High Dimensional Statistic

Xin Wen

Contents

I Part One

Part One

1. Basic tail and concentration bounds

From Markov to Chernoff

The most elementary **tail bound** is **Markov's inequality**: given a non-negative random variable X with finite mean, we have

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t} \quad \text{for all } t > 0$$

This is a simple instance of an upper tail bound. For a random variable X that also has a finite variance, we have Chebyshev's inequality :

$$\mathbb{P}[|X - \mu| \geq t] \leq \frac{\operatorname{var}(X)}{t^2} \quad \text{for all } t > 0$$

This is a simple form of **concentration inequality**, guaranteeing that X is close to its mean $\mu = \mathbb{E}[X]$ whenever its variance is small. Observe that Chebyshev's inequality follows by applying Markov's inequality to the non-negative random variable $Y = (X - \mu)^2$. Both Markov's and Chebyshev's inequalities are sharp, meaning that they cannot be improved in general (see Exercise 2.1).

There are various extensions of Markov's inequality applicable to random variables with higher-order moments. For instance, whenever X has a central moment of order k, an application of Markov's inequality to the random variable $|X - \mu|^k$ yields that

$$\mathbb{P}[|X - \mu| \geq t] \leq \frac{\mathbb{E}\left[|X - \mu|^k\right]}{t^k} \quad \text{for all } t > 0 \tag{1.1}$$

Of course, the same procedure can be applied to functions other than polynomials $|X - \mu|^k$. For instance, suppose that the random variable X has a moment generating function in a neighborhood of zero, meaning that there is some constant $b > 0$ such that the function $\varphi(\lambda) = \mathbb{E}\left[e^{\lambda(X-\mu)}\right]$ exists for all $\lambda \leq |b|$. In this case, for any $\lambda \in [0,b]$, we may apply Markov's inequality to the random variable $Y = e^{\lambda(X-\mu)}$, thereby obtaining the upper bound

$$\mathbb{P}[(X-\mu) \geq t] = \mathbb{P}\left[e^{\lambda(X-\mu)} \geq e^{\lambda t}\right] \leq \frac{\mathbb{E}\left[e^{\lambda(X-\mu)}\right]}{e^{\lambda t}}$$

Optimizing our choice of λ so as to obtain the tightest result yields the **Chernoff bound**, namely, the inequality

$$\log \mathbb{P}[(X-\mu) \geq t] \leq \inf_{\lambda \in [0,b]} \left\{ \log \mathbb{E}\left[e^{\lambda(X-\mu)}\right] - \lambda t \right\} \tag{1.2}$$

As we explore in Exercise 2.3, the moment bound (1.1) with an optimal choice of k is never worse than the bound (1.2) based on the moment generating function. Nonetheless, the Chernoff bound is most widely used in practice, possibly due to the ease of manipulating moment generating functions. Indeed, a variety of important tail bounds can be obtained as particular cases of inequality (1.2).

Sub-Gaussian variables and Hoeffaing bounds

■ **Example 1.1 — Gaussian tail bounds.** Let $X \sim \mathcal{N}\left(\mu,\sigma^2\right)$ be a Gaussian random variable with mean μ and variance σ^2. By a straightforward calculation, we find that X has the moment generating function

$$\mathbb{E}\left[e^{\lambda X}\right] = e^{\mu\lambda + \frac{\sigma^2\lambda^2}{2}}, \quad \text{valid for all } \lambda \in \mathbb{R}$$

Substituting this expression into the optimization problem defining the optimized Chernoff bound (1.2), we obtain

$$\inf_{\lambda \geq 0} \left\{ \log \mathbb{E}\left[e^{\lambda(X-\mu)}\right] - \lambda t \right\} = \inf_{\lambda \geq 0} \left\{ \frac{\lambda^2\sigma^2}{2} - \lambda t \right\} = -\frac{t^2}{2\sigma^2}$$

where we have taken derivatives in order to find the optimum of this quadratic function about λ. Returning to the Chernoff bound (1.2), we conclude that any $\mathcal{N}\left(\mu,\sigma^2\right)$ random variable satisfies the **upper deviation inequality**

$$\mathbb{P}[X \geq \mu + t] \leq e^{-\frac{t^2}{2\sigma^2}} \quad \text{for all } t \geq 0 \tag{1.3}$$

In fact, this bound is sharp up to polynomial-factor corrections, as shown by our exploration of the Mills ratio in Exercise 2.2.

Definition 1.0.1 — sub-Gaussian. A random variable X with mean $\mu = \mathbb{E}[X]$ is **sub-Gaussian** if there is a positive number σ such that

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\sigma^2\lambda^2/2} \quad \text{for all } \lambda \in \mathbb{R} \tag{1.4}$$

The constant σ is referred to as the sub-Gaussian parameter. Naturally, any Gaussian variable with variance σ^2 is sub-Gaussian with parameter σ, as should be clear from the calculation described in Example 2.1.

The condition (1.4), when combined with the Chernoff bound as in Example 2.1, shows that, if X is sub-Gaussian with parameter σ, then it satisfies the upper deviation inequality (1.3). Moreover, by the symmetry of the definition, the variable $-X$ is sub-Gaussian if and only if X is sub-Gaussian, so that we also have the lower deviation inequality $\mathbb{P}[X \leq \mu - t] \leq e^{-\frac{t^2}{2\sigma^2}}$, valid for all $t \geq 0$. Combining the pieces, we conclude that any sub-Gaussian variable satisfies the concentration inequality

$$\mathbb{P}[|X-\mu| \geq t] \leq 2e^{-\frac{\lambda^2}{2\sigma^2}} \quad \text{for all } t \in \mathbb{R}$$

■ **Example 1.2 — Rademacher variables.** A **Rademacher random variable** ε takes the values {-1,+1} equiprobably. We claim that it is sub-Gaussian with parameter $\sigma = 1$. By taking expectations and using the power-series expansion for the exponential, we obtain

$$\begin{aligned}\mathbb{E}\left[e^{\lambda\varepsilon}\right] = \frac{1}{2}\left\{e^{-\lambda} + e^{\lambda}\right\} &= \frac{1}{2}\left\{\sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!} + \sum_{k=0}^{\infty}\frac{(\lambda)^k}{k!}\right\} \\ &= \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{(2k)!} \\ &\leq 1 + \sum_{k=1}^{\infty}\frac{\lambda^{2k}}{2^k k!} \\ &= e^{\lambda^2/2}\end{aligned}$$

which shows that ε is sub-Gaussian with parameter $\sigma = 1$ as claimed.

We now generalize the preceding example to show that any bounded random variable is also sub-Gaussian.

■ **Example 1.3 — Bounded random variables.** Let X be zero-mean, and supported on some interval $[a,b]$. Letting X' be an independent copy, for any $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}_X\left[e^{\lambda X}\right] = \mathbb{E}_X\left[e^{\lambda(X - \mathbb{E}_{X'}[X'])}\right] \leq \mathbb{E}_{X,X'}\left[e^{\lambda(X-X')}\right]$$

where the inequality follows from the convexity of the exponential, and Jensen's inequality. Letting ε be an independent Rademacher variable, note that the distribution of $(X - X')$ is the same as that of $\varepsilon(X - X')$, so that we have

$$\mathbb{E}_{X,X'}\left[e^{\lambda(X-X')}\right] = \mathbb{E}_{X,X'}\left[\mathbb{E}_{\varepsilon}\left[e^{\lambda\varepsilon(X-X')}\right]\right] \leq^{(i)} \mathbb{E}_{X,X'}\left[e^{\frac{\lambda^2(X-X')^2}{2}}\right]$$

where step (i) follows from the result of Example 1.2 , applied conditionally with (X, X') held fixed. Since $|X - X'| \leq b - a$, we are guaranteed that

$$\mathbb{E}_{X,X'}\left[e^{\frac{\lambda^2(X-X')^2}{2}}\right] \leq e^{\frac{\lambda^2(b-a)^2}{2}}$$

Putting together the pieces, we have shown that X is sub-Gaussian with parameter at most $\sigma = b - a$. This result is useful but can be sharpened. In Exercise 2.4, we work through a more involved argument to show that X is sub-Gaussian with parameter at most $\sigma = \frac{b-a}{2}$

The technique used in Example 1.3 is a simple example of a symmetrization argument, in which we first introduce an independent copy X', and then symmetrize the problem with a Rademacher variable. Such symmetrization arguments are useful in a variety of contexts, as will be seen in later chapters.

Just as the property of Gaussianity is preserved by linear operations, so is the property of sub-Gaussianity. For instance, if X_1 and X_2 are independent sub-Gaussian variables with parameters σ_1 and σ_2, then $X_1 + X_2$ is sub-Gaussian with parameter $\sqrt{\sigma_1^2 + \sigma_2^2}$. See Exercise 2.13 for verification of this fact, as well as some related properties.

Proposition 1.0.1 — **Hoeffding bound-sums of independent sub-Gaussian random variables.** Suppose that the variables $X_i, i = 1, \ldots, n$, are independent, and X_i has mean μ_i and sub-Gaussian parameter σ_i. Then for all $t \geq 0$, we have

$$\mathbb{P}\left[\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right] \leq \exp\left\{-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right\}$$

The Hoeffding bound is often stated only for the special case of bounded random variables. In particular, if $X_i \in [a, b]$ for all $i = 1, 2, \ldots, n$, then from the result of Exercise 2.4, it is sub-Gaussian with parameter $\sigma = \frac{b-a}{2}$, so that we obtain the bound

$$\mathbb{P}\left[\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right] \leq e^{-\frac{2t^2}{n(b-a)^2}}$$

Although the Hoeffding bound is often stated in this form, the basic idea applies somewhat more generally to sub-Gaussian variables, as we have given here.

We conclude our discussion of sub-Gaussianity with a result that provides three different characterizations of sub-Gaussian variables.

1. First, the most direct way in which to establish sub-Gaussianity is by computing or bounding the moment generating function, as we have done in Example 1.1.
2. A second intuition is that any sub-Gaussian variable is dominated in a certain sense by a Gaussian variable.
3. Third, sub-Gaussianity also follows by having suitably tight control on the moments of the random variable.

The following result shows that all three notions are equivalent in a precise sense.

Theorem 1.0.2 — Equivalent characterizations of sub-Gaussian variables. Given any zero-mean random variable X, the following properties are equivalent:

1. There is a constant $\sigma \geq 0$ such that
$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}} \quad \text{for all } \lambda \in \mathbb{R}$$
2. There is a constant $c \geq 0$ and Gaussian random variable $Z \sim \mathcal{N}\left(0,\tau^2\right)$ such that
$$\mathbb{P}[|X| \geq s] \leq c\mathbb{P}[|Z| \geq s] \quad \text{for all } s \geq 0$$
3. There is a constant $\theta \geq 0$ such that
$$\mathbb{E}\left[X^{2k}\right] \leq \frac{(2k)!}{2^k k!}\theta^{2k} \quad \text{for all } k = 1,2,\dots$$
4. There is a constant $\sigma \geq 0$ such that
$$\mathbb{E}\left[e^{\frac{\lambda x^2}{2\sigma^2}}\right] \leq \frac{1}{\sqrt{1-\lambda}} \quad \text{for all } \lambda \in [0,1)$$

Proof. In this appendix, we prove Theorem 1.0.2. We establish the equivalence by proving the circle of implications (I) ⇒ (II) ⇒ (III) ⇒ (I), followed by the equivalence (I) ⇔ (IV)

1. Implication (I) ⇒ (II) : If X is zero-mean and sub-Gaussian with parameter σ, then we claim that, for $Z \sim \mathcal{N}\left(0,2\sigma^2\right)$
$$\frac{\mathbb{P}[X \geq t]}{\mathbb{P}[Z \geq t]} \leq \sqrt{8}e$$
for all $t \geq 0$ showing that X is majorized by Z with constant $c = \sqrt{8}e$. On one hand, by the subGaussianity of X, we have $\mathbb{P}[X \geq t] \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$ for all $t \geq 0$. On the other hand, by the Mills ratio for Gaussian tails, if $Z \sim \mathcal{N}\left(0,2\sigma^2\right)$, then we have
$$\mathbb{P}[Z \geq t] \geq \left(\frac{\sqrt{2}\sigma}{t} - \frac{(\sqrt{2}\sigma)^3}{t^3}\right)e^{-\frac{t^2}{4\sigma^2}} \quad \text{for all } t > 0 \tag{1.5}$$
(See Exercise 2.2 for a derivation of this inequality.) We split the remainder of our analysis into two cases.
   (a) Case 1: First, suppose that $t \in [0,2\sigma]$. since the function $\Phi(t) = \mathbb{P}[Z \geq t]$ is decreasing, for all t in this interval,
$$\mathbb{P}[Z \geq t] \geq \mathbb{P}[Z \geq 2\sigma] \geq \left(\frac{1}{\sqrt{2}} - \frac{1}{2\sqrt{2}}\right)e^{-1} = \frac{1}{\sqrt{8}e}$$
   Since $\mathbb{P}[X \geq t] \leq 1$, we conclude that $\frac{\mathbb{P}[X\geq t]}{\mathbb{P}[Z\geq t]} \leq \sqrt{8}e$ for all $t \in [0,2\sigma]$
   (b) Case 2 : Otherwise, we may assume that $t > 2\sigma$. In this case, by combining the Mills ratio (2.53) and the sub-Gaussian tail bound and making the substitution

$s = t/\sigma$, we find that

$$\begin{aligned}\sup_{t>2\sigma} \frac{\mathbb{P}[X \geq t]}{\mathbb{P}[Z \geq t]} &\leq \sup_{s>2} \frac{e^{-\frac{s^2}{4}}}{\left(\frac{\sqrt{2}}{s} - \frac{(\sqrt{2})^3}{s^3}\right)} \\ &\leq \sup_{s>2} s^3 e^{-\frac{s^2}{4}} \\ &\leq \sqrt{8}e\end{aligned}$$

where the last step follows from a numerical calculation.

