associated chain as a sum of increments. Given any other $\tilde{\theta} \in \mathbb{T}$, we can define the chain $\{\tilde{\gamma}^1, \dots, \tilde{\gamma}^L\}$, and then derive an analogous bound for the increment $|X_{\tilde{\theta}} - X_{\tilde{\gamma}^1}|$. By appropriately adding and subtracting terms and then applying triangle inequality, we obtain

$$\begin{aligned} |X_\theta - X_{\tilde{\theta}}| &= |X_{\gamma^1} - X_{\tilde{\gamma}^1} + (X_\theta - X_{\gamma^1}) + (X_{\tilde{\gamma}^1} - X_{\tilde{\theta}})| \\ &\leq |X_{\gamma^1} - X_{\tilde{\gamma}^1}| + |X_\theta - X_{\gamma^1}| + |X_{\tilde{\theta}} - X_{\tilde{\gamma}^1}|. \end{aligned}$$

Taking maxima over $\theta, \theta' \in \mathbb{T}$ on the left-hand side and using our upper bounds on the

³In particular, for the smallest integer such that $N_L = |\mathbb{T}|$, we can simply choose $\mathbb{T}_L = \mathbb{T}$.

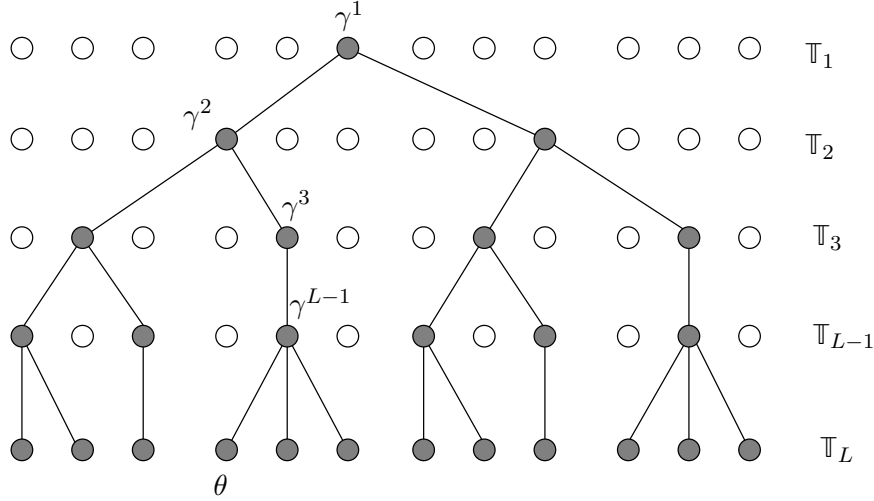


Figure 5-3. Illustration of the chaining relation in the case $L = 5$. The set \mathbb{T} , shown at the bottom of the figure, has a finite number of elements. For each $m = 1, 2, 3, \dots, 5$, we let \mathbb{T}_m be an $D\epsilon^{-m}$ cover of the set \mathbb{T} ; the elements of the cover are shaded in gray at each level. For each element $\theta \in \mathbb{T}$, we construct the chain by setting $\gamma^5 = \theta$, and then recursively $\gamma^{m-1} = \pi_{m-1}(\gamma^m)$ for $m = 5, 4, \dots, 2$. We can then decompose the difference $X_\theta - X_{\gamma^1}$ as a sum (5.49) of terms along the edges of tree.

right-hand side, we obtain

$$\max_{\theta, \theta' \in \mathbb{T}} |X_\theta - X_{\theta'}| \leq \max_{\gamma, \tilde{\gamma} \in \mathbb{T}_1} |X_\gamma - X_{\tilde{\gamma}}| + 2 \sum_{m=2}^L \max_{\beta \in \mathbb{T}_m} |X_\beta - X_{\pi_{m-1}(\beta)}|,$$

We first upper bound the finite maximum over \mathbb{T}_1 , which has $N(D/2) := N_X(D/2; \mathbb{T})$ elements. By the sub-Gaussian nature of the process, the increment $X_\gamma - X_{\tilde{\gamma}}$ is sub-Gaussian with parameter at most $\rho_X(\gamma, \tilde{\gamma}) \leq D$. Consequently, by our earlier results on finite Gaussian maxima (see Exercise (2.11)), we have

$$\mathbb{E} \left[\max_{\gamma, \tilde{\gamma} \in \mathbb{T}_1} |X_\gamma - X_{\tilde{\gamma}}| \right] \leq 3D \sqrt{\log N(D/2)}.$$

Similarly, for each $m = 2, 3, \dots, L$, the set \mathbb{T}_m has $N(D2^{-m})$ elements, and moreover, $\max_{\beta \in \mathbb{T}_m} \rho_X(\beta, \pi_{m-1}(\beta)) \leq D2^{-(m-1)}$, whence

$$\mathbb{E} \left[\max_{\beta \in \mathbb{T}_m} |X_\beta - X_{\pi_{m-1}(\beta)}| \right] \leq D2^{-(m-1)} \sqrt{2 \log N(D2^{-m})}.$$

Combining the pieces, we conclude that

$$\mathbb{E} \left[\max_{\theta, \theta' \in \mathbb{T}} |X_\theta - X_{\theta'}| \right] \leq \sum_{m=1}^L 3 D 2^{-(m-1)} \sqrt{\log N(D 2^{-m})}.$$

Since the metric entropy $\log N(t)$ is non-increasing in t , we have

$$D 2^{-(m-1)} \sqrt{2 \log(N(D 2^{-m}))} \leq 4 \int_{D 2^{-(m+1)}}^{D 2^{-m}} \sqrt{2 \log N(u)} du,$$

$$\text{and hence } \mathbb{E} \left[\sup_{\theta, \theta' \in \mathbb{T}} |X_\theta - X_{\theta'}| \right] \leq 24 \int_{\delta/2}^D \sqrt{\log N(u)} du. \quad \square \quad 1$$

Let us illustrate the Dudley entropy bound with some examples. 2

3

Example 5.16. In Example 5.14, we showed that the Gaussian complexity of the parametric function class \mathcal{P} was upper bounded by $\mathcal{O}(\sqrt{\frac{\log n}{n}})$, a result obtained by the naive discretization bound. Here we show that the Dudley entropy integral yields the sharper upper bound $\mathcal{O}(1/\sqrt{n})$. In particular, since the L_∞ -norm metric entropy is upper bounded as $\log N(\delta; \mathcal{P}, \|\cdot\|_\infty) = \mathcal{O}(\log(1 + 1/\delta))$, the Dudley bound implies that

$$\mathcal{G}\left(\frac{\mathcal{P}(x_1^n)}{n}\right) \leq \frac{c}{\sqrt{n}} \int_0^2 \sqrt{\log(1 + 1/u)} du = \frac{c'}{\sqrt{n}}.$$

Thus, we have removed the logarithmic factor from the naive discretization bound. ♣ 4

Recall from Chapter 4 our discussion of the Vapnik-Chervonenkis dimension. As we now show, the Dudley integral can be used to obtain a sharp result for any finite VC class. 5

6

7

Example 5.17 (Bounds for Vapnik-Chervonenkis classes). Let \mathcal{F} be a b -uniformly bounded class of functions with finite VC dimension ν , and suppose that we are interested in establishing a uniform law of large numbers for \mathcal{F} —that is, in controlling the random variable $\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f] \right|$, where $X_i \sim \mathbb{P}$ are i.i.d. samples. As discussed in Chapter 4, by exploiting concentration and symmetrization results, the study of this random variable can be reduced to controlling the expectation $\mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right]$, where ε_i are i.i.d. Rademacher variables (random signs), and the observations x_i are fixed for the moment. 8

9

10

11

12

13

14

15

To see how Dudley's entropy integral may be applied, define the zero-mean random variable $Z_f := \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(x_i)$, and consider the stochastic process $\{Z_f \mid f \in \mathcal{F}\}$. It

is straightforward to verify that the increment $Z_f - Z_g$ is sub-Gaussian with parameter

$$\|f - g\|_{\mathbb{P}_n}^2 := \frac{1}{n} \sum_{i=1}^n (f(x_i) - g(x_i))^2.$$

Consequently, by Dudley's entropy integral, we have

$$\mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right] \leq \frac{24}{\sqrt{n}} \int_0^{2b} \sqrt{\log N(t; \mathcal{F}, \|\cdot\|_{\mathbb{P}_n})} dt, \quad (5.50)$$

where we have used the fact that $\sup_{f, g \in \mathcal{F}} \|f - g\|_{\mathbb{P}_n} \leq 2b$. Now by known results on VC classes and metric entropy (see the bibliographic section for further details), there is a universal constant C such that

$$N(\varepsilon; \mathcal{F}, \|\cdot\|_{\mathbb{P}_n}) \leq C \nu (16e)^\nu \left(\frac{b}{\varepsilon}\right)^{2\nu}. \quad (5.51)$$

See Exercise 5.4 for the proof of a weaker claim of this form. Substituting the metric entropy bound (5.51) into the entropy integral (5.50), we find that there are universal constants c_0 and c_1 , depending on b but not on (ν, n) , such that

$$\begin{aligned} \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right] &\leq c_0 \frac{\nu}{n} + c_1 \sqrt{\frac{\nu}{n}} \int_0^{2b} \sqrt{\log(b/t)} dt \\ &= c_0 \frac{\nu}{n} + c'_1 \sqrt{\frac{\nu}{n}}. \end{aligned} \quad (5.52)$$

1



Note that the bound (5.52) is sharper earlier $\sqrt{\frac{\nu \log(n+1)}{n}}$ bound that we proved in Lemma 4.1. It leads to various improvements of previous results that we have stated. For example, if we consider the classical Glivenko-Cantelli setting, which amounts to bounding $\|\hat{F}_n - F\|_\infty = \sup_{u \in \mathbb{R}} |\hat{F}_n(u) - F(u)|$. Since the set of indicator functions has VC dimension $\nu = 1$, the bound (5.52), combined with Theorem 4.2, yields that

$$\mathbb{P} \left[\|\hat{F}_n - F\|_\infty \geq \frac{c}{\sqrt{n}} + \delta \right] \leq 2e^{-\frac{n\delta^2}{8}} \quad \text{for all } \delta \geq 0, \quad (5.53)$$

2 where c is a universal constant. Apart from better constants, this bound is unimprov-
3 able.

■ 5.5 Some Gaussian comparison inequalities

Suppose that we are given a pair of Gaussian processes, say $\{Y_\theta, \theta \in \mathbb{T}\}$ and $\{Z_\theta, \theta \in \mathbb{T}\}$, both indexed by the same set \mathbb{T} . It is often useful to compare the two processes in some sense, possibly in terms of the expected value of some real-valued function F defined on the processes. One important example is the supremum $F(X) := \sup_{\theta \in \mathbb{T}} X_\theta$. Under what conditions can we say that $F(X)$ is larger (or smaller) than $F(Y)$? Results that allow us to deduce such properties are known as *Gaussian comparison inequalities*, and there are a large number of them. In this section, we derive a few of the standard ones, and illustrate them via a number of examples.

Recall that we have defined the suprema of Gaussian processes by taking limits of maxima over finite subsets. For this reason, it is sufficient to consider the case where \mathbb{T} is finite, say $\mathbb{T} = \{1, \dots, N\}$ for some integer N . We focus on this case throughout this section, adopting the notation $[N] = \{1, \dots, N\}$ as a convenient shorthand.

■ 5.5.1 A general comparison result

We begin by stating and proving a fairly general Gaussian comparison principle:

Theorem 5.2. Let (X_1, \dots, X_N) and (Y_1, Y_2, \dots, Y_N) be a pair of centered Gaussian random vectors, and suppose that there exist disjoint subsets A and B of $[N] \times [N]$

$$\mathbb{E}[X_i X_j] \leq \mathbb{E}[Y_i Y_j] \quad \text{for all } (i, j) \in A, \quad (5.54a)$$

$$\mathbb{E}[X_i X_j] \geq \mathbb{E}[Y_i Y_j] \quad \text{for all } (i, j) \in B, \quad (5.54b)$$

$$\mathbb{E}[X_i X_j] = \mathbb{E}[Y_i Y_j] \quad \text{for all } (i, j) \notin A \cup B. \quad (5.54c)$$

Let $F : \mathbb{R}^N \rightarrow \mathbb{R}$ be a twice differentiable function, and suppose that

$$\frac{\partial^2 F}{\partial u_i \partial u_j}(u) \geq 0 \quad \text{for all } (i, j) \in A, \text{ and} \quad (5.55a)$$

$$\frac{\partial^2 F}{\partial u_i \partial u_j}(u) \leq 0 \quad \text{for all } (i, j) \in B. \quad (5.55b)$$

Then we are guaranteed that

$$\mathbb{E}[F(X)] \leq \mathbb{E}[F(Y)]. \quad (5.56)$$

Proof. We may assume without loss of generality that X and Y are independent. We

proceed via an interpolation argument: define the Gaussian random vector

$$Z(t) = (\sqrt{1-t})X + \sqrt{t}Y, \quad \text{for each } t \in [0, 1], \quad (5.57)$$

and consider the function $\phi : [0, 1] \rightarrow \mathbb{R}$ given by $\phi(t) = \mathbb{E}[F(Z(t))]$. If we can show that $\phi'(t) \geq 0$ for all $t \in (0, 1)$, then we may conclude that

$$\mathbb{E}[F(Y)] = \mathbb{E}[\phi(1)] \geq \mathbb{E}[\phi(0)] = \mathbb{E}[F(X)].$$

With this goal in mind, we begin by using chain rule to compute the first derivative

$$\phi'(t) = \sum_{j=1}^N \mathbb{E} \left[\frac{\partial F}{\partial z_j}(Z(t)) Z'_j(t) \right].$$

For any $t \in (0, 1)$, we have

$$\begin{aligned} \mathbb{E}[Z_i(t)Z'_j(t)] &= \mathbb{E} \left[(\sqrt{1-t}X_i + \sqrt{t}Y_i) \left(-\frac{1}{2\sqrt{1-t}}X_j + \frac{1}{2\sqrt{t}}Y_j \right) \right] \\ &= \frac{1}{2} \{ \mathbb{E}[Y_i Y_j] - \mathbb{E}[X_i X_j] \}. \end{aligned}$$

- 1 Consequently, for each $i = 1, \dots, N$, we can write $Z_i(t) = \alpha_{ij}Z'_j(t) + W_{ij}$, where
 2 $\alpha_{ij} \geq 0$ for $(i, j) \in A$, $\alpha_{ij} \leq 0$ for $(i, j) \in B$, and $\alpha_{ij} = 0$ if $(i, j) \notin A \cup B$. More-
 3 over, due to the Gaussian assumption, we are guaranteed that the random vector
 4 $W(j) := (W_{1j}, \dots, W_{Nj})$ is independent of $Z'_j(t)$.