2. Implication (II) $\Rightarrow$ (III) : Suppose that X is majorized by a zero-mean Gaussian with variance τ^2. since X^{2k} is a non-negative random variable, we have

$$\mathbb{E}\left[X^{2k}\right] = \int_0^\infty \mathbb{P}\left[X^{2k} > s\right] ds = \int_0^\infty \mathbb{P}\left[|X| > s^{1/(2k)}\right] ds$$

Under the majorization assumption, there is some constant $c \geq 1$ such that

$$\int_0^\infty \mathbb{P}\left[|X| > s^{1/(2k)}\right] ds \leq c \int_0^\infty \mathbb{P}\left[|Z| > s^{1/(2k)}\right] ds = c\mathbb{E}\left[Z^{2k}\right]$$

where $Z \sim \mathcal{N}\left(0, \tau^2\right)$. The polynomial moments of Z are given by

$$\mathbb{E}\left[Z^{2k}\right] = \frac{(2k)!}{2^k k!}\tau^{2k}, \quad \text{for } k = 1,2,\ldots$$

whence

$$\mathbb{E}\left[X^{2k}\right] \leq c\mathbb{E}\left[Z^{2k}\right] = c\frac{(2k)!}{2^k k!}\tau^{2k} \leq \frac{(2k)!}{2^k k!}(c\tau)^{2k}$$

for all $k = 1,2,\ldots$

Consequently, the moment bound III holds with $\theta = c\tau$.

3. Implication (III) $\Rightarrow$ (I) : For each $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}\left[e^{\lambda X}\right] \leq 1 + \sum_{k=2}^\infty \frac{|\lambda|^k \mathbb{E}\left[|X|^k\right]}{k!}$$

where we have used the fact $\mathbb{E}[X] = 0$ to eliminate the term involving $k = 1$. If X is symmetric around zero, then all of its odd moments vanish, and by applying our assumption on $\theta(X)$, we obtain

$$\mathbb{E}\left[e^{\lambda X}\right] \leq 1 + \sum_{k=1}^\infty \frac{\lambda^{2k}}{(2k)!}\frac{(2k)!\theta^{2k}}{2^k k!} = e^{\frac{\lambda^2\theta^2}{2}}$$

which shows that X is sub-Gaussian with parameter θ. When X is not symmetric, we can bound the odd moments in terms of the even ones as

$$\mathbb{E}\left[|\lambda X|^{2k+1}\right] \leq^{(i)} \left(\mathbb{E}\left[|\lambda X|^{2k}\right]\mathbb{E}\left[|\lambda X|^{2k+2}\right]\right)^{1/2} \leq^{(ii)} \frac{1}{2}\left(\lambda^{2k}\mathbb{E}\left[X^{2k}\right] + \lambda^{2k+2}\mathbb{E}\left[X^{2k+2}\right]\right)$$

where step (i) follows from the Cauchy-Schwarz inequality; and step (ii) follows from the arithmetic-geometric mean inequality. Applying this bound to the power-series expansion (2.55), we obtain

$$\begin{aligned}\mathbb{E}\left[e^{\lambda X}\right] &\leq 1+\left(\frac{1}{2}+\frac{1}{2\cdot 3!}\right)\lambda^2\mathbb{E}\left[X^2\right]+\sum_{k=2}^{\infty}\left(\frac{1}{(2k)!}+\frac{1}{2}\left[\frac{1}{(2k-1)!}+\frac{1}{(2k+1)!}\right]\right)\lambda^{2k}\mathbb{E}\left[X^{2k}\right]\\ &\leq \sum_{k=0}^{\infty}2^k\frac{\lambda^{2k}\mathbb{E}\left[X^{2k}\right]}{(2k)!}\\ &\leq e^{\frac{(\sqrt{2}\lambda\theta)^2}{2}}\end{aligned}$$

which establishes the claim.

4. Implication (I) $\Rightarrow$ (IV) : This result is obvious for $s=0$. For $s\in(0,1)$, we begin with the sub-Gaussian inequality $\mathbb{E}\left[e^{\lambda X}\right]\leq e^{\frac{\lambda^2\sigma^2}{2}}$, and multiply both sides by $e^{-\frac{\lambda^2\sigma^2}{2s}}$, thereby obtaining

$$\mathbb{E}\left[e^{\lambda X-\frac{\lambda^2\sigma^2}{2s}}\right]\leq e^{\frac{\lambda^2\sigma^2(s-1)}{2s}}$$

since this inequality holds for all $\lambda\in\mathbb{R}$, we may integrate both sides over $\lambda\in\mathbb{R}$, using Fubini's theorem to justify exchanging the order of integration. On the right-hand side, we have

$$\int_{-\infty}^{\infty}\exp\left(\frac{\lambda^2\sigma^2(s-1)}{2s}\right)d\lambda=\frac{1}{\sigma}\sqrt{\frac{2\pi s}{1-s}}$$

Turning to the left-hand side, for each fixed $x\in\mathbb{R}$, we have

$$\int_{-\infty}^{\infty}\exp\left(\lambda x-\frac{\lambda^2\sigma^2}{2s}\right)d\lambda=\frac{\sqrt{2\pi s}}{\sigma}e^{\frac{sx^2}{2\sigma^2}}$$

Taking expectations with respect to X, we conclude that

$$\mathbb{E}\left[e^{\frac{sX^2}{2\sigma^2}}\right]\leq\frac{\sigma}{\sqrt{2\pi s}}\frac{1}{\sigma}\sqrt{\frac{2\pi s}{1-s}}=\frac{1}{\sqrt{1-s}}$$

which establishes the claim.

5. Implication (IV) $\Rightarrow$ (I) : Applying the bound $e^u\leq u+e^{9u^2/16}$ with $u=\lambda X$ and then taking expectations, we find that

$$\mathbb{E}\left[e^{\lambda X}\right]\leq\mathbb{E}[\lambda X]+\mathbb{E}\left[e^{\frac{92^2\chi^2}{16}}\right]=\mathbb{E}\left[e^{\frac{s^2}{2\sigma^2}}\right]\leq\frac{1}{\sqrt{1-s}}$$

valid whenever $s=\frac{9}{8}\lambda^2\sigma^2$ is strictly less than 1. Noting that $\frac{1}{\sqrt{1-s}}\leq e^s$ for all $s\in\left[0,\frac{1}{2}\right]$ and that $s<\frac{1}{2}$ whenever $|\lambda|<\frac{2}{3\sigma}$, we conclude that

$$\mathbb{E}\left[e^{\lambda X}\right]\leq e^{\frac{9}{8}\lambda^2\sigma^2}\quad\text{for all }|\lambda|<\frac{2}{3\sigma}\tag{1.6}$$

It remains to establish a similar upper bound for $|\lambda| \geq \frac{2}{3\sigma}$. Note that, for any $\alpha > 0$, the functions $f(u) = \frac{u^2}{2\alpha}$ and $f^*(v) = \frac{\alpha v^2}{2}$ are conjugate duals. Thus, the Fenchel-Young inequality implies that $uv \leq \frac{u^2}{2\alpha} + \frac{\alpha v^2}{2}$, valid for all $u,v \in \mathbb{R}$ and $\alpha > 0$. We apply this inequality with $u = \lambda, v = X$ and $\alpha = c/\sigma^2$ for a constant $c > 0$ to be chosen; doing so yields

$$\mathbb{E}\left[e^{\lambda X}\right] \leq \mathbb{E}\left[e^{\frac{\lambda^2\sigma^2}{2c} + \frac{c^2}{2\sigma^2}}\right] = e^{\frac{\lambda^2\sigma^2}{2c}}\mathbb{E}\left[e^{\frac{cX^2}{2\sigma^2}}\right] \leq e^{\frac{\lambda^2\sigma^2}{2c}}e^{c}$$

where step (ii) is valid for any $c \in (0,1/2)$, using the same argument that led to the bound (1.6). In particular, setting $c = 1/4$ yields $\mathbb{E}\left[e^{\lambda X}\right] \leq e^{2\lambda^2\sigma^2}e^{1/4}$ Finally, when $|\lambda| \geq \frac{2}{3\sigma}$, then we have $\frac{1}{4} \leq \frac{9}{16}\lambda^2\sigma^2$, and hence

$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{2\lambda^2\sigma^2 + \frac{9}{16}\lambda^2\sigma^2} \leq e^{3\lambda^2\sigma^2}$$

This inequality, combined with the bound (1.6), completes the proof.

■

Sub-exponential variables and Bernstein bounds

Definition 1.0.2 — sub -exponential. A random variable X with mean $\mu = \mathbb{E}[X]$ is **sub -exponential** if there are non-negative parameters (v, α) such that

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{v^2\lambda^2}{2}} \quad \text{for all } |\lambda| < \frac{1}{\alpha} \tag{1.7}$$

It follows immediately from this definition that any sub-Gaussian variable is also subexponential-in particular, with $v = \sigma$ and $\alpha = 0$, where we interpret $1/0$ as being the same as $+\infty$. However, the converse statement is not true, as shown by the following calculation:

■ **Example 1.4 — Sub-exponential but not sub-Gaussian.** Let $Z \sim \mathcal{N}(0,1)$, and consider the random variable $X = Z^2$ $(X \sim \chi^2(1))$. For $\lambda < \frac{1}{2}$, we have

$$\begin{aligned}\mathbb{E}\left[e^{\lambda(X-1)}\right] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{\lambda\left(z^2-1\right)}e^{-z^2/2}dz \\ &= \frac{e^{-\lambda}}{\sqrt{1-2\lambda}}\end{aligned}$$

For $\lambda > \frac{1}{2}$, the moment generating function is infinite, which reveals that X is not sub-Gaussian.

As will be seen momentarily, **the existence of the moment generating function in a neighborhood of zero is actually an equivalent definition of a sub-exponential variable**. Let us verify directly that condition (1.7) is satisfied. Following some calculus, we find that

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2} = e^{4\lambda^2/2}, \quad \text{for all } |\lambda| < \frac{1}{4}$$

which shows that X is sub-exponential with parameters $(v,\alpha) = (2,4)$.

As with sub-Gaussianity, the control (1.7) on the moment generating function, when combined with the Chernoff technique, yields deviation and concentration inequalities for sub-exponential variables. When t is small enough, these bounds are sub-Gaussian in nature (i.e., with the exponent quadratic in t), whereas for larger t, the exponential component of the bound scales linearly in t. We summarize in the following:

Proposition 1.0.3 — Sub-exponential tail bound. Suppose that X is sub-exponential with parameters (v,α). Then

$$\mathbb{P}[X-\mu \geq t] \leq \begin{cases} e^{-\frac{t^2}{2v^2}} & \text{if } 0 \leq t \leq \frac{v^2}{\alpha} \\ e^{-\frac{1}{2\alpha}} & \text{for } t > \frac{v^2}{\alpha} \end{cases}$$

As with the Hoeffding inequality, similar bounds can be derived for the left-sided event $\{X-\mu \leq -t\}$, as well as the two-sided event $\{|X-\mu| \geq t\}$, with an additional factor of 2 in the latter case.