Since F is twice differentiable, we may apply a first-order Taylor series to the function $\partial F / \partial z_j$ between the points $W(j)$ and $Z(t)$, thereby obtaining

$$\frac{\partial F}{\partial z_j}(Z(t)) = \frac{\partial F}{\partial z_j}(W) + \sum_{i=1}^n \frac{\partial^2 F}{\partial z_j \partial z_i}(U) \alpha_{ij} Z'_j(t),$$

where $U \in \mathbb{R}^N$ is some intermediate point between $W(j)$ and $Z(t)$. Taking expectations then yields

$$\begin{aligned} \mathbb{E} \left[\frac{\partial F}{\partial z_j}(Z(t)) Z'_j(t) \right] &= \mathbb{E} \left[\frac{\partial F}{\partial z_j}(W(j)) Z'_j(t) \right] + \sum_{i=1}^N \mathbb{E} \left[\frac{\partial^2 F}{\partial z_j \partial z_i}(U) \alpha_{ij} (Z'_j(t))^2 \right] \\ &= \sum_{i=1}^n \mathbb{E} \left[\frac{\partial^2 F}{\partial z_j \partial z_i}(U) \alpha_{ij} (Z'_j(t))^2 \right], \end{aligned}$$

- 5 where the first term vanishes since $W(j)$ and $Z'_j(t)$ are independent, and $Z'_j(t)$ is zero-
 6 mean. By our assumptions on the second derivatives of f and the previously stated

conditions on α_{ij} , we have $\frac{\partial^2 F}{\partial z_j \partial z_i}(U) \alpha_{ij} \geq 0$, so that we may conclude that $\phi'(t) \geq 0$ for all $t \in (0, 1)$, which completes the proof. \square

■ 5.5.2 Slepian and Sudakov-Fernique inequalities

An important corollary of Theorem 5.2 is *Slepian's inequality*.

Corollary 5.1 (Slepian's inequality). Let $X \in \mathbb{R}^N$ and $Y \in \mathbb{R}^N$ be zero-mean Gaussian random vectors such that

$$\mathbb{E}[X_i X_j] \geq \mathbb{E}[Y_i Y_j] \quad \text{for all } i \neq j, \text{ and} \quad (5.58a)$$

$$\mathbb{E}[X_i^2] = \mathbb{E}[Y_i^2] \quad \text{for all } i = 1, 2, \dots, N. \quad (5.58b)$$

Then we are guaranteed

$$\mathbb{E}[\max_{i=1, \dots, N} X_i] \leq \mathbb{E}[\max_{i=1, \dots, N} Y_i]. \quad (5.59)$$

Proof. In order to study the maximum, let us introduce, for each $\beta > 0$, a real-valued function on \mathbb{R}^N via $F_\beta(x) := \beta^{-1} \log \left\{ \sum_{j=1}^N \exp(\beta x_j) \right\}$. By a straightforward calculation, we find the useful relation

$$\max_{j=1, \dots, N} x_j \leq F_\beta(x) \leq \max_{j=1, \dots, N} x_j + \frac{\log N}{\beta}, \quad \text{valid for all } \beta > 0, \quad (5.60)$$

so that bounds on F_β can be used to control the maximum by taking $\beta \rightarrow +\infty$. Note that F_β is twice differentiable for each $\beta > 0$; moreover, some calculus shows that $\frac{\partial^2 F_\beta}{\partial x_i \partial x_j} \leq 0$ for all $i \neq j$. Consequently, applying Theorem 5.2 with $A = \emptyset$ and $B = \{(i, j), i \neq j\}$ yields that $\mathbb{E}[F_\beta(X)] \leq \mathbb{E}[F_\beta(Y)]$. Combining this inequality with the sandwich relation (5.60), we conclude that

$$\mathbb{E}[\max_{j=1, \dots, N} X_j] \leq \mathbb{E}[\max_{j=1, \dots, N} Y_j] + \frac{\log N}{\beta},$$

and taking the limit $\beta \rightarrow +\infty$ yields the claim. \square

Note that Theorem 5.2 and Corollary 5.1 are stated in terms of the variances and correlations of the random vector. In many cases, it is more convenient to compare two Gaussian processes in terms of their associated pseudo-metrics

$$\rho_X^2(i, j) = \mathbb{E}(X_i - X_j)^2, \quad \text{and} \quad \rho_Y^2(i, j) = \mathbb{E}(Y_i - Y_j)^2.$$

The Sudakov-Fernique comparison is stated in exactly this way.

Theorem 5.3 (Sudakov-Fernique). Given a pair of centered Gaussian vectors $\{X_j, j = 1, \dots, N\}$ and $\{Y_j, j = 1, \dots, N\}$, suppose that

$$\mathbb{E}(X_i - X_j)^2 \leq \mathbb{E}(Y_i - Y_j)^2 \quad \text{for all } (i, j) \in [N] \times [N]. \quad (5.61)$$

Then $\mathbb{E}[\max_{j=1, \dots, N} X_j] \leq \mathbb{E}[\max_{j=1, \dots, N} Y_j]$.

Remark: It is worth noting that the Sudakov-Fernique theorem also yields Slepian's inequality as a corollary. In particular, if the Slepian conditions (5.58a) hold, then it can be seen that the Sudakov-Fernique condition (5.61) also holds. The proof of Theorem 5.3 requires more effort than Slepian's inequality; see the bibliographic section for references to proofs.

5.5.3 Gaussian contraction inequality

One important consequence of the Sudakov-Fernique comparison is the Gaussian contraction inequality, which applies to functions $\phi_j : \mathbb{R} \rightarrow \mathbb{R}$ that are 1-Lipschitz, meaning that $|\phi_j(s) - \phi_j(t)| \leq |s - t|$ for all $s, t \in \mathbb{R}$, and satisfy the centering relation $\phi_j(0) = 0$. Given a vector $\theta \in \mathbb{R}^d$, we define (with a minor abuse of notation) the vector

$$\phi(\theta) := (\phi_1(\theta_1) \quad \phi_2(\theta_2) \quad \cdots \quad \phi_d(\theta_d)) \in \mathbb{R}^d$$

Lastly, given a set $\mathbb{T} \subseteq \mathbb{R}^d$, we let $\phi(\mathbb{T}) = \{\phi(\theta) \mid \theta \in \mathbb{T}\}$ denote its image under the mapping ϕ . The following result shows that the Gaussian complexity of this image is never larger than the Gaussian complexity $\mathcal{G}(\mathbb{T})$ of the original set.

Proposition 5.2 (Gaussian contraction inequality). For any set $\mathbb{T} \subseteq \mathbb{R}^d$ and any family of centered 1-Lipschitz functions $\{\phi_j, j = 1, \dots, d\}$, we have

$$\underbrace{\mathbb{E}\left[\sup_{\theta \in \mathbb{T}} \sum_{j=1}^d w_j \phi_j(\theta_j)\right]}_{\mathcal{G}(\phi(\mathbb{T}))} \leq \underbrace{\mathbb{E}\left[\sup_{\theta \in \mathbb{T}} \sum_{j=1}^d w_j \theta_j\right]}_{\mathcal{G}(\mathbb{T})} \quad (5.62)$$

We leave the proof of this claim for the reader (see Exercise 5.12). For future reference, we also note that, with an additional factor of two, an analogous result holds for the Rademacher complexity—namely

$$\mathcal{R}(\phi(\mathbb{T})) \leq 2\mathcal{R}(\mathbb{T}) \quad (5.63)$$

for any family of centered 1-Lipschitz functions. The proof of this result is more difficult than the Gaussian case; see the bibliographic section for further discussion.

Let us illustrate the use of the Gaussian contraction inequality (5.62) with some examples.

Example 5.18. Given a function class \mathcal{F} and a collection of design points x_1^n , we have previously studied the Gaussian complexity of the set $\mathcal{F}(x_1^n) \subset \mathbb{R}^n$ from equation (5.32). In various statistical problems, it is often more natural to consider the Gaussian complexity of the set

$$\mathcal{F}^2(x_1^n) := \left\{ (f^2(x_1), f^2(x_2), \dots, f^2(x_n)) \mid f \in \mathcal{F} \right\} \subseteq \mathbb{R}^n,$$

where $f^2(x) = [f(x)]^2$ are the squared function values. The contraction inequality allow us to upper bound the Gaussian complexity of this set in terms of the original set $\mathcal{F}(x_1^n)$. In particular, suppose that the function class is b -uniformly bounded, meaning that $b := \sup_{f \in \mathcal{F}} \|f\|_\infty = \sup_{f \in \mathcal{F}} \sup_{x \in \mathcal{X}} |f(x)|$ is finite. Define the function $\phi_b : \mathbb{R} \rightarrow \mathbb{R}$ via

$$\phi_b(t) := \begin{cases} \frac{t^2}{2b} & \text{if } |t| \leq b, \text{ and} \\ b/2 & \text{otherwise.} \end{cases}$$

It is straightforward to verify that ϕ_b is a contraction according to our definition; moreover, since $|f(x_i)| \leq b$, we have $\phi_b(f(x_i)) = \frac{f^2(x_i)}{2b}$ for all $f \in \mathcal{F}$ and $i = 1, 2, \dots, n$, and hence

$$\begin{aligned} \frac{1}{2b} \mathcal{G}(\mathcal{F}^2(x_1^n)) &= \mathbb{E} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n w_i \frac{f^2(x_i)}{2b} \right] \\ &= \mathbb{E} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n w_i \phi_b(f(x_i)) \right] \leq \mathbb{E} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n w_i f(x_i) \right] = \mathcal{G}(\mathcal{F}(x_1^n)), \end{aligned}$$

where the inequality follows from Proposition 5.2. ♣

■ 5.6 Sudakov's lower bound

In previous sections, we have derived two upper bounds on the expected supremum of a sub-Gaussian process indexed by a set \mathbb{T} : the simple 1-step discretization in Proposition 5.1, and the more refined Dudley integral bound in Theorem 5.1. In this section, we turn to the complementary question of deriving lower bounds. In contrast to the upper bounds in the preceding sections, these lower bounds are specialized to the case of *Gaussian* processes, since a general sub-Gaussian process might have different behavior than its Gaussian analog. (For instance, compare the Rademacher and Gaussian complexity of the ℓ_1 -ball, as discussed in Example 5.10.)

This section is devoted to the exploration of a lower bound known as the *Sudakov minoration*, which is obtained by exploiting the Gaussian comparison inequalities discussed in the previous section.

Theorem 5.4 (Sudakov minoration). Let $\{X_\theta, \theta \in \mathbb{T}\}$ be a zero-mean Gaussian process defined on the non-empty set \mathbb{T} . Then

$$\mathbb{E}[\sup_{\theta \in \mathbb{T}} X_\theta] \geq \sup_{\delta > 0} \frac{\delta}{2} \sqrt{\log M_X(\delta; \mathbb{T})}, \quad (5.64)$$

where $M_X(\delta; \mathbb{T})$ is the δ -packing number of the set \mathbb{T} in the metric $\rho_X(\theta, \theta') := \sqrt{\text{var}(X_\theta - X_{\theta'})}$.

Proof. For any $\delta > 0$, let $\{\theta^1, \dots, \theta^M\}$ be an δ -packing of \mathbb{T} . Any such packing yields the lower bound $\mathbb{E}[\sup_{\theta \in \mathbb{T}} X_\theta] \geq \mathbb{E}[\max_{i=1, \dots, M} Y_i]$, where $Y_i := X_{\theta^i}$. Note that by construction, we have $\mathbb{E}[(Y_i - Y_j)^2] = \rho_X^2(\theta^i, \theta^j) > \delta^2$ for all $i \neq j$. Now let us define an i.i.d. sequence of Gaussian random variables $Z_i \sim \mathcal{N}(0, \delta^2/2)$ for $i = 1, \dots, M$. Since $\mathbb{E}[(Z_i - Z_j)^2] = \delta^2$ for all $i \neq j$, the pair of random vectors Y and Z satisfy the Sudakov-Fernique condition (5.61), so that we may conclude that

$$\mathbb{E}[\sup_{\theta \in \mathbb{T}} X_\theta] \geq \mathbb{E}[\max_{i=1, \dots, M} Y_i] \geq \mathbb{E}[\max_{i=1, \dots, M} Z_i].$$

Since the variables $\{Z_i\}_{i=1}^M$ are zero-mean Gaussian and i.i.d., we can apply standard results on i.i.d. Gaussian maxima (viz. Exercise 2.11) to obtain $\mathbb{E}[\max_{i=1, \dots, M} Z_i] \geq \frac{\delta}{2} \sqrt{\log M}$, thereby completing the proof. \square

Let us illustrate the Sudakov lower bound with some examples.

Example 5.19 (Gaussian complexity of ℓ_2 -ball). We have shown previously that the Gaussian complexity $\mathcal{G}(\mathbb{B}_2^d)$ of the d -dimensional Euclidean ball satisfies $\mathcal{G}(\mathbb{B}_2^d) \leq \sqrt{d}$. We have verified this fact both by direct calculation, and through use of the upper bound in Proposition 5.1. Here let us show how the Sudakov minoration captures the complementary lower bound. From Example 5.12, the metric entropy of the ball \mathbb{B}_2^d in ℓ_2 -norm satisfies the lower bound $\log N_2(\delta; \mathbb{B}_2^d) \geq d \log(1/\delta)$. By Lemma 5.1, the packing number has the same scaling up to a factor of two. Therefore, the Sudakov bound (5.64) implies that

$$\mathcal{G}(\mathbb{B}_2^d) \geq \sup_{\delta > 0} \left\{ \frac{\sqrt{d} \delta \sqrt{\log(1/2\delta)}}{\sqrt{2}} \right\} \geq \sqrt{\frac{\log 2}{32}} \sqrt{d},$$

where we have set $\delta = 1/4$. Thus, in this simple case, the Sudakov lower bound recovers the correct scaling as a function of \sqrt{d} , albeit with poor control of the constant. ♣

We can also use the Sudakov minoration to upper bound the metric entropy of a set \mathbb{T} , assuming that we have an upper bound on its Gaussian complexity, as illustrated in the following example.

Example 5.20 (Metric entropy of ℓ_1 -ball). Let us use the Sudakov minoration to upper bound the metric entropy of the ℓ_1 -ball $\mathbb{B}_1^d = \{\theta \in \mathbb{R}^d \mid \sum_{i=1}^d |\theta_i| \leq 1\}$. We first observe that (for $d \geq 10$) its Gaussian complexity can be upper bounded as

$$\mathcal{G}(\mathbb{B}_1) = \mathbb{E} \left[\sup_{\|\theta\|_1 \leq 1} \langle w, \theta \rangle \right] = \mathbb{E} \|w\|_\infty \leq 3\sqrt{\log d},$$

where we have used the duality between the ℓ_1 and ℓ_∞ norms, and standard results on Gaussian maxima (see Exercise 2.11). Applying Sudakov's minoration, we conclude that the metric entropy of the d -dimensional ball \mathbb{B}_1^d in the ℓ_2 -norm is upper bounded as

$$\log N(\delta; \mathbb{B}_1^d, \|\cdot\|_2) \leq c(1/\delta)^2 \log d. \quad (5.65)$$

It is known that (for the most relevant range of δ) this upper bound on the metric entropy of \mathbb{B}_1^d is tight up to constant factors; see the bibliographic section for further discussion. We thus see in a different way how the ℓ_1 -ball is *much* smaller than the ℓ_2 -ball, since its metric entropy scales logarithmically in dimension, as opposed to linearly. ♣

As another example, let us now return to some analysis of the singular values of Gaussian random matrices.