Proof. By recentering as needed, we may assume without loss of generality that $\mu = 0$. We follow the usual Chernoff-type approach: combining it with the definition (1.7) of a sub-exponential variable (i) yields the upper bound

$$\mathbb{P}[X \geq t] \leq e^{-\lambda t}\mathbb{E}\left[e^{\lambda X}\right] \leq^{(i)} \exp\underbrace{\left(-\lambda t + \frac{\lambda^2 v^2}{2}\right)}_{g(\lambda,t)}, \quad \text{valid for all } \lambda \in \left[0,\alpha^{-1}\right)$$

In order to complete the proof, it remains to compute, for each fixed $t \geq 0$, the quantity $g^*(t) := \inf_{\lambda\in[0,\alpha^{-1})} g(\lambda,t)$. Note that the unconstrained minimum of the function $g(\cdot,t)$ occurs at $\lambda^* = t/v^2$. Consider whether λ^* lies in the intervl $[0,\frac{1}{\alpha})$. If $0 \leq t < \frac{v^2}{\alpha}$, then this unconstrained optimum corresponds to the constrained minimum as well, so that $g^*(t) = -\frac{t^2}{2v^2}$ over this interval.

Otherwise, we may assume that $t \geq \frac{v^2}{\alpha}$. In this case, since the function $g(\cdot,t)$ is monotonically decreasing in the interval $[0,\lambda^*)$, the constrained minimum is achieved at the boundary point $\lambda^\dagger = \alpha^{-1}$, and we have

$$g^*(t) = g\left(\lambda^\dagger,t\right) = -\frac{t}{\alpha} + \frac{1}{2\alpha}\frac{v^2}{\alpha} \leq^{(i)} -\frac{t}{2\alpha}$$

where inequality (i) uses the fact that $\frac{v^2}{\alpha} \leq t$ ■

This direct calculation may be impracticable in many settings, so it is natural to seek alternative approaches. One such method is based on control of the polynomial moments of X.

Theorem 1.0.4 — Bernstein's condition. Given a random variable X with mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \mathbb{E}\left[X^2\right] - \mu^2$, we say that **Bernstein's condition** with parameter b holds if

$$\left|\mathbb{E}\left[(X-\mu)^k\right]\right| \leq \frac{1}{2}k!\sigma^2 b^{k-2} \quad \text{for } k = 2,3,4,\dots \tag{1.8}$$

One sufficient condition for Bernstein's condition to hold is that X be bounded; in particular, if $|X-\mu| \leq b$, then it is straightforward to verify that condition (1.8) holds. Even for bounded variables, our next result will show that the Bernstein condition can be used to obtain tail bounds that may be tighter than the Hoeffding bound. Moreover, Bernstein's condition is also satisfied by various unbounded variables, a property which lends it much broader applicability.

When X satisfies the Bernstein condition, then it is sub-exponential with parameters determined by σ^2 and b. Indeed, by the power-series expansion of the exponential, we have

$$\begin{aligned}\mathbb{E}\left[e^{\lambda(X-\mu)}\right] &= 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \lambda^k \frac{\mathbb{E}\left[(X-\mu)^k\right]}{k!} \\ &\leq^{(i)} 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2}\sum_{k=3}^{\infty}(|\lambda|b)^{k-2}\end{aligned}$$

where the inequality (i) makes use of the Bernstein condition (1.8). For any $|\lambda| < 1/b$, we can sum the geometric series so as to obtain

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq 1 + \frac{\lambda^2\sigma^2/2}{1-b|\lambda|} \leq^{(ii)} e^{\frac{\lambda^2\sigma^2/2}{1-b|\lambda|}} \tag{1.9}$$

where inequality (ii) follows from the bound $1 + t \leq e^t$. Consequently, we conclude that

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\lambda^2(\sqrt{2}\sigma)^2}{2}} \quad \text{for all } |\lambda| < \frac{1}{2b}$$

showing that X is sub-exponential with parameters $(\sqrt{2}\sigma, 2b)$. As a consequence, an application of Proposition 1.0.3 leads directly to tail bounds on a random variable satisfying the Bernstein condition (1.8). However, the resulting tail bound can be sharpened slightly, at least in terms of constant factors, by making direct use of the upper bound (1.9). We summarize in the following:

Proposition 1.0.5 — Bernstein-type bound. For any random variable satisfying the Bernstein condition (1.8), we have

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\lambda^2\sigma^2/2}{1-b|\lambda|}} \quad \text{for all } |\lambda| < \frac{1}{b} \tag{1.10}$$

and, moreover, the concentration inequality

$$\mathbb{P}[|X-\mu| \geq t] \leq 2e^{-\frac{t^2}{2(\sigma^2+bt)}} \quad \text{for all } t \geq 0 \tag{1.11}$$

We proved inequality (1.10) in the discussion preceding this proposition. Using this bound on the moment generating function, the tail bound (1.11) follows by setting $\lambda = \frac{t}{bt+\sigma^2} \in \left[0, \frac{1}{b}\right)$ in the Chernoff bound, and then simplifying the resulting expression.

> Proposition 1.0.5 has an important consequence even for bounded random variables (i.e., those satisfying $|X-\mu| \leq b$). The most straightforward way to control such variables is by exploiting the boundedness to show that $(X-\mu)$ is sub-Gaussian with parameter b (see Exercise 2.4), and then applying a Hoeffding-type inequality (see Proposition 1.0.1). Alternatively, using the fact that any bounded variable satisfies the Bernstein condition (1.9), we can also apply Proposition 1.0.5 , thereby obtaining the tail bound (1.11), that involves both the variance σ^2 and the bound b. This tail bound shows that for suitably small t, the variable X has sub-Gaussian behavior with parameter σ, as opposed to the parameter b that would arise from a Hoeffding approach. Since $\sigma^2 = \mathbb{E}\left[(X-\mu)^2\right] \leq b^2$, this bound is never worse; moreover, it is substantially better when $\sigma^2 \ll b^2$, as would be the case for a random variable that occasionally takes on large values, but has relatively small variance. Such variance-based control frequently plays a key role in obtaining optimal rates in statistical problems, as will be seen in later chapters. For bounded random variables, Bennett's inequality can be used to provide sharper control on the tails (see Exercise 2.7).

Proposition 1.0.6 — **Sum of independent sub-exponential random variables.** In particular, consider an independent sequence $\{X_k\}_{k=1}^n$ of random variables, such that X_k has mean μ_k, and is sub-exponential with parameters (v_k, α_k). We compute the moment generating function

$$\mathbb{E}\left[e^{\lambda \sum_{k=1}^{n}(X_k-\mu_k)}\right] \overset{\text{(i)}}{=} \prod_{k=1}^{n} \mathbb{E}\left[e^{\lambda(X_k-\mu_k)}\right] \leq^{(ii)} \prod_{k=1}^{n} e^{\lambda^2 v_k^2/2}$$

valid for all $|\lambda| < \left(\max_{k=1,\ldots,n} \alpha_k\right)^{-1}$, where equality (i) follows from independence, and inequality (ii) follows since X_k is sub-exponential with parameters (v_k, α_k). Thus, we conclude that the variable $\sum_{k=1}^{n}(X_k-\mu_k)$ is sub-exponential with the parameters (v_*, α_*), where

$$\alpha_* := \max_{k=1,\ldots,n} \alpha_k \quad \text{and} \quad v_* := \sqrt{\sum_{k=1}^{n} v_k^2}$$

Using the same argument as in Proposition 1.0.3, this observation leads directly to the upper tail bound

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}(X_k-\mu_k) \geq t\right] \leq \begin{cases} e^{-\frac{nt^2}{2\left(v_*^2/n\right)}} & \text{for } 0 \leq t \leq \frac{v_*^2}{n\alpha_*} \\ e^{-\frac{nt}{2\alpha*}} & \text{for } t > \frac{v_*^2}{n\alpha_*} \end{cases}$$

along with similar two-sided tail bounds.

Let us illustrate our development thus far with some examples.

■ **Example 1.5 — (χ^2 -variables).** A chi-squared $\left(\chi^2\right)$ random variable with n degrees of freedom, denoted by $Y \sim \chi_n^2$, can be represented as the sum $Y=\sum_{k=1}^n Z_k^2$ where $Z_k \sim \mathcal{N}(0,1)$ are i.i.d. variates. As discussed in Example 2.8, the variable Z_k^2 is sub-exponential with parameters $(2,4)$. Consequently, since the variables $\left\{Z_k\right\}_{k=1}^n$ are independent, the χ^2 -variate Y is sub-exponential with parameters $(v,\alpha)=(2\sqrt{n},4)$, and the preceding discussion yields the two-sided tail bound

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{k=1}^n Z_k^2-1\right| \geq t\right] \leq 2e^{-nt^2/8}, \quad \text{for all } t \in (0,1) \tag{1.12}$$

The concentration of χ^2 -variables plays an important role in the analysis of procedures based on taking random projections. A classical instance of the random projection method is the Johnson-Lindenstrauss analysis of metric embedding.

■ **Example 1.6 — Johnson-Lindenstrauss embedding.** As one application of the concentration of χ^2 -variables, consider the following problem. Suppose that we are given $N \geq 2$ distinct vectors $\left\{u^1,\ldots,u^N\right\}$, with each vector lying in $\mathbb{R}^d$. If the data dimension d is large, the idea of dimensionality reduction is to construct a mapping $F: \mathbb{R}^d \rightarrow \mathbb{R}^m-$ with the projected dimension m substantially smaller than $d-$ that preserves some "essential" features of the data set. We might consider preserving pairwise distances, or equivalently norms and inner products. With these motivations in mind, given some tolerance $\delta \in (0,1)$, we might be interested in a mapping F with the guarantee that

$$(1-\delta) \leq \frac{\left\|F\left(u^i\right)-F\left(u^j\right)\right\|_2^2}{\left\|u^i-u^j\right\|_2^2} \leq (1+\delta) \quad \text{for all pairs } u^i \neq u^j \tag{1.13}$$

In words, the projected data set $\left\{F\left(u^1\right),\ldots,F\left(u^N\right)\right\}$ preserves all pairwise squared distances up to a multiplicative factor of δ. Of course, this is always possible if the projected dimension m is large enough, but the goal is to do it with relatively small m

Constructing such a mapping that satisfies the condition (1.13) with high probability turns out to be straightforward as long as the projected dimension is lower bounded as $m \gtrsim \frac{1}{\delta^2}\log N$. Observe that the projected dimension is independent of the ambient dimension d, and scales only logarithmically with the number of data points N.

The construction is probabilistic: first form a random matrix $\mathbf{X} \in \mathbb{R}^{m\times d}$ filled with independent $\mathcal{N}(0,1)$ entries, and use it to define a linear mapping $F: \mathbb{R}^d \rightarrow \mathbb{R}^m$ via $u \mapsto \mathbf{X}u/\sqrt{m}$. We now verify that F satisfies condition (1.13) with high probability. Let $x_i \in \mathbb{R}^d$ denote the i th row of $\mathbf{X}$, and consider some fixed $u \neq 0$. Since x_i is a standard normal vector, the variable $\left\langle x_i, u/\|u\|_2\right\rangle$ follows a $\mathcal{N}(0,1)$ distribution, and hence the quantity

$$Y:=\frac{\|\mathbf{X}u\|_2^2}{\|u\|_2^2}=\sum_{i=1}^m\left\langle x_i, u/\|u\|_2\right\rangle^2$$

follows a χ^2 distribution with m degrees of freedom, using the independence of the rows. Therefore, applying the tail bound (1.12), we find that

$$\mathbb{P}\left[\left|\frac{\|\mathbf{X}u\|_2^2}{m\|u\|_2^2}-1\right| \geq \delta\right] \leq 2e^{-m\delta^2/8} \quad \text{for all } \delta \in (0,1)$$

Rearranging and recalling the definition of F yields the bound

$$\mathbb{P}\left[\frac{\|F(u)\|_2^2}{\|u\|_2^2} \notin [(1-\delta),(1+\delta)]\right] \leq 2e^{-m\delta^2/8}$$

for any fixed $0 \neq u \in \mathbb{R}^d$. Noting that there are $\begin{pmatrix} N \\ 2 \end{pmatrix}$ distinct pairs of data points, we apply the union bound to conclude that

$$\mathbb{P}\left[\frac{\left\|F\left(u^i-u^j\right)\right\|_2^2}{\left\|u^i-u^j\right\|_2^2} \notin [(1-\delta),(1+\delta)] \text{ for some } u^i \neq u^j\right] \leq 2\begin{pmatrix} N \\ 2 \end{pmatrix} e^{-m\delta^2/8}$$

For any $\epsilon \in (0,1)$, this probability can be driven below ϵ by choosing $m > \frac{16}{\delta^2}\log(N/\epsilon)$.