Example 5.21 (Lower bounds on maximum singular value). As a continuation of Example 5.13, let us use the Sudakov minoration to *lower* bound the maximum singular value of a standard Gaussian random matrix $\mathbf{W} \in \mathbb{R}^{n \times d}$. Recall that we can write

$$\mathbb{E}[\|\mathbf{W}\|_{\text{op}}] = \mathbb{E} \left[\sup_{\Theta \in \mathbb{M}^{n,d}(1)} \langle \mathbf{W}, \Theta \rangle \right],$$

where the set $\mathbb{M}^{n,d}(1)$ was previously defined (5.39). Consequently, in order to lower bound $\mathbb{E}[\|\mathbf{W}\|_{\text{op}}]$ via Sudakov minoration, it suffices to lower bound the metric entropy of $\mathbb{M}^{n,d}(1)$ in the Frobenius norm. In Exercise 5.13, we show that there is a universal constant c_1 such that

$$\log M(\delta; \mathbb{M}^{n,d}(1); \|\cdot\|_F) \geq c_1(n+d) \log(1/\delta)$$

for all $\delta \in (0, 1/2)$. By concavity of the square root, we have $\sqrt{n+d} \geq \frac{1}{\sqrt{2}}(\sqrt{n} + \sqrt{d})$, so that the Sudakov minoration implies that

$$\mathbb{E}[\|\mathbf{W}\|_{\text{op}}] \geq c_2(\sqrt{n} + \sqrt{d}),$$

1 for some constant $c_2 > 0$. ♣

2 ■ 5.7 Chaining and Orlicz processes

3 In Section 5.4, we introduced the idea of chaining, and showed how it can be used to
 4 obtain upper bounds on the expected supremum of a sub-Gaussian process. When the
 5 process is actually Gaussian, then classical concentration results can be used to show
 6 that the supremum is sharply concentrated around this expectation (see Exercise 5.10).
 7 For more general sub-Gaussian processes, it is useful to be able to derive similar results
 8 that bound the probability of deviations above the tail. Moreover, there are many
 9 processes that do not have sub-Gaussian tails, but rather instead are sub-exponential
 10 in nature. It would also be useful to obtain bounds on the expected supremum and
 11 associated deviation bounds for such processes.

12 The notion of an *Orlicz norm* allows us to treat both sub-Gaussian and sub-
 13 exponential processes in a unified manner. For a given parameter $q \in [1, 2]$, consider
 14 the function $\psi_q(t) := \exp(t^q) - 1$. This function can be used to define a norm on the
 15 space of random variables as follows:

16 **Definition 5.4** (Orlicz norm). The ψ_q -Orlicz norm of a zero random variable X is given by

$$17 \quad \|X\|_{\psi_q} := \inf \{ \lambda > 0 \mid \mathbb{E}[\psi_q(|X|/\lambda)] \leq 1 \}. \quad (5.66)$$

18 The Orlicz norm is infinite if there is no $\lambda \in \mathbb{R}$ for which the given expectation is
 19 finite.

Any random variable with a bounded Orlicz norm satisfies a concentration inequality specified in terms of the function ψ_q . In particular, we have

$$\mathbb{P}[|X| \geq t] \stackrel{(i)}{=} \mathbb{P}\left[\psi_q(|X|/\|X\|_{\psi_q}) \geq \psi_q(t/\|X\|_{\psi_q})\right] \stackrel{(ii)}{\leq} \frac{1}{\psi_q(t/\|X\|_{\psi_q})},$$

20 where the equality (i) follows because ψ_q is an increasing function, and the bound (ii)
 21 follows from Markov's inequality. In the case $q = 2$, this bound is essentially equivalent
 22 to our usual sub-Gaussian tail bound; see Exercise 2.18 for further details.

23

Based on the notion of the Orlicz norm, we can now define an interesting generalization of a sub-Gaussian process:

Definition 5.5. A zero-mean stochastic process $\{X_\theta, \theta \in \mathbb{T}\}$ is a ψ_q -process with respect to a metric ρ if

$$\|X_\theta - X_{\theta'}\|_{\psi_q} \leq \rho(\theta, \theta') \quad \text{for all } \theta, \theta' \in \mathbb{T}. \quad (5.67)$$

As a particular example, in this new terminology, it can be verified that the canonical Gaussian process is a ψ_2 -process with respect to the (scaled) Euclidean metric $\rho(\theta, \theta') = 2\|\theta - \theta'\|_2$. We define the generalized Dudley entropy integral

$$\mathcal{J}_q(D) := \int_0^D \psi_q^{-1}(N(u; \mathbb{T}, \rho)) du. \quad (5.68)$$

where ψ_q^{-1} is the inverse function of ψ_q , and $D = \sup_{\theta, \theta' \in \mathbb{T}} \rho(\theta, \theta')$ is the diameter of the set \mathbb{T} under ρ . For the exponential-type functions considered here, note that we have

$$\psi_q^{-1}(u) = [\log(1 + u)]^{1/q}. \quad (5.69)$$

With this set-up, we have the following result:

Theorem 5.5. Let $\{X_\theta, \theta \in \mathbb{T}\}$ be an ψ_q -process with respect to ρ . Then there is a universal constant c_1 such that

$$\mathbb{P} \left[\sup_{\theta, \theta' \in \mathbb{T}} |X_\theta - X_{\theta'}| \geq c_1 (\mathcal{J}_q(D) + \delta) \right] \leq 2e^{-\frac{\delta^q}{D^q}} \quad \text{for all } \delta > 0. \quad (5.70)$$

We now turn to the proof of Theorem 5.5. We begin by stating an auxiliary lemma that is of independent interest. For any measurable set A and random variable Y , let us introduce the shorthand notation $\mathbb{E}_A[Y] = \int_A Y d\mathbb{P}$. Note that we have $\mathbb{E}_A[Y]/\mathbb{P}(A) = \mathbb{E}[Y \mid Y \in A]$ by construction.

Lemma 5.3. Suppose that Y_1, \dots, Y_N are non-negative random variables such that $\|Y_i\|_{\psi_q} \leq 1$. Then for any measurable set A , we have

$$\mathbb{E}_A[Y_i] \leq \mathbb{P}[A] \psi_q^{-1}(1/\mathbb{P}(A)) \quad \text{for all } i = 1, 2, \dots, N, \quad (5.71)$$

and moreover

$$\mathbb{E}_A \left[\max_{i=1, \dots, N} Y_i \right] \leq \mathbb{P}[A] \psi_q^{-1} \left(\frac{N}{\mathbb{P}(A)} \right). \quad (5.72)$$

Proof. Let us first establish the inequality (5.71). By definition, we have

$$\begin{aligned}\mathbb{E}_A[Y] &= \mathbb{P}[A] \frac{1}{\mathbb{P}[A]} \mathbb{E}_A[\psi_q^{-1}(\psi_q(Y))] \\ &\stackrel{(i)}{\leq} \mathbb{P}[A] \psi_q^{-1}\left(\mathbb{E}_A[\psi_q(Y)] \frac{1}{\mathbb{P}[A]}\right) \\ &\stackrel{(ii)}{\leq} \mathbb{P}[A] \psi_q^{-1}\left(\frac{1}{\mathbb{P}[A]}\right),\end{aligned}$$

- 1 where step (i) uses concavity of ψ_q^{-1} and Jensen's inequality, whereas step (ii) uses the
2 fact that $\mathbb{E}_A[\psi_q(Y)] \leq \mathbb{E}[\psi_q(Y)] \leq 1$, which follows since $\psi_q(Y)$ is non-negative, and
3 the Orlicz norm of Y is at most one, combined with the fact that ψ_q^{-1} is an increasing
4 function.