In parallel to Theorem 1.0.7 , there are a number of equivalent ways to characterize a subexponential random variable. The following theorem provides a summary:

Theorem 1.0.7 — Equivalent characterizations of sub-exponential variables. For a zeromean random variable X, the following statements are equivalent:

1. There are non-negative numbers (v,α) such that t

$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\frac{\nu^2\lambda^2}{2}} \quad \text{for all } |\lambda| < \frac{1}{\alpha}$$

2. There is a positive number $c_0 > 0$ such that $\mathbb{E}\left[e^{\lambda X}\right] < \infty$ for all $|\lambda| \leq c_0$
3. There are constants $c_1, c_2 > 0$ such that

$$\mathbb{P}[|X| \geq t] \leq c_1 e^{-c_2 t} \quad \text{for all } t > 0$$

4. The quantity $\gamma := \sup_{k\geq 2}\left[\frac{\mathbb{E}\left[X^k\right]}{k!}\right]^{1/k}$ is finite.

Proof. This appendix is devoted to the proof of Theorem 1.0.7. In particular, we prove the chain of equivalences $I \Leftrightarrow II \Leftrightarrow III$, followed by the equivalence $II \Leftrightarrow IV$.

1. (II) $\Rightarrow$ (I) : The existence of the moment generating function for $|\lambda| < c_0$ implies that $\mathbb{E}\left[e^{\lambda X}\right] = 1 + \frac{\lambda^2\mathbb{E}\left[X^2\right]}{2} + o\left(\lambda^2\right)$ as $\lambda \to 0$. Moreover, an ordinary Taylor-series expansion implies that $e^{\frac{\sigma^2\lambda^2}{2}} = 1 + \frac{\sigma^2\lambda^2}{2} + o\left(\lambda^2\right)$ as $\lambda \to 0$. Therefore, as long as $\sigma^2 > \mathbb{E}\left[X^2\right]$, there exists some $b \geq 0$ such that $\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\frac{\sigma^2\lambda^2}{2}}$ for all $|\lambda| \leq \frac{1}{b}$
2. (I) $\Rightarrow$ (II) : This implication is immediate.

3. (III) $\Rightarrow$ (II) : For an exponent $a > 0$ and truncation level $T > 0$ to be chosen, we have

$$\mathbb{E}\left[e^{a|X|}\mathbb{D}\left[e^{a|X|} \leq e^{aT}\right]\right] \leq \int_0^{e^{aT}} \mathbb{P}\left[e^{a|X|} \geq t\right] dt \leq 1 + \int_1^{e^{aT}} \mathbb{P}\left[|X| \geq \frac{\log t}{a}\right] dt$$

Applying the assumed tail bound, we obtain

$$\mathbb{E}\left[e^{a|X|}[\left[e^{a|X|} \leq e^{aT}\right]\right] \leq 1 + c_1 \int_1^{e^{aT}} e^{-\frac{c_2 \log t}{a}} dt = 1 + c_1 \int_1^{e^{aT}} t^{-c_2/a} dt$$

Thus, for any $a \in \left[0, \frac{c_2}{2}\right]$, we have

$$\mathbb{E}\left[e^{a|X|}\mathbb{D}\left[e^{a|X|} \leq e^{aT}\right]\right] \leq 1 + \frac{c_1}{2}\left(1 - e^{-aT}\right) \leq 1 + \frac{c_1}{2}$$

By taking the limit as $T \to \infty$, we conclude that $\mathbb{E}\left[e^{a|X|}\right]$ is finite for all $a \in \left[0, \frac{c_2}{2}\right]$. Since both e^{aX} and e^{-aX} are upper bounded by $e^{|a||X|}$, it follows that $\mathbb{E}\left[e^{aX}\right]$ is finite for all $|a| \leq \frac{c_2}{2}$

4. (II) $\Rightarrow$ (III) : By the Chernoff bound with $\lambda = c_0/2$, we have

$$\mathbb{P}[X \geq t] \leq \mathbb{E}\left[e^{\frac{c_0 x}{2}}\right] e^{-\frac{c_0 t}{2}}$$

Applying a similar argument to $-X$, we conclude that $\mathbb{P}[|X| \geq t] \leq c_1 e^{-c_2 t}$ with $c_1 = \mathbb{E}\left[e^{c_0 X/2}\right] + \mathbb{E}\left[e^{-c_0 X/2}\right]$ and $c_2 = c_0/2$

5. (II) $\Leftrightarrow$ (IV) : since the moment generating function exists in an open interval around zero, we can consider the power-series expansion

$$\mathbb{E}\left[e^{\lambda X}\right] = 1 + \sum_{k=2}^{\infty} \frac{\lambda^k \mathbb{E}\left[X^k\right]}{k!} \qquad \text{for all } |\lambda| < a$$

By definition, the quantity $\gamma(X)$ is the radius of convergence of this power series, from which the equivalence between (II) and (IV) follows.

■

Some one-sided results

Theorem 1.0.8 — One-sided Bernstein's inequality. If $X \leq b$ almost surely, the

$$\mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right] \leq \exp\left(\frac{\lambda^2 \mathbb{E}\left[X^2\right]}{1 - \frac{b\lambda}{3}}\right) \qquad \text{for all } \lambda \in [0, 3/b). \tag{1.14}$$

Consequently, given n independent random variables such that $X_i \leq b$ almost surely, we have

$$\mathbb{P}\left[\sum_{i=1}^{n} (X_i - \mathbb{E}[X_i]) \geq n\delta\right] \leq \exp\left(-\frac{n\delta^2}{2\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right] + \frac{b\delta}{3}\right)}\right) \tag{1.15}$$

Of course, if a random variable is bounded from below, then the same result can be used to derive bounds on its lower tail; we simply apply the bound (1.15) to the random variable $-X$. In the special case of independent non-negative random variables $Y_i \geq 0$, we find that

$$\mathbb{P}\left[\sum_{i=1}^{n}(Y_i - \mathbb{E}[Y_i]) \leq -n\delta\right] \leq \exp\left(-\frac{n\delta^2}{\frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left[Y_i^2\right]}\right)$$

Thus, we see that the lower tail of any non-negative random variable satisfies a bound of the sub-Gaussian type, albeit with the second moment instead of the variance.

The proof of Proposition 1.0.8 is quite straightforward given our development thus far.

Proof. Defining the function

$$h(u) := 2\frac{e^u - u - 1}{u^2} = 2\sum_{k=2}^{\infty}\frac{u^{k-2}}{k!}$$

we have the expansion

$$\mathbb{E}\left[e^{\lambda X}\right] = 1 + \lambda\mathbb{E}[X] + \frac{1}{2}\lambda^2\mathbb{E}\left[X^2 h(\lambda X)\right]$$

Observe that for all scalars $x < 0, x' \in [0,b]$ and $\lambda > 0$, we have

$$h(\lambda x) \leq h(0) \leq h\left(\lambda x'\right) \leq h(\lambda b)$$

Consequently, since $X \leq b$ almost surely, we have $\mathbb{E}\left[X^2 h(\lambda X)\right] \leq \mathbb{E}\left[X^2\right] h(\lambda b)$, and hence

$$\begin{aligned}\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] &\leq e^{-\lambda\mathbb{E}[X]}\left\{1 + \lambda\mathbb{E}[X] + \frac{1}{2}\lambda^2\mathbb{E}\left[X^2\right]h(\lambda b)\right\} \\ &\leq^{(i)} \exp\left\{\frac{\lambda^2\mathbb{E}\left[X^2\right]}{2}h(\lambda b)\right\}\end{aligned}$$

where (i) is from $1 + t \leq e^t$. Consequently, the bound (1.14) will follow if we can show that $h(\lambda b) \leq \left(1 - \frac{\lambda b}{3}\right)^{-1}$ for $\lambda b < 3$. By applying the inequality $k! \geq 2\left(3^{k-2}\right)$, valid for all $k \geq 2$, we find that

$$h(\lambda b) = 2\sum_{k=2}^{\infty}\frac{(\lambda b)^{k-2}}{k!} \leq \sum_{k=2}^{\infty}\left(\frac{\lambda b}{3}\right)^{k-2} = \frac{1}{1 - \frac{\lambda b}{3}}$$

where the condition $\frac{\lambda b}{3} \in [0,1)$ allows us to sum the geometric series. In order to prove the upper tail bound (1.15), we apply the Chernoff bound, exploiting independence to apply the moment generating function bound (1.14) separately, and thereby find that

$$\mathbb{P}\left[\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq n\delta\right] \leq \exp\left(-\lambda n\delta + \frac{\frac{\lambda^2}{2}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right]}{1 - \frac{b\lambda}{3}}\right), \quad \text{valid for } b\lambda \in [0,3)$$

Substituting

$$\lambda = \frac{n\delta}{\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right] + \frac{n\delta b}{3}} \in [0,3/b)$$

and simplifying yields the bound. ■

1.0.1 Martingale-based methods

In this section, we describe some of the results briefly in martingale decompositions area along with some examples.

Background

Let $\{X_k\}_{k=1}^n$ be a sequence of independent random variables, and consider the random variable $f(X) = f(X_1 \ldots, X_n)$, for some function $f : \mathbb{R}^n \to \mathbb{R}$. Suppose that our goal is to obtain bounds on the deviations of f from its mean. In order to do so, we consider the sequence of random variables given by $Y_0 = \mathbb{E}[f(X)], Y_n = f(X)$, and

$$Y_k = \mathbb{E}\left[f(X) \mid X_1, \ldots, X_k\right] \quad \text{for } k = 1, \ldots, n-1$$

where we assume that all conditional expectations exist. Note that Y_0 is a constant, and the random variables Y_k will tend to exhibit more fluctuations as we move along the sequence from Y_0 to Y_n. Based on this intuition, the martingale approach to tail bounds is based on the telescoping decomposition

$$f(X) - \mathbb{E}[f(X)] = Y_n - Y_0 = \sum_{k=1}^{n} \underbrace{(Y_k - Y_{k-1})}_{D_k}$$

in which the deviation $f(X) - \mathbb{E}[f(X)]$ is written as a sum of increments $\{D_k\}_{k=1}^n$. As we will see, the sequence $\{Y_k\}_{k=1}^n$ is a particular example of a martingale sequence, known as the **Dobb martingale**, whereas the sequence $\{D_k\}_{k=1}^n$ is an example of a martingale difference sequence.

With this example in mind, we now turn to the general definition of a martingale sequence. Let $\{\mathcal{F}_k\}_{k=1}^\infty$ be a sequence of σ -fields that are nested, meaning that $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$ for all $k \geq 1$ such a sequence is known as a filtration. In the Doob martingale described above, the σ -field $\sigma(X_1, \ldots, X_k)$ generated by the first k variables plays the role of $\mathcal{F}_k$. Let $\{Y_k\}_{k=1}^\infty$ be a sequence of random variables such that Y_k is measurable with respect to the σ -field $\mathcal{F}_k$. In this case, we say that $\{Y_k\}_{k=1}^\infty$ is adapted to the filtration $\{\mathcal{F}_k\}_{k=1}^\infty$. In the Doob martingale, the random variable Y_k is a measurable function of $(X_1, \ldots, X_k)$, and hence the sequence is adapted to the filtration defined by the σ -fields. We are now ready to define a general martingale:

Definition 1.0.3 Given a sequence $\{Y_k\}_{k=1}^\infty$ of random variables adapted to a filtration $\{\mathcal{F}_k\}_{k=1}^\infty$, the pair $\{(Y_k, \mathcal{F}_k)\}_{k=1}^\infty$ is a martingale if, for all $k \geq 1$

$$\mathbb{E}\left[|Y_k|\right] < \infty \quad \text{and} \quad \mathbb{E}\left[Y_{k+1} \mid \mathcal{F}_k\right] = Y_k \tag{1.16}$$

It is frequently the case that the filtration is defined by a second sequence of random variables $\{X_k\}_{k=1}^\infty$ via the canonical σ -fields $\mathcal{F}_k := \sigma(X_1, \ldots, X_k)$. In this case, we say that $\{Y_k\}_{k=1}^\infty$ is a martingale sequence with respect to $\{X_k\}_{k=1}^\infty$. The Doob construction is an instance of such a martingale sequence. If a sequence is martingale with respect to itself (i.e., with $\mathcal{F}_k = \sigma(Y_1, \ldots, Y_k)$), then we say simply that $\{Y_k\}_{k=1}^{\infty}$ forms a martingale sequence.

Let us consider some examples to illustrate:

■ **Example 1.7 — Partial sums as martingales.** Perhaps the simplest instance of a martingale is provided by considering partial sums of an i.i.d. sequence. Let $\{X_k\}_{k=1}^{\infty}$ be a sequence of i.i.d. random variables with mean μ, and define the partial sums $S_k := \sum_{j=1}^{k} X_j$. Defining $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$, the random variable S_k is measurable with respect to $\mathcal{F}_k$, and, moreover, we have

$$\begin{aligned} \mathbb{E}[S_{k+1} \mid \mathcal{F}_k] &= \mathbb{E}[X_{k+1} + S_k \mid X_1, \ldots, X_k] \\ &= \mathbb{E}[X_{k+1}] + S_k \\ &= \mu + S_k \end{aligned}$$

Here we have used the facts that X_{k+1} is independent of $X_1^k := (X_1, \ldots, X_k)$, and that S_k is a function of X_1^k. Thus, while the sequence $\{S_k\}_{k=1}^{\infty}$ itself is not a martingale unless $\mu = 0$, the recentered variables $Y_k := S_k - k\mu$ for $k \geq 1$ define a martingale sequence with respect to $\{X_k\}_{k=1}^{\infty}$

Let us now show that the Doob construction does lead to a martingale, as long as the underlying function f is absolutely integrable.

■ **Example 1.8 — Doob construction.** Given a sequence of independent random variables $\{X_k\}_{k=1}^{n}$, recall the sequence $Y_k = \mathbb{E}[f(X) \mid X_1, \ldots, X_k]$ previously defined, and suppose that $\mathbb{E}[|f(X)|] < \infty$. We claim that $\{Y_k\}_{k=0}^{n}$ is a martingale with respect to $\{X_k\}_{k=1}^{n}$. Indeed, in terms of the shorthand $X_1^k = (X_1, X_2, \ldots, X_k)$, we have

$$\mathbb{E}[|Y_k|] = \mathbb{E}\left[\left|\mathbb{E}\left[f(X) \mid X_1^k\right]\right|\right] \leq \mathbb{E}[|f(X)|] < \infty$$

where the bound follows from Jensen's inequality. Turning to the second property, we have

$$\mathbb{E}\left[Y_{k+1} \mid X_1^k\right] = \mathbb{E}\left[\mathbb{E}\left[f(X) \mid X_1^{k+1}\right] \mid X_1^k\right] \overset{(i)}{=} \mathbb{E}\left[f(X) \mid X_1^k\right] = Y_k$$

where we have used the tower property of conditional expectation in step (i).

The following martingale plays an important role in analyzing stopping rules for sequential hypothesis tests:

■ **Example 1.9 — Likelihood ratio.** Let f and g be two mutually absolutely continuous densities, and let $\{X_k\}_{k=1}^{\infty}$ be a sequence of random variables drawn i.i.d. according to f. For each $k \geq 1$, let $Y_k := \prod_{\ell=1}^{k} \frac{g(X_t)}{f(X_t)}$ be the likelihood ratio based on the first k samples. Then the sequence $\{Y_k\}_{k=1}^{\infty}$ is a martingale with respect to $\{X_k\}_{k=1}^{\infty}$. Indeed, we have

$$\mathbb{E}[Y_{n+1} \mid X_1, \ldots, X_n] = \mathbb{E}\left[\frac{g(X_{n+1})}{f(X_{n+1})}\right] \prod_{k=1}^{n} \frac{g(X_k)}{f(X_k)} = Y_n$$

(the previous n terms are known, which are not random variables any more.)using the fact that $\mathbb{E}\left[\frac{g(X_{n+1})}{f(X_{n+1})}\right] = 1$.

A closely related notion is that of martingale difference sequence, meaning an adapted sequence $\{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ such that, for all $k \geq 1$

$$\mathbb{E}\left[|D_k|\right] < \infty \quad \text{and} \quad \mathbb{E}\left[D_{k+1} \mid \mathcal{F}_k\right] = 0$$

As suggested by their name, such difference sequences arise in a natural way from martingales. In particular, given a martingale $\{(Y_k, \mathcal{F}_k)\}_{k=0}^{\infty}$, let us define $D_k = Y_k - Y_{k-1}$ for $k \geq 1$.

We then have

$$\begin{aligned}\mathbb{E}\left[D_{k+1} \mid \mathcal{F}_k\right] &= \mathbb{E}\left[Y_{k+1} \mid \mathcal{F}_k\right] - \mathbb{E}\left[Y_k \mid \mathcal{F}_k\right] \\ &= \mathbb{E}\left[Y_{k+1} \mid \mathcal{F}_k\right] - Y_k = 0\end{aligned}$$

using the martingale property (1.16) and the fact that Y_k is measurable with respect to $\mathcal{F}_k$ Thus, for any martingale sequence $\{Y_k\}_{k=0}^{\infty}$, we have the telescoping decomposition

$$Y_n - Y_0 = \sum_{k=1}^{n} D_k$$

where $\{D_k\}_{k=1}^{\infty}$ is a martingale difference sequence. This decomposition plays an important role in our development of concentration inequalities to follow.

Loncentration bounds for martingale difference sequences

We now turn to the derivation of concentration inequalities for martingales. These inequalities can be viewed in one of two ways: either as bounds for the difference $Y_n - Y_0$, or as bounds for the sum $\sum_{k=1}^{n} D_k$ of the associated martingale difference sequence. Throughout this section, we present results mainly in terms of martingale differences, with the understanding that such bounds have direct consequences for martingale sequences. Of particular interest to us is the Doob martingale described in Example 2.17, which can be used to control the deviations of a function from its expectation.

We begin by stating and proving a general Bernstein-type bound for a martingale difference sequence, based on imposing a sub-exponential condition on the martingale differences.

Theorem 1.0.9 Let $\{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ be a martingale difference sequence, and suppose that $\mathbb{E}\left[e^{\lambda D_k} \mid \mathcal{F}_{k-1}\right] \leq e^{\lambda^2 v_k^2/2}$ almost surely for any $|\lambda| < 1/\alpha_k$. Then the following hold:

1. The sum $\sum_{k=1}^{n} D_k$ is sub-exponential with parameters $\left(\sqrt{\sum_{k=1}^{n} v_k^2}, \alpha_*\right)$ where $\alpha_* := \max_{k=1,\ldots,n} \alpha_k$

2. The sum satisfies the concentration inequality

$$\mathbb{P}\left[\left|\sum_{k=1}^{n} D_k\right| \geq t\right] \leq \begin{cases} 2e^{-\frac{t^2}{2\sum_{k=1}^{n} v_k^2}} & \text{if } 0 \leq t \leq \frac{\sum_{k=1}^{n} v_k^2}{\alpha_*} \\ 2e^{-\frac{t}{2\alpha_*}} & \text{if } t > \frac{\sum_{k=1}^{n} v_k^2}{\alpha_*} \end{cases} \tag{1.17}$$

Proof. We follow the standard approach of controlling the moment generating function of $\sum_{k=1}^{n} D_k$, and then applying the Chernoff bound. For any scalar λ such that $|\lambda| < \frac{1}{\alpha_*}$, conditioning on $\mathcal{F}_{n-1}$ and applying iterated expectation yields

$$\begin{aligned} \mathbb{E}\left[e^{\lambda\left(\sum_{k=1}^{n} D_k\right)}\right] &= \mathbb{E}\left[e^{\lambda\left(\sum_{k=1}^{n-1} D_k\right)} \mathbb{E}\left[e^{\lambda D_n} \mid \mathcal{F}_{n-1}\right]\right] \\ &\leq \mathbb{E}\left[e^{\lambda \sum_{k=1}^{n-1} D_k}\right] e^{\lambda^2 v_n^2/2} \end{aligned} \tag{1.18}$$

where the inequality follows from the stated assumption on D_n. Iterating this procedure yields the bound $\mathbb{E}\left[e^{\lambda \sum_{k=1}^{n} D_k}\right] \leq e^{\lambda^2 \sum_{k=1}^{n} v_k^2/2}$, valid for all $|\lambda| < \frac{1}{\alpha_*}$. By definition, we conclude that $\sum_{k=1}^{n} D_k$ is sub-exponential with parameters $\left(\sqrt{\sum_{k=1}^{n} v_k^2}, \alpha_*\right)$, as claimed. The tail bound (1.17) follows by applying Proposition 1.0.3. ■

In order for Theorem 1.0.9 to be useful in practice, we need to isolate sufficient and easily checkable conditions for the differences D_k to be almost surely sub-exponential (or sub-Gaussian when $\alpha = 0$). As discussed previously, bounded random variables are subGaussian, which leads to the following corollary:

Corollary 1.0.10 — Azuma-Hoeffding. Let $\left(\{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty}\right)$ be a martingale difference sequence for which there are constants $\{(a_k, b_k)\}_{k=1}^{n}$ such that $D_k \in [a_k, b_k]$ almost surely for all $k = 1, \ldots, n$. Then, for all $t \geq 0$

$$\mathbb{P}\left[\left|\sum_{k=1}^{n} D_k\right| \geq t\right] \leq 2e^{-\frac{2t^2}{\sum_{k=1}^{n}(b_k - a_k)^2}}$$

Proof. Recall the decomposition (1.18) in the proof of Theorem 1.0.9; from the structure of this argument, it suffices to show that $\mathbb{E}\left[e^{\lambda D_k} \mid \mathcal{F}_{k-1}\right] \leq e^{\lambda^2 (b_k - a_k)^2/8}$ almost surely for each $k = 1, 2, \ldots, n$. But since $D_k \in [a_k, b_k]$ almost surely, the conditioned variable $(D_k \mid \mathcal{F}_{k-1})$ also belongs to this interval almost surely, and hence from the result of Exercise 2.4, it is sub-Gaussian with parameter at most $\sigma = (b_k - a_k)/2$ ■

An important application of Corollary 1.0.10 concerns functions that satisfy a bounded difference property. Let us first introduce some convenient notation. Given vectors $x, x' \in \mathbb{R}^n$ and an index $k \in \{1, 2, \ldots, n\}$, we define a new vector $x^{\setminus k} \in \mathbb{R}^n$ via

$$x_j^{\setminus k} := \begin{cases} x_j & \text{if } j \neq k \\ x'_k & \text{if } j = k \end{cases} \tag{1.19}$$

With this notation, we say that $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the bounded difference inequality with parameters $(L_1, \ldots, L_n)$ if, for each index $k = 1, 2, \ldots, n$

$$\left| f(x) - f\left(x^{\backslash k}\right) \right| \leq L_k \quad \text{for all } x, x' \in \mathbb{R}^n \tag{1.20}$$

For instance, if the function f is L -Lipschitz with respect to the Hamming norm $d_H(x,y) = \sum_{i=1}^n \mathbb{I}\left[x_i \neq y_i\right]$, which counts the number of positions in which x and y differ, then the bounded difference inequality holds with parameter L uniformly across all coordinates.

Corollary 1.0.11 — Bounded differences inequality. Suppose that f satisfies the bounded difference property (1.20) with parameters $(L_1, \ldots, L_n)$ and that the random vector $X = (X_1, X_2, \ldots, X_n)$ has independent components. Then

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2e^{-\frac{2^2}{\sum_{k=1}^n L_k^2}} \quad \text{for all } t \geq 0$$

Proof. Recalling the Doob martingale introduced in Example 1.8, consider the associated martingale difference sequence

$$D_k = \mathbb{E}\left[f(X) \mid X_1, \ldots, X_k\right] - \mathbb{E}\left[f(X) \mid X_1, \ldots, X_{k-1}\right]$$

We claim that D_k lies in an interval of length at most L_k almost surely. In order to prove this claim, define the random variables and

$$A_k := \inf_x \mathbb{E}\left[f(X) \mid X_1, \ldots, X_{k-1}, x\right] - \mathbb{E}\left[f(X) \mid X_1, \ldots, X_{k-1}\right]$$

$$B_k := \sup_x \mathbb{E}\left[f(X) \mid X_1, \ldots, X_{k-1}, x\right] - \mathbb{E}\left[f(X) \mid X_1, \ldots, X_{k-1}\right]$$

On one hand, we have

$$D_k - A_k = \mathbb{E}\left[f(X) \mid X_1, \ldots, X_k\right] - \inf_x \mathbb{E}\left[f(X) \mid X_1, \ldots, X_{k-1}, x\right]$$

so that $D_k \geq A_k$ almost surely. A similar argument shows that $D_k \leq B_k$ almost surely. We now need to show that $B_k - A_k \leq L_k$ almost surely. Observe that by the independence of $\{X_k\}_{k=1}^n$, we have

$$\mathbb{E}\left[f(X) \mid x_1, \ldots, x_k\right] = \mathbb{E}_{k+1}\left[f\left(x_1, \ldots, x_k, X_{k+1}^n\right)\right] \quad \text{for any vector } (x_1, \ldots, x_k)$$

where $\mathbb{E}_{k+1}$ denotes expectation over $X_{k+1}^n := (X_{k+1}, \ldots, X_n)$. Consequently, we have

$$\begin{aligned} B_k - A_k &= \sup_x \mathbb{E}_{k+1}\left[f\left(X_1, \ldots, X_{k-1}, x, X_{k+1}^n\right]\right. - \inf_x \mathbb{E}_{k+1}\left[f\left(X_1, \ldots, X_{k-1}, x, X_{k+1}^n\right]\right. \\ &\leq \sup_{x,y} \left|\mathbb{E}_{k+1}\left[f\left(X_1, \ldots, X_{k-1}, x, X_{k+1}^n\right) - f\left(X_1, \ldots, X_{k-1}, y, X_{k+1}^n\right)\right]\right| \\ &\leq L_k \end{aligned}$$

using the bounded differences assumption. Thus, the variable D_k lies within an interval of length L_k at most surely, so that the claim follows as a corollary of the Azuma-Hoeffding inequality. ■

R In the special case when f is L -Lipschitz with respect to the Hamming norm, Corollary 1.0.11 implies that

$$\mathbb{P}[|f(X)-\mathbb{E}[f(X)]|\geq t]\leq 2e^{-\frac{\lambda^2}{nL^2}} \quad \text{for all } t\geq 0 \tag{1.21}$$

Let us consider some examples to illustrate.