We now prove its extension (5.72). Any measurable set A can be partitioned into a disjoint union of sets A_i , $i = 1, 2, \dots, N$ such that $Y_i = \max_{j=1, \dots, N} Y_j$ on A_i . Using this partition, we have

$$\begin{aligned}\mathbb{E}_A\left[\max_{i=1, \dots, N} Y_i\right] &= \sum_{i=1}^N \mathbb{E}_{A_i}[Y_i] \leq \mathbb{P}[A] \sum_{i=1}^N \frac{\mathbb{P}[A_i]}{\mathbb{P}[A]} \psi_q^{-1}\left(\frac{1}{\mathbb{P}[A_i]}\right) \\ &\leq \mathbb{P}[A] \psi_q^{-1}\left(\frac{N}{\mathbb{P}[A]}\right),\end{aligned}$$

- 5 where the last step uses the concavity of ψ_q^{-1} , and Jensen's inequality with the weights
6 $\mathbb{P}[A_i]/\mathbb{P}[A]$.
7 □

To see the relevance of this lemma for Theorem 5.5, we will use it to show that the supremum $Z := \sup_{\theta, \theta' \in \mathbb{T}} |X_\theta - X_{\theta'}|$ satisfies the inequality

$$\mathbb{E}_A[Z] \leq 8\mathbb{P}[A] \int_0^D \psi_q^{-1}\left(\frac{N(u; \mathbb{T}, \rho)}{\mathbb{P}[A]}\right) du. \quad (5.73)$$

Choosing A to be the full probability space immediately yields an upper bound on the expected supremum—namely $\mathbb{E}[Z] \leq 8\mathcal{J}_q(D)$. On the other hand, if we choose $A = \{Z > t\}$, then we have

$$\mathbb{P}[Z > t] \stackrel{(i)}{\leq} \frac{1}{t} \mathbb{E}_A[Z] \stackrel{(ii)}{\leq} 8 \frac{\mathbb{P}[Z > t]}{t} \int_0^D \psi_q^{-1}\left(\frac{N(u; \mathbb{T}, \rho)}{\mathbb{P}[Z > t]}\right) du,$$

where step (i) follows from a version of Markov's inequality, and step (ii) follows from the bound (5.73). Cancelling out a factor of $\mathbb{P}[Z > t]$ from both sides, and using the

inequality $\psi_q^{-1}(st) \leq c(\psi_q^{-1}(s) + \psi_q^{-1}(t))$, we obtain

$$t \leq 8c \left\{ \int_0^D \psi_q^{-1}(N(u; \mathbb{T}, \rho)) du + D \psi_q^{-1}\left(\frac{1}{\mathbb{P}[Z > t]}\right) \right\}.$$

Let $\delta > 0$ be arbitrary, and set $t = 8c(\mathcal{J}_q(D) + \delta)$. Some algebra then yields the inequality $\delta \leq D \psi_q^{-1}\left(\frac{1}{\mathbb{P}[Z > t]}\right)$, or equivalently,

$$\mathbb{P}[Z \geq 8c(\mathcal{J}_q(D) + \delta)] \leq \frac{1}{\psi_q(\delta/D)},$$

as claimed. 1

In order to prove Theorem 5.5, it suffices to establish the bound (5.73). We do so by combining Lemma 5.3 with a chaining argument. For each integer $m = 0, 1, 2, \dots, L$, let \mathbb{T}_m be a minimal $\epsilon_m = D2^{-m}$ covering set of \mathbb{T} in the metric ρ_X , so that at the m^{th} step, the set \mathbb{T}_m has $N_m = N_X(\epsilon_m; \mathbb{T})$ elements. By definition of the diameter D , note that S_0 consists of a singleton, say $S_0 = \{\theta_0\}$. Since \mathbb{T} is finite, there is some finite integer L for which $\mathbb{T} = \cup_{m=1}^L \mathbb{T}_m$. For each $m = 1, \dots, L$, define the mapping $\pi_m : \mathbb{T} \rightarrow \mathbb{T}_m$ via

$$\pi_m(\theta) = \arg \min_{\gamma \in \mathbb{T}_m} \rho_X(\theta, \gamma),$$

so that $\pi_m(\theta)$ is the best approximation of $\theta \in \mathbb{T}$ from the set \mathbb{T}_m . Using this notation, we can decompose the difference $X_\theta - X_{\theta_0}$ into a sum of increments in terms of an associated sequence $(\theta_0, \dots, \theta_L)$, where $\theta_L = \theta$, and recursively for $m = L, L-1, \dots, 1$, we define $\theta_{m-1} := \pi_{m-1}(\theta_m)$. This leads to the chaining relation $X_\theta - X_{\theta_0} = \sum_{m=1}^L (X_{\theta_m} - X_{\theta_{m-1}})$, and thus the upper bound

$$\sup_{\theta \in \mathbb{T}} |X_\theta - X_{\theta_0}| \leq \sum_{m=1}^L \sup_{\gamma \in \mathbb{T}_m} |X_\gamma - X_{\pi_{m-1}(\gamma)}|. \quad (5.74)$$

For each $\gamma \in \mathbb{T}_m$, we are guaranteed that

$$\|X_\gamma - X_{\pi_{m-1}(\gamma)}\|_{\psi_q} \leq \rho_X(\gamma, \pi_{m-1}(\gamma)) \leq D 2^{-m+1}.$$

Since $|\mathbb{T}_m| = N(D2^{-m})$, Lemma 5.3 implies that

$$\mathbb{E}_A \left[\sup_{\gamma \in \mathbb{T}_m} |X_\gamma - X_{\pi_{m-1}(\gamma)}| \right] \leq \mathbb{P}[A] D 2^{-m+1} \psi_q^{-1}\left(\frac{N(D2^{-m})}{\mathbb{P}(A)}\right),$$

for every measurable set A . Consequently, from the upper bound (5.74), we obtain

$$\begin{aligned} \mathbb{E}_A \left[\sup_{\theta \in \mathbb{T}} |X_\theta - X_{\theta_0}| \right] &\leq \sum_{m=1}^L \mathbb{E}_A \left[\sup_{\gamma \in \mathbb{T}_m} |X_\gamma - X_{\pi_{m-1}(\gamma)}| \right] \\ &\leq \mathbb{P}[A] \sum_{m=1}^L D 2^{-m+1} \psi_q^{-1} \left(\frac{N(D2^{-m})}{\mathbb{P}(A)} \right) \\ &\leq \mathbb{P}[A] \int_0^D \psi_q^{-1} \left(\frac{N_X(u; \mathbb{T})}{\mathbb{P}(A)} \right) du, \end{aligned}$$

1 since the sum can be upper bounded by the integral.

2 ■ 5.8 Bibliographic details and background

3 The notion of metric entropy was introduced by Kolmogorov [Kol56, Kol58] and further
 4 developed by various authors; see the paper by Kolmogorov and Tikhomirov [KT59]
 5 for an overview and some discussion of the early history. Metric entropy, along with
 6 related notions of the “sizes” of various function classes, are central objects of study in
 7 the field of approximation theory; see the books [DL93, Pin85, CS90] for further details
 8 on approximation and operator theory. Examples 5.6 and 5.7 are discussed in depth
 9 by Kolmogorov and Tikhomirov [KT59], as is the metric entropy bound for the special
 10 ellipse given in Example 5.8. Mitjagin [Mit61] proves a more general result, giving a
 11 sharp characterization of the metric entropy for any ellipse; see also Lorentz [Lor66] for
 12 related and more general results.

13 The pioneering work of Dudley [Dud67] established the connection between the
 14 entropy integral and the behavior of Gaussian processes. The idea of chaining itself
 15 dates back to Kolmogorov and others. Upper bounds based on entropy integrals are
 16 not always the best possible. Sharp upper and lower bounds for Gaussian suprema
 17 provided by the generic chaining method of Talagrand [Tal00]. The proof of the Orlicz-
 18 norm generalization of Dudley’s entropy integral in Theorem 5.5 is based on Ledoux
 19 and Talagrand [LT91].