■ **Example 1.10 — Classical Hoeffding from bounded differences.** As a warm-up, let us show how the classical Hoeffding bound (2.11) for bounded variables - say $X_i \in [a,b]$ almost surely-follows as an immediate corollary of the bound (1.21). Consider the function $f\left(x_1,\ldots,x_n\right)=\sum_{i=1}^n\left(x_i-\mu_i\right)$, where $\mu_i=\mathbb{E}\left[X_i\right]$ is the mean of the i th random variable. For any index $k\in\{1,\ldots,n\}$, we have

$$\begin{aligned}\left|f(x)-f\left(x^{\backslash k}\right)\right| &= \left|\left(x_k-\mu_k\right)-\left(x_k'-\mu_k\right)\right| \\ &= \left|x_k-x_k'\right|\leq b-a\end{aligned}$$

showing that f satisfies the bounded difference inequality in each coordinate with parameter $L=b-a$. Consequently, it follows from the bounded difference inequality (1.21) that

$$\mathbb{P}\left[\left|\sum_{i=1}^n\left(X_i-\mu_i\right)\right|\geq t\right]\leq 2e^{-\frac{2t^2}{n(b-a)^2}}$$

which is the classical Hoeffding bound for independent random variables.

■ **Example 1.11 — U -statistics.** Let $g:\mathbb{R}^2\to\mathbb{R}$ be a symmetric function of its arguments. Given an i.i.d. sequence $X_k, k\geq 1$, of random variables, the quantity

$$U:=\frac{1}{\left(\begin{array}{l}n\\2\end{array}\right)}\sum_{j<k}g\left(X_j,X_k\right)$$

is known as a pairwise U -statistic. For instance, if $g(s,t)=|s-t|$, then U is an unbiased estimator of the mean absolute pairwise deviation $\mathbb{E}\left[\left|X_1-X_2\right|\right]$. Note that, while U is not a sum of independent random variables, the dependence is relatively weak, and this fact can be revealed by a martingale analysis.

If g is bounded (say $\|g\|_\infty\leq b$), then Corollary 1.0.11 can be used to establish the concentration of U around its mean. Viewing U as a function $f(x)=f\left(x_1,\ldots,x_n\right)$, for any given coordinate k, we have

$$\begin{aligned}\left|f(x)-f\left(x^{\backslash k}\right)\right| &\leq \frac{1}{\left(\begin{array}{l}n\\2\end{array}\right)}\sum_{j\neq k}\left|g\left(x_j,x_k\right)-g\left(x_j,x_k'\right)\right| \\ &\leq \frac{(n-1)(2b)}{\left(\begin{array}{l}n\\2\end{array}\right)}=\frac{4b}{n}\end{aligned}$$

so that the bounded differences property holds with parameter $L_k = \frac{4b}{n}$ in each coordinate. Thus, we conclude that

$$\mathbb{P}[|U - \mathbb{E}[U]| \geq t] \leq 2e^{-\frac{n^2}{8b^2}}$$

This tail inequality implies that U is a consistent estimate of $\mathbb{E}[U]$, and also yields finite sample bounds on its quality as an estimator. Similar techniques can be used to obtain tail bounds on U -statistics of higher order, involving sums over k -tuples of variables.

Martingales and the bounded difference property also play an important role in analyzing the properties of random graphs, and other random combinatorial structures.

■ **Example 1.12 — Clique number in random graphs.** An undirected graph is a pair $G = (V, E)$, composed of a vertex set $V = \{1, \ldots, d\}$ and an edge set E, where each edge $e = (i, j)$ is an unordered pair of distinct vertices $(i \neq j)$. A graph clique C is a subset of vertices such that $(i, j) \in E$ for all $i, j \in C$. The clique number $C(G)$ of the graph is the cardinality of the largest clique-note that $C(G) \in [1, d]$. When the edges E of the graph are drawn according to some random process, then the clique number $C(G)$ is a random variable, and we can study its concentration around its mean $\mathbb{E}[C(G)]$.

The Erdös-Rényi ensemble of random graphs is one of the most well-studied models: it is defined by a parameter $p \in (0,1)$ that specifies the probability with which each edge (i, j) is included in the graph, independently across all $\left(\begin{array}{c} d \\ 2 \end{array} \right)$ edges. More formally, for each $i < j$, let us introduce a Bernoulli edge-indicator variable X_{ij} with parameter p, where $X_{ij} = 1$ means that edge (i, j) is included in the graph, and $X_{ij} = 0$ means that it is not included.

Note that the $\left(\begin{array}{c} d \\ 2 \end{array} \right)$ -dimensional random vector $Z := \left\{ X_{ij} \right\}_{i<j}$ specifies the edge set; thus, we may view the clique number $C(G)$ as a function $Z \mapsto f(Z)$. Let Z' denote a vector in which a single coordinate of Z has been changed, and let G' and G be the associated graphs. It is easy to see that $C(G')$ can differ from $C(G)$ by at most 1, so that $|f(Z') - f(Z)| \leq 1$. Thus, the function $C(G) = f(Z)$ satisfies the bounded difference property in each coordinate with parameter $L = 1$, so that

$$\mathbb{P}\left[\frac{1}{n}|C(G) - \mathbb{E}[C(G)]| \geq \delta\right] \leq 2e^{-2n\delta^2}$$

Consequently, we see that the clique number of an Erdös-Rényi random graph is very sharply concentrated around its expectation.

Definition 1.0.4 — Rademacher complexity. Let $\{\varepsilon_k\}_{k=1}^{n}$ be an i.i.d. sequence of Rademacher variables (i.e., taking the values {-1,+1} equiprobably, as in Example 1.2). Given a collection of vectors $\mathcal{A} \subset \mathbb{R}^n$, define the random variable

$$Z := \sup_{a \in \mathcal{A}} \left[\sum_{k=1}^{n} a_k \varepsilon_k \right] = \sup_{a \in \mathcal{A}} [\langle a, \varepsilon \rangle]$$

The random variable Z measures the size of $\mathcal{A}$ in a certain sense, and its expectation $\mathcal{R}(\mathcal{A}) := \mathbb{E}[Z(\mathcal{A})]$ is known as the Rademacher complexity of the set $\mathcal{A}$.

■ **Example 1.13 — Rademacher complexity.** Let us now show how Corollary 1.0.11 can be used to establish that $Z(\mathcal{A})$ is sub-Gaussian. Viewing $Z(\mathcal{A})$ as a function $(\varepsilon_1, \dots, \varepsilon_n) \mapsto f(\varepsilon_1, \dots, \varepsilon_n) = \sup_{a \in \mathcal{A}} [\langle a, \varepsilon \rangle]$, we need to bound the maximum change when coordinate k is changed. Given two Rademacher vectors $\varepsilon, \varepsilon' \in \{-1, +1\}^n$, recall our definition (1.19) of the modified vector $\varepsilon^{\setminus k}$. Since $f\left(\varepsilon^{\setminus k}\right) \geq \left\langle a, \varepsilon^{\setminus k} \right\rangle$ for any $a \in \mathcal{A}$, we have

$$\langle a, \varepsilon \rangle - f\left(\varepsilon^{\setminus k}\right) \leq \left\langle a, \varepsilon - \varepsilon^{\setminus k} \right\rangle = a_k \left(\varepsilon_k - \varepsilon'_k\right) \leq 2 |a_k|$$

Taking the supremum over $\mathcal{A}$ on both sides, we obtain the inequality

$$f(\varepsilon) - f\left(\varepsilon^{\setminus k}\right) \leq 2 \sup_{a \in \mathcal{A}} |a_k|$$

Since the same argument applies with the roles of ε and $\varepsilon^{\setminus k}$ reversed, we conclude that f satisfies the bounded difference inequality in coordinate k with parameter $2 \sup_{a \in \mathcal{A}} |a_k|$. Consequently, Corollary 1.0.11 implies that the random variable $Z(\mathcal{A})$ is sub-Gaussian with parameter at most $2\sqrt{\sum_{k=1}^{n} \sup_{a \in \mathcal{A}} a_k^2}$. This sub-Gaussian parameter can be reduced to the (potentially much) smaller quantity $\sqrt{\sup_{a \in \mathcal{A}} \sum_{k=1}^{n} a_k^2}$ using alternative techniques; in particular, see Example 3.5 in Chapter 3 for further details.

1.0.2 Lipschitz functions of Gaussian variables

We conclude this chapter with a classical result on the concentration properties of Lipschitz functions of Gaussian variables. These functions exhibit a particularly attractive form of dimension-free concentration.

Definition 1.0.5 — L-Lipschitz. Let us say that a function $f : \mathbb{R}^n \to \mathbb{R}$ is L -Lipschitz with respect to the Euclidean norm $\|\cdot\|_2$ if

$$|f(x) - f(y)| \leq L \|x - y\|_2 \quad \text{for all } x, y \in \mathbb{R}^n \tag{1.22}$$

The following result guarantees that any such function is sub-Gaussian with parameter at most L :

Theorem 1.0.12 Let $(X_1, \dots, X_n)$ be a vector of i.i.d. standard Gaussian variables, and let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz with respect to the Euclidean norm. Then the variable

$f(X) - \mathbb{E}[f(X)]$ is sub-Gaussian with parameter at most L, and hence

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2e^{-\frac{t^2}{2L^2}} \quad \text{for all } t \geq 0 \tag{1.23}$$

Note that this result is truly remarkable: it guarantees that any L -Lipschitz function of a standard Gaussian random vector, regardless of the dimension, exhibits concentration like a scalar Gaussian variable with variance L^2.

With the aim of keeping the proof as simple as possible, let us prove a version of the concentration bound (1.23) with a weaker constant in the exponent. (See the bibliographic notes for references to proofs of the sharpest results.) We also prove the result for a function that is both Lipschitz and differentiable; since any Lipschitz function is differentiable almost everywhere, it is then straightforward to extend this result to the general setting. For a differentiable function, the Lipschitz property guarantees that $\|\nabla f(x)\|_2 \leq L$ for all $x \in \mathbb{R}^n$. In order to prove this version of the theorem, we begin by stating an auxiliary technical lemma:

Lemma 1.1 Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable. Then for any convex function $\phi : \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{E}[\phi(f(X) - \mathbb{E}[f(X)])] \leq \mathbb{E}\left[\phi\left(\frac{\pi}{2}\langle \nabla f(X), Y\rangle\right)\right] \tag{1.24}$$

where $X, Y \sim \mathcal{N}(0, \mathbf{I}_n)$ are standard multivariate Gaussian, and independent.