20 The metric entropy of the ℓ_1 -ball was discussed in Example 5.20; more generally,
 21 sharp upper and lower bounds on the entropy numbers of ℓ_q -balls for $q \in (0, 1]$ are
 22 obtained by Schütt [Sch84] and Kühn [Kö1]. Raskutti et al. [RWY11] convert these
 23 estimates to upper and lower bounds on the metric entropy; see Lemma 2 in their
 24 paper.

25 Gaussian comparison inequalities have a lengthy history in probability theory. Ledoux
 26 and Talagrand [LT91] provide a detailed discussion of Gaussian comparison inequalities,
 27 including Slepian’s inequality, the Sudakov-Fernique inequality, and Gordon’s inequal-
 28 ities. The proofs of Theorem 5.2 and Theorem 5.5 follow this development. Chatter-

jee [Cha05] provides a self-contained proof of the Sudakov-Fernique inequality, including control on the slack in the bound. See §4.2 of Ledoux and Talagrand [LT91] for a proof of the contraction inequality (5.63) for the Rademacher complexity.

The bound (5.51) on the metric entropy of a VC class is proved in Theorem 2.6.7 of van der Vaart and Wellner [vdVW96]. Exercise 5.4, adapted from this same book, works through the proof of a weaker bound.

■ 5.9 Exercises

Exercise 5.1 (Failure of total boundedness). Let $\mathcal{C}([0, 1], b)$ denote the class of all convex functions f defined on the unit interval such that $\|f\|_\infty \leq b$. Show that $\mathcal{C}([0, 1], b)$ is *not* totally bounded in the sup-norm. (*Hint*: Try to construct an infinite collection of functions $\{f^j\}_{j=1}^\infty$ such that $\|f^j - f^k\|_\infty \geq 1/2$ for all $j \neq k$.)

Exercise 5.2 (Packing and covering). Prove the following relationships between packing and covering numbers:

$$M(2\delta; \mathbb{T}, \rho) \stackrel{(a)}{\leq} N(\delta; \mathbb{T}, \rho) \stackrel{(b)}{\leq} M(\delta; \mathbb{T}, \rho).$$

Exercise 5.3 (Packing of Boolean hypercube). Recall from Example 5.2 the binary hypercube $\mathbb{H}^d = \{0, 1\}^d$ equipped with the rescaled Hamming metric (5.2). Prove that the packing number satisfies the bound

$$\frac{\log M(\delta; \mathbb{H}^d)}{d} \leq D(\delta/2 \| 1/2) + \frac{\log(d+1)}{d},$$

where $D(\delta/2 \| 1/2) = \frac{\delta}{2} \log \frac{\delta/2}{1/2} + (1 - \frac{\delta}{2}) \log \frac{1-\delta/2}{1/2}$ is the Kullback-Leibler divergence between the Bernoulli distributions with parameter $\delta/2$ and $1/2$. (*Hint*: You may find the result of Exercise 2.10 to be useful.)

Exercise 5.4. In this exercise, we explore the connection between VC dimension and metric entropy. Given a set class \mathbb{S} with finite VC dimension ν , we show that the function class $\mathcal{F}_\mathbb{S} := \{\mathbb{I}_S, S \in \mathbb{S}\}$ of indicator functions has metric entropy at most

$$N(\delta; \mathcal{F}_\mathbb{S}, L^1(\mathbb{P})) \leq K(\nu) \left(\frac{3}{\delta}\right)^{2\nu}, \quad \text{for a constant } K(\nu). \quad (5.75)$$

Let $\{\mathbb{I}_{S^1}, \dots, \mathbb{I}_{S^N}\}$ be a maximal δ -packing in the $L^1(\mathbb{P})$ norm, so that

$$\|\mathbb{I}_{S_i} - \mathbb{I}_{S_j}\|_1 = \mathbb{E}[|\mathbb{I}_{S_i}(X) - \mathbb{I}_{S_j}(X)|] > \delta \quad \text{for all } i \neq j.$$

By Exercise 5.2, this N is an upper bound on the δ -covering number.

- 1 (a) Suppose that we generate n samples X_i , $i = 1, \dots, n$ drawn i.i.d. from \mathbb{P} . Show
 2 that the probability that every set S_i picks out a different subset of $\{X_1, \dots, X_n\}$
 3 is at least $1 - \binom{N}{2}(1 - \delta)^n$.
- 4 (b) Using part (a), show that for $N \geq 2$ and $n = \frac{3 \log N}{\delta}$, there exists a set of n points
 5 from which \mathbb{S} picks out at least N subsets, and conclude that $N \leq \left(\frac{3 \log N}{\delta}\right)^\nu$.
- 6 (c) Use part (b) to show that the bound (5.75) holds with $K(\nu) := (2\nu)^{2\nu-1}$.

7 **Exercise 5.5** (Gaussian and Rademacher complexity). In this problem, we explore the
 8 connection between the Gaussian and Rademacher complexity of a set.

- 9 (a) Show that for any set $\mathbb{T} \subseteq \mathbb{R}^d$, the Rademacher complexity satisfies the upper
 10 bound $\mathcal{R}(\mathbb{T}) \leq \sqrt{\frac{\pi}{2}} \mathcal{G}(\mathbb{T})$. Give an example of a set for which this bound is met
 11 with equality.
- 12 (b) Show that for any set $\mathbb{T} \subseteq \mathbb{R}^d$ with $d \geq 10$, we have $\mathcal{G}(\mathbb{T}) \leq 3\sqrt{\log d} \mathcal{R}(\mathbb{T})$. Give
 13 an example of a set \mathbb{T} for which this upper bound is tight up to the constant pre-
 14 factor. (*Hint:* In proving this bound, you may assume the Rademacher analogue
 15 of the contraction inequality, namely that $\mathcal{R}(\phi(\mathbb{T})) \leq \mathcal{R}(\mathbb{T})$ for any contraction.)

16 **Exercise 5.6** (Gaussian complexity for ℓ_q -balls). The ℓ_q ball of unit radius is given
 17 by $\mathbb{B}_q^d(1) = \{\theta \in \mathbb{R}^d \mid \|\theta\|_q \leq 1\}$, where $\|\theta\|_q = (\sum_{j=1}^d |\theta_j|^q)^{1/q}$ for $q \in [1, \infty)$ and
 18 $\|\theta\|_\infty = \max_j |\theta_j|$.

- (a) For $q \in (1, \infty)$, show that there are constants c_q such that

$$\sqrt{\frac{2}{\pi}} \leq \frac{\mathcal{G}(\mathbb{B}_q^d(1))}{d^{1-\frac{1}{q}}} \leq c_q.$$

- 19 (b) Give an exact expression for $\mathcal{G}(\mathbb{B}_\infty^d(1))$.

Exercise 5.7 (Upper bounds for ℓ_0 -“balls”). Consider the set

$$\mathbb{T}^d(s) := \{\theta \in \mathbb{R}^d \mid \|\theta\|_0 \leq s, \|\theta\|_2 \leq 1\},$$

corresponding to all s -sparse vectors contained within the Euclidean unit ball. In this exercise, we prove that its Gaussian complexity satisfies the upper bound

$$\mathcal{G}(\mathbb{T}^d(s)) \lesssim \sqrt{s \log\left(\frac{e d}{s}\right)}. \quad (5.76)$$

- (a) Show that $\mathcal{G}(\mathbb{T}^d(s)) = \mathbb{E}[\max_{|S|=s} \|w_S\|_2]$ where $w_S \in \mathbb{R}^{|S|}$ is the sub-vector of (w_1, \dots, w_d) indexed by the subset $S \subset \{1, 2, \dots, d\}$.

- (b) Show that for any fixed subset S of cardinality s ,

$$\mathbb{P}[\|w_S\|_2 \geq \sqrt{s} + \delta] \leq e^{-\delta^2/2}.$$

- (c) Use part (b) to establish the bound (5.76).