Proof. We now prove the theorem using this lemma. For any fixed $\lambda \in \mathbb{R}$, applying inequality (1.24) to the convex function $t \mapsto e^{\lambda t}$ yields

$$\begin{aligned}\mathbb{E}_X[\exp(\lambda\{f(X) - \mathbb{E}[f(X)]\})] &\leq \mathbb{E}_{X,Y}\left[\exp\left(\frac{\lambda\pi}{2}\langle Y, \nabla f(X)\rangle\right)\right] \\ &= \mathbb{E}_X\left[\exp\left(\frac{\lambda^2\pi^2}{8}\|\nabla f(X)\|_2^2\right)\right]\end{aligned}$$

where we have used the independence of X and Y to first take the expectation over Y marginally, and the fact that $\langle Y, \nabla f(x)\rangle$ is a zero-mean Gaussian variable with variance $\|\nabla f(x)\|_2^2$. Due to the Lipschitz condition on f, we have $\|\nabla f(x)\|_2 \leq L$ for all $x \in \mathbb{R}^n$, whence

$$\mathbb{E}[\exp(\lambda\{f(X) - \mathbb{E}[f(X)]\})] \leq e^{\frac{1}{8}\lambda^2\pi^2L^2}$$

which shows that $f(X) - \mathbb{E}[f(X)]$ is sub-Gaussian with parameter at most $\frac{\pi L}{2}$. The tail bound

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2\exp\left(-\frac{2t^2}{\pi^2L^2}\right) \quad \text{for all } t \geq 0$$

follows from Proposition 1.0.1. ■

Proof. It remains to prove Lemma 1.1, and we do so via a classical interpolation method that exploits the rotation invariance of the Gaussian distribution. For each $\theta \in [0, \pi/2]$, consider the random vector $Z(\theta) \in \mathbb{R}^n$ with components

$$Z_k(\theta) := X_k \sin\theta + Y_k \cos\theta \quad \text{for } k = 1, 2, \ldots, n$$

By the convexity of ϕ, we have

$$\mathbb{E}_X\left[\phi\left(f(X) - \mathbb{E}_Y[f(Y)]\right)\right] \leq \mathbb{E}_{X,Y}[\phi(f(X) - f(Y))] \tag{1.25}$$

Now since $Z_k(0) = Y_k$ and $Z_k(\pi/2) = X_k$ for all $k = 1, \ldots, n$, we have

$$f(X) - f(Y) = \int_0^{\pi/2} \frac{d}{d\theta} f(Z(\theta)) d\theta = \int_0^{\pi/2} \langle \nabla f(Z(\theta)), Z'(\theta) \rangle \, d\theta \tag{1.26}$$

where $Z'(\theta) \in \mathbb{R}^n$ denotes the elementwise derivative, a vector with the components $Z_k'(\theta) = X_k \cos\theta - Y_k \sin\theta$. Substituting the integral representation (1.26) into our earlier bound (1.25) yields

$$\begin{aligned}\mathbb{E}_X\left[\phi\left(f(X) - \mathbb{E}_Y[f(Y)]\right)\right] &\leq \mathbb{E}_{X,Y}\left[\phi\left(\int_0^{\pi/2} \langle \nabla f(Z(\theta)), Z'(\theta) \rangle \, d\theta\right)\right] \\ &= \mathbb{E}_{X,Y}\left[\phi\left(\frac{1}{\pi/2}\int_0^{\pi/2} \frac{\pi}{2} \langle \nabla f(Z(\theta)), Z'(\theta) \rangle \, d\theta\right)\right] \\ &\leq \frac{1}{\pi/2}\int_0^{\pi/2} \mathbb{E}_{X,Y}\left[\phi\left(\frac{\pi}{2} \langle \nabla f(Z(\theta)), Z'(\theta) \rangle\right)\right] d\theta\end{aligned} \tag{1.27}$$

where the final step again uses convexity of ϕ (just taking the integral out of expectation). By the rotation invariance of the Gaussian distribution, for each $\theta \in [0, \pi/2]$, the pair $\left(Z_k(\theta), Z_k'(\theta)\right)$ is a jointly Gaussian vector, with zero mean and identity covariance $\mathbf{I}_2$. Therefore, the expectation inside the integral in equation (1.27) does not depend on θ, and hence

$$\frac{1}{\pi/2}\int_0^{\pi/2} \mathbb{E}_{X,Y}\left[\phi\left(\frac{\pi}{2} \langle \nabla f(Z(\theta)), Z'(\theta) \rangle\right)\right] d\theta = \mathbb{E}\left[\phi\left(\frac{\pi}{2} \langle \nabla f(\widetilde{X}), \widetilde{Y} \rangle\right)\right]$$

where $(\widetilde{X}, \widetilde{Y})$ are independent standard Gaussian n -vectors. This completes the proof of the bound (1.24) ■

Note that the proof makes essential use of various properties specific to the standard Gaussian distribution. However, similar concentration results hold for other non-Gaussian distributions, including the uniform distribution on the sphere and any strictly log-concave distribution (see Chapter 3 for further discussion of such distributions). However, without additional structure of the function $f($ such as convexity), dimension-free concentration for Lipschitz functions need not hold for an arbitrary sub-Gaussian distribution; see the bibliographic section for further discussion of this fact.

Theorem 1.0.12 is useful for a broad range of problems; let us consider some examples to illustrate.

■ **Example 1.14 — χ^2 concentration.** For a given sequence $\{Z_k\}_{k=1}^n$ of i.i.d. standard normal variates, the random variable $Y := \sum_{k=1}^n Z_k^2$ follows a χ^2 -distribution with n degrees of freedom. The most direct way to obtain tail bounds on Y is by noting that Z_k^2 is sub-exponential, and exploiting independence (see Example 1.5). In this example, we pursue an alternative approach - namely, via concentration for Lipschitz functions of Gaussian variates. Indeed, defining the variable $V = \sqrt{Y}/\sqrt{n}$, we can write $V = \|(Z_1, \ldots, Z_n)\|_2 / \sqrt{n}$, and since the Euclidean norm is a 1 -Lipschitz function, Theorem 1.0.12 implies that

$$\mathbb{P}[V \geq \mathbb{E}[V] + \delta] \leq e^{-n\delta^2/2} \quad \text{for all } \delta \geq 0$$

Using concavity of the square-root function and Jensen's inequality, we have

$$\mathbb{E}[V] \leq \sqrt{\mathbb{E}\left[V^2\right]} = \left\{\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[Z_k^2\right]\right\}^{1/2} = 1$$

Recalling that $V = \sqrt{Y}/\sqrt{n}$ and putting together the pieces yields

$$\mathbb{P}\left[Y/n \geq (1+\delta)^2\right] \leq e^{-n\delta^2/2} \quad \text{for all } \delta \geq 0$$

Since $(1+\delta)^2 = 1 + 2\delta + \delta^2 \leq 1 + 3\delta$ for all $\delta \in [0,1]$, we conclude that

$$\mathbb{P}[Y \geq n(1+t)] \leq e^{-nt^2/18} \quad \text{for all } t \in [0,3]$$

where we have made the substitution $t = 3\delta$. It is worthwhile comparing this tail bound to those that can be obtained by using the fact that each Z_k^2 is sub-exponential, as discussed in Example 1.5.

■ **Example 1.15 — Order statistics.** Given a random vector $(X_1, X_2, \ldots, X_n)$, its order statistics are obtained by reordering its entries in a non-decreasing manner-namely as

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n-1)} \leq X_{(n)}$$

As particular cases, we have $X_{(n)} = \max_{k=1,\ldots,n} X_k$ and $X_{(1)} = \min_{k=1,\ldots,n} X_k$. Given another random vector $(Y_1, \ldots, Y_n)$, it can be shown that $\left|X_{(k)} - Y_{(k)}\right| \leq \|X - Y\|_2$ for all $k = 1, \ldots, n$ so that each order statistic is a 1 -Lipschitz function. (We leave the verification of this inequality as an exercise for the reader.) Consequently, when X is a Gaussian random vector, Theorem 1.0.12 implies that

$$\mathbb{P}\left[\left|X_{(k)} - \mathbb{E}\left[X_{(k)}\right]\right| \geq \delta\right] \leq 2e^{-\frac{\delta^2}{2}} \quad \text{for all } \delta \geq 0$$

■ **Example 1.16 — Gaussian complexity.** This example is closely related to our earlier discussion of Rademacher complexity in Example 1.13. Let $\{W_k\}_{k=1}^n$ be an i.i.d. sequence of $\mathcal{N}(0,1)$ variables. Given a collection of vectors $\mathcal{A} \subset \mathbb{R}^n$, define the random variable

$$Z := \sup_{a \in \mathcal{A}} \left[\sum_{k=1}^n a_k W_k\right] = \sup_{a \in \mathcal{A}} \langle a, W \rangle$$

As with the Rademacher complexity, the variable $Z = Z(\mathcal{A})$ is one way of measuring the size of the set $\mathcal{A}$, and will play an important role in later chapters. Viewing Z as a function $(w_1, \ldots, w_n) \mapsto f(w_1, \ldots, w_n)$, let us verify that f is Lipschitz (with respect to Euclidean norm) with parameter $\sup_{a \in \mathcal{A}} \|a\|_2$. Let $w, w' \in \mathbb{R}^n$ be arbitrary, and let $a^* \in \mathcal{A}$ be any vector that achieves the maximum defining $f(w)$. Following the same argument as Example 1.13, we have the upper bound

$$f(w) - f(w') \leq \langle a^*, w - w' \rangle \leq D(\mathcal{A}) \left\| w - w' \right\|_2$$

where $D(\mathcal{A}) = \sup_{a \in \mathcal{A}} \|a\|_2$ is the Euclidean width of the set. The same argument holds with the roles of w and w' reversed, and hence

$$\left| f(w) - f(w') \right| \leq D(\mathcal{A}) \left\| w - w' \right\|_2$$

Consequently, Theorem 1.0.12 implies that

$$\mathbb{P}[|Z - \mathbb{E}[Z]| \geq \delta] \leq 2 \exp\left(-\frac{\delta^2}{2D^2(\mathcal{A})} \right)$$

■ **Example 1.17 — Gaussian chaos variables.** As a generalization of the previous example, let $\mathbf{Q} \in \mathbb{R}^{n \times n}$ be a symmetric matrix, and let $w, \widetilde{w}$ be independent zero-mean Gaussian random vectors with covariance matrix $\mathbf{I}_n$. The random variable

$$Z := \sum_{i,j=1}^{n} Q_{ij} w_i \widetilde{w}_j = w^{\mathrm{T}} \mathbf{Q} \widetilde{w}$$

is known as a (decoupled) Gaussian chaos. By the independence of w and $\widetilde{w}$, we have $\mathbb{E}[Z] = 0$, so it is natural to seek a tail bound on Z.

Conditioned on $\widetilde{w}$, the variable Z is a zero-mean Gaussian variable with variance $\|\mathbf{Q}\widetilde{w}\|_2^2 = \widetilde{w}^{\mathrm{T}} \mathbf{Q}^2 \widetilde{w}$, whence

$$\mathbb{P}[|Z| \geq \delta \mid \widetilde{w}] \leq 2e^{-\frac{\delta^2}{\|2Q\tilde{w}\|_2^2}}$$

Let us now control the random variable $Y := \|Q\widetilde{w}\|_2$. Viewed as a function of the Gaussian vector $\widetilde{w}$, it is Lipschitz with constant

$$\|\mathbf{Q}\|_2 := \sup_{\|u\|_2 = 1} \|\mathbf{Q}u\|_2$$

corresponding to the ℓ_2 -operator norm of the matrix $\mathbf{Q}$. Moreover, by Jensen's inequality, we have $\mathbb{E}[Y] \leq \sqrt{\mathbb{E}\left[\widetilde{w}^{\mathrm{T}} \mathbf{Q}^2 \widetilde{w}\right]} = \|\mathbf{Q}\|_{\mathrm{F}}$, where

$$\|\mathbf{Q}\|_F := \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} Q_{ij}^2}$$

is the Frobenius norm of the matrix Q. Putting together the pieces yields the tail bound

$$\mathbb{P}\left[\|\mathbf{Q}\widetilde{w}\|_2 \geq \|\mathbf{Q}\|_F + t\right] \leq 2\exp\left(-\frac{t^2}{2\|\mathbf{Q}\|_2^2}\right)$$

Note that $(\|\mathbf{Q}\|_\mathrm{F} + t)^2 \leq 2\|\mathbf{Q}\|_\mathrm{F}^2 + 2t^2$. Consequently, setting $t^2 = \delta\|\mathbf{Q}\|_2$ and simplifying yields

$$\mathbb{P}\left[\widetilde{w}^T\mathbf{Q}^2\widetilde{w} \geq 2\|\mathbf{Q}\|_\mathrm{F}^2 + 2\delta\|\mathbf{Q}\|_2\right] \leq 2\exp\left(-\frac{\delta}{2\|\mathbf{Q}\|_2}\right)$$

Putting together the pieces, we find that

$$\begin{aligned}\mathbb{P}[|Z| \geq \delta] &\leq 2\exp\left(-\frac{\delta^2}{4\|\mathbf{Q}\|_\mathrm{F}^2 + 4\delta\|\mathbf{Q}\|_2}\right) + 2\exp\left(-\frac{\delta}{2\|\mathbf{Q}\|_2}\right)\\ &\leq 4\exp\left(-\frac{\delta^2}{4\|\mathbf{Q}\|_\mathrm{F}^2 + 4\delta\|\mathbf{Q}\|_2}\right)\end{aligned}$$

We have thus shown that the Gaussian chaos variable satisfies a sub-exponential tail bound.