Exercise 5.8 (Lower bounds for ℓ_0 -“balls”). In Exercise 5.7, we established an upper bound on the Gaussian complexity of the set $\mathbb{T}^d(s) := \{\theta \in \mathbb{R}^d \mid \|\theta\|_0 \leq s, \|\theta\|_2 \leq 1\}$. The point of this exercise is to establish the matching lower bound.

- (a) Derive a lower bound on the $1/\sqrt{2}$ covering number of $\mathbb{T}^d(s)$ in the Euclidean norm. (*Hint:* The Gilbert-Varshamov lemma could be useful to you).

- (b) Use part (a) and a Gaussian comparison result to show that $\mathcal{G}(\mathbb{T}^d(s)) \gtrsim \sqrt{s \log(\frac{ed}{s})}$.

Exercise 5.9 (Gaussian complexity of ellipses). Recall that the space $\ell^2(\mathbb{N})$ consists of all real sequences $(\theta_j)_{j=1}^\infty$ such that $\sum_{j=1}^\infty \theta_j^2 < \infty$. Given a strictly positive sequence $(\mu_j)_{j=1}^\infty \in \ell^2(\mathbb{N})$, consider the associated ellipse $\mathcal{E} := \{(\theta_j)_{j=1}^\infty \mid \sum_{j=1}^\infty \theta_j^2 / \mu_j^2 \leq 1\}$. Ellipses of this form will play an important role in our subsequent analysis of the statistical properties of reproducing kernel Hilbert spaces.

- (a) Prove that the Gaussian complexity satisfies the bounds

$$\sqrt{\frac{2}{\pi}} \left(\sum_{j=1}^\infty \mu_j^2 \right)^{1/2} \leq \mathcal{G}(\mathcal{E}) \leq \left(\sum_{j=1}^\infty \mu_j^2 \right)^{1/2},$$

Hint: Parts of previous problems may be helpful to you.

- (b) For a given radius $r > 0$ consider the truncated set

$$\tilde{\mathcal{E}}(r) := \mathcal{E} \cap \left\{ (\theta_j)_{j=1}^\infty \mid \sum_{j=1}^\infty \theta_j^2 \leq r^2 \right\}.$$

Obtain upper and lower bounds on the Gaussian complexity $\mathcal{G}(\tilde{\mathcal{E}}(r))$ that are tight up to universal constants, independent of r and $(\mu_j)_{j=1}^\infty$. *Hint:* Try to reduce the problem to an instance of (a).

Exercise 5.10 (Concentration of Gaussian suprema). Let $\{X_\theta, \theta \in \mathbb{T}\}$ be a zero-mean Gaussian process, and define $Z = \sup_{\theta \in \mathbb{T}} X_\theta$. Prove that

$$\mathbb{P}[|Z - \mathbb{E}[Z]| \geq \delta] \leq 2e^{-\frac{\delta^2}{2\sigma^2}},$$

1 where $\sigma^2 := \sup_{\theta \in \mathbb{T}} \text{var}(X_\theta)$ is the maximal variance of the process.

2 **Exercise 5.11.** In this exercise, we work through the details of Example 5.13.

3 (a) Show that the maximum singular value $\|\mathbf{W}\|_{\text{op}}$ has the variational representa-
4 tion (5.40).

5 (b) Defining the random variable $X_\Theta = \langle \mathbf{W}, \Theta \rangle$, show that the stochastic process
6 $\{X_\Theta, \Theta \in \mathbb{M}^{n,d}(1)\}$ is zero-mean, and sub-Gaussian with respect to the Frobenius
7 norm $\|\Theta - \Theta'\|_F$.

8 (c) Prove the upper bound (5.42).

9 (d) Prove the upper bound (5.43) on the metric entropy.

Exercise 5.12 (Gaussian contraction inequality). For each $j = 1, \dots, d$, let $\phi_j : \mathbb{R} \rightarrow \mathbb{R}$ be a 1-Lipschitz function, meaning that $|\phi_j(s) - \phi_j(t)| \leq |s - t|$ for all $s, t \in \mathbb{R}$, and suppose moreover that $\phi_j(0) = 0$. Given a set $\mathbb{T} \subseteq \mathbb{R}^d$, consider the set

$$\phi(\mathbb{T}) := \left\{ (\phi_1(\theta_1), \phi_2(\theta_2), \dots, \phi_d(\theta_d)) \mid \theta \in \mathbb{T} \right\} \subseteq \mathbb{R}^d.$$

10 Prove the Gaussian contraction inequality $\mathcal{G}(\phi(\mathbb{T})) \leq \mathcal{G}(\mathbb{T})$.

Exercise 5.13 (Details of Example 5.21). Recall the set $\mathbb{M}^{n,d}(1)$ from Example 5.21. Show that

$$\log M(\delta; \mathbb{M}^{n,d}(1); \|\cdot\|_F) \lesssim (n + d) \log(1/\delta) \quad \text{for all } \delta \in (0, 1/2).$$

11 **Exercise 5.14** (Maximum singular value of Gaussian random matrices). In this exer-
12 cise, we explore one method for obtaining tail bounds on the maximal singular value of
13 a Gaussian random matrix $\mathbf{W} \in \mathbb{R}^{n \times d}$ with i.i.d. $N(0, 1)$ entries.

14 (a) To build intuition, let us begin by doing a simple simulation. Write a short
15 computer program to generate Gaussian random matrices $\mathbf{W} \in \mathbb{R}^{n \times d}$ for $n = 1000$
16 and $d = \alpha n$, and to compute the maximum singular value of \mathbf{W}/\sqrt{n} , denoted
17 by $\gamma_{\max}(\mathbf{W})/\sqrt{n}$. Perform $T = 20$ trials for each value of α in the set $\{0.1 +$
18 $k(0.025), k = 1, \dots, 100\}$. Plot the resulting curve of α versus the average of
19 $\gamma_{\max}(\mathbf{W})/\sqrt{n}$.

(b) Now let's do some analysis to understand this behavior. Prove that

$$\gamma_{\max}(\mathbf{W}) = \sup_{u \in \mathbb{S}^{n-1}} \sup_{v \in \mathbb{S}^{d-1}} u^T \mathbf{W} v,$$

where $\mathbb{S}^{d-1} = \{y \in \mathbb{R}^d \mid \|y\|_2 = 1\}$ is the d -dimensional Euclidean sphere. 1

(c) Observe that $Z_{u,v} := u^T \mathbf{W} v$ defines a Gaussian process indexed by the Cartesian product $\mathbb{T} := \mathbb{S}^{n-1} \times \mathbb{S}^{d-1}$. Prove the upper bound

$$\mathbb{E}[\gamma_{\max}(\mathbf{W})] = \mathbb{E}\left[\sup_{(u,v) \in \mathbb{T}} u^T \mathbf{W} v\right] \leq \sqrt{n} + \sqrt{d}.$$

Hint: For $(u, v) \in \mathbb{S}^{n-1} \times \mathbb{S}^{d-1}$, consider the zero-mean Gaussian variable

$$Y_{u,v} = \langle g, u \rangle + \langle h, v \rangle,$$

where $g \in N(0, I_{n \times n})$ and $h \sim N(0, I_{d \times d})$ are independent Gaussian random 2
vectors. We thus obtain a second Gaussian process $\{Y_{u,v}, (u, v) \in \mathbb{S}^{n-1} \times \mathbb{S}^{d-1}\}$, 3
and you may find it useful to compare $\{Z_{u,v}\}$ and $\{Y_{u,v}\}$. 4

(d) Prove that

$$\mathbb{P}\left[\gamma_{\max}(\mathbf{W})/\sqrt{n} \geq 1 + \sqrt{\frac{d}{n}} + t\right] \leq 2e^{-\frac{nt^2}{2}}.$$