■ **Example 1.18 — Singular values of Gaussian random matrices.** For integers $n > d$, let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a random matrix with i.i.d. $\mathcal{N}(0,1)$ entries, and let

$$\sigma_1(\mathbf{X}) \geq \sigma_2(\mathbf{X}) \geq \cdots \geq \sigma_d(\mathbf{X}) \geq 0$$

denote its ordered singular values. By Weyl's theorem (see Exercise 8.3), given another matrix $\mathbf{Y} \in \mathbb{R}^{n\times d}$, we have

$$\max_{k=1,\ldots,d} |\sigma_k(\mathbf{X}) - \sigma_k(\mathbf{Y})| \leq \|\mathbf{X}-\mathbf{Y}\|_2 \leq \|\mathbf{X}-\mathbf{Y}\|_\mathrm{F} \tag{1.28}$$

where $\|\cdot\|_\mathrm{F}$ denotes the Frobenius norm. The inequality (1.28) shows that each singular value $\sigma_k(\mathbf{X})$ is a 1 -Lipschitz function of the random matrix, so that Theorem 1.0.12 implies that, for each $k = 1,\ldots,d$, we have

$$\mathbb{P}\left[|\sigma_k(\mathbf{X}) - \mathbb{E}\left[\sigma_k(\mathbf{X})\right]| \geq \delta\right] \leq 2e^{-\frac{\delta^2}{2}} \quad \text{for all } \delta \geq 0$$

Consequently, even though our techniques are not yet powerful enough to characterize the expected value of these random singular values, we are guaranteed that the expectations are representative of the typical behavior. See Chapter 6 for a more detailed discussion of the singular values of random matrices.

2. Concentration of measure

We begin in Section 3.1 with a discussion of the entropy method for concentration, and illustrate its use in deriving tail bounds for Lipschitz functions of independent random variables. In Section 3.2, we turn to some geometric aspects of concentration inequalities, a viewpoint that is historically among the oldest. Section 3.3 is devoted to the use of transportation cost inequalities for deriving concentration inequalities, a method that is in some sense dual to the entropy method, and well suited to certain types of dependent random variables. We conclude in Section 3.4 by deriving some tail bounds for empirical processes, including versions of the functional Hoeffding and Bernstein inequalities. These inequalities play an especially important role in our later treatment of nonparametric problems.

2.0.1 Concentration by entropic techniques

Entropy and its properties

Given a convex function $\phi : \mathbb{R} \to \mathbb{R}$, it can be used to define a functional on the space of probability distributions via

$$\mathbb{H}_\phi(X) := \mathbb{E}[\phi(X)] - \phi(\mathbb{E}[X])$$

where $X \sim \mathbb{P}$. This quantity, which is well defined for any random variable such that both X and $\phi(X)$ have finite expectations, is known as the ϕ -entropy of the random variable X. By Jensen's inequality and the convexity of ϕ, the ϕ -entropy is always non-negative.

As the name suggests, it serves as a measure of variability. For instance, in the most extreme case, we have $\mathbb{H}_\phi(X) = 0$ for any random variable such that X is equal to its expectation $\mathbb{P}$ -almosteverywhere.

There are various types of entropies, depending on the choice of the underlying convex function ϕ. For example, the convex function $\phi(u) = u^2$ yields

$$\mathbb{H}_{\phi}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = \text{var}(X)$$

corresponding to the usual variance of the random variable X.

Another interesting choice is the convex function $\phi(u) = -\log u$ defined on the positive real line. When applied to the positive random variable $Z := e^{\lambda X}$, this choice of ϕ yields

$$\mathbb{H}_{\phi}\left(e^{\lambda X}\right) = -\lambda\mathbb{E}[X] + \log\mathbb{E}\left[e^{\lambda X}\right] = \log\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right]$$

a type of entropy corresponding to the centered cumulant generating function.

Throughout the remainder of this chapter, we focus on a slightly different choice of entropy functional, namely the convex function $\phi : [0,\infty) \to \mathbb{R}$ defined as

$$\phi(u) := u\log u \quad \text{for } u > 0, \quad \text{and} \quad \phi(0) := 0 \tag{2.1}$$

For any non-negative random variable $Z \geq 0$, it defines the ϕ -entropy given by

$$\mathbb{H}(Z) = \mathbb{E}[Z\log Z] - \mathbb{E}[Z]\log\mathbb{E}[Z] \tag{2.2}$$

assuming that all relevant expectations exist. In the remainder of this chapter, we omit the subscript ϕ, since the choice (2.1) is to be implicitly understood.

The reader familiar with information theory may observe that the entropy (2.2) is closely related to the **Shannon entropy**, as well as the **Kullback-Leibler divergence**; see Exercise 3.1 for an exploration of this connection. As will be clarified in the sequel, the most attractive property of the ϕ -entropy (2.2) is its so-called tensorization when applied to functions of independent random variables.

For the random variable $Z := e^{\lambda X}$, the entropy has an explicit expression as a function of the moment generating function $\varphi_{\mathrm{x}}(\lambda) = \mathbb{E}\left[e^{\lambda X}\right]$ (variable is λ) and its first derivative. In particular, a short calculation (2.2) yields

$$\mathbb{H}\left(e^{\lambda X}\right) = \lambda\varphi_{\mathrm{x}}'(\lambda) - \varphi_{\mathrm{x}}(\lambda)\log\varphi_{\mathrm{x}}(\lambda) \tag{2.3}$$

Consequently, if we know the moment generating function of X, then it is straightforward to compute the entropy $\mathbb{H}\left(e^{\lambda X}\right)$. Let us consider a simple example to illustrate:

■ **Example 2.1 — Entropy of a Gaussian random variable.** For the scalar Gaussian variable $X \sim \mathcal{N}\left(0,\sigma^2\right)$, we have $\varphi_{\mathrm{x}}(\lambda) = e^{\lambda^2\sigma^2/2}$. By taking derivatives, we find that $\varphi_{\mathrm{x}}'(\lambda) = \lambda\sigma^2\varphi_{\mathrm{x}}(\lambda)$ and hence

$$\mathbb{H}\left(e^{\lambda X}\right) = \lambda^2\sigma^2\varphi_{\mathbf{x}}(\lambda) - \frac{1}{2}\lambda^2\sigma^2\varphi_{\mathbf{x}}(\lambda) = \frac{1}{2}\lambda^2\sigma^2\varphi_{\mathbf{x}}(\lambda)$$

First, we find the connection between the entropy and tail bounds. We then show how the entropy based on $\phi(u) = u \log u$ has a certain tensorization property that makes it particularly well suited to dealing with general Lipschitz functions of collections of random variables.

Herbst argument and its extensions

Suppose that there is a constant $\sigma > 0$ such that the entropy of $e^{\lambda X}$ satisfies an upper bound of the form

$$\mathbb{H}\left(e^{\lambda X}\right) \leq \frac{1}{2}\sigma^2\lambda^2\varphi_{\mathrm{x}}(\lambda) \tag{2.4}$$

Note that by our earlier calculation in Example 2.1, any Gaussian variable $X \sim \mathcal{N}\left(0,\sigma^2\right)$ satisfies this condition with equality for all $\lambda \in \mathbb{R}$. Moreover, as shown in Exercise 3.7, any bounded random variable satisfies an inequality of the form (2.4).

What does the entropy bound (2.4) imply about the tail behavior of the random variable? Answer: Any such variable must have sub-Gaussian tail behavior.

Proposition 2.0.1 — Herbst argument. Suppose that the entropy $\mathbb{H}\left(e^{\lambda X}\right)$ satisfies inequality (2.4) for all $\lambda \in I$, where I can be either of the intervals $[0,\infty)$ or $\mathbb{R}$. Then X satisfies the bound

$$\log \mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq \frac{1}{2}\lambda^2\sigma^2 \quad \text{for all } \lambda \in I \tag{2.5}$$

When $I = \mathbb{R}$, then the inequality (2.5) is equivalent to asserting that the centered variable $X - \mathbb{E}[X]$ is sub-Gaussian with parameter σ. Via an application of the usual Chernoff argument, the bound (2.5) with $I = [0,\infty)$ implies the one-sided tail bound

$$\mathbb{P}[X \geq \mathbb{E}[X] + t] \leq e^{-\frac{t^2}{2\sigma^2}}$$

and with $I = \mathbb{R}$, it implies the two-sided bound $\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq 2e^{-\frac{t^2}{2\sigma^2}}$. Of course, these are the familiar tail bounds for sub-Gaussian variables discussed previously in Chapter 2 .

Proof. Recall the representation (2.3) of entropy in terms of the moment generating function. Combined with the assumed upper bound (2.4), we conclude that the moment generating function $\varphi \equiv \varphi_{\mathrm{x}}$ satisfies the differential inequality

$$\lambda\varphi'(\lambda) - \varphi(\lambda)\log\varphi(\lambda) \leq \frac{1}{2}\sigma^2\lambda^2\varphi(\lambda), \quad \text{valid for all } \lambda \geq 0 \tag{2.6}$$

Define the function $G(\lambda) = \frac{1}{\lambda}\log\varphi(\lambda)$ for $\lambda \neq 0$ ($\varphi_{\mathrm{x}}(\lambda) = \mathbb{E}\left[e^{\lambda X}\right]$), and extend the definition by continuity to

$$G(0) := \lim_{\lambda \to 0} G(\lambda) = \mathbb{E}[X] \tag{2.7}$$

Note that we have $G'(\lambda) = \frac{1}{\lambda}\frac{\varphi'(\lambda)}{\varphi(\lambda)} - \frac{1}{\lambda^2}\log\varphi(\lambda)$, so that the inequality (2.6) can be rewritten in the simple form $G'(\lambda) \leq \frac{1}{2}\sigma^2$ for all $\lambda \in I$. For any $\lambda_0 > 0$, we can integrate both sides of the inequality to obtain

$$G(\lambda) - G(\lambda_0) \leq \frac{1}{2}\sigma^2(\lambda - \lambda_0)$$

Letting $\lambda_0 \to 0^+$ and using the relation (2.7), we conclude that

$$G(\lambda) - \mathbb{E}[X] \leq \frac{1}{2}\sigma^2\lambda$$

which is equivalent to the claim (2.5). We leave the extension of this proof to the case $I = \mathbb{R}$ as an exercise for the reader. ■

Thus far, we have seen how a particular upper bound (2.4) on the entropy $\mathbb{H}\left(e^{\lambda X}\right)$ translates into a bound on the cumulant generating function (2.5), and hence into sub-Gaussian tail bounds via the usual Chernoff argument. It is natural to explore to what extent this approach may be generalized. As seen previously in Chapter 2, a broader class of random variables are those with sub-exponential tails, and the following result is the analog of Proposition 2.0.1 in this case.

Proposition 2.0.2 — Bernstein entropy bound. Suppose that there are positive constants b and σ such that the entropy $\mathbb{H}\left(e^{\lambda X}\right)$ satisfies the bound

$$\mathbb{H}\left(e^{\lambda X}\right) \leq \lambda^2\left\{b\varphi'_{\mathbf{x}}(\lambda) + \varphi_{\mathbf{x}}(\lambda)\left(\sigma^2 - b\mathbb{E}[X]\right)\right\} \quad \text{for all } \lambda \in [0, 1/b) \tag{2.8}$$

Then X satisfies the bound

$$\log\mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right] \leq \sigma^2\lambda^2(1 - b\lambda)^{-1} \quad \text{for all } \lambda \in [0, 1/b) \tag{2.9}$$

As a consequence of the usual Chernoff argument, Proposition 2.0.2 implies that X satisfies the upper tail bound

$$\mathbb{P}[X \geq \mathbb{E}[X] + \delta] \leq \exp\left(-\frac{\delta^2}{4\sigma^2 + 2b\delta}\right) \quad \text{for all } \delta \geq 0$$

which (modulo non-optimal constants) is the usual Bernstein-type bound to be expected for a variable with sub-exponential tails. See Proposition 1.0.5 from Chapter 2 for further details on such Bernstein bounds.

We now turn to the proof of Proposition 2.0.2.

Proof. As before, we omit the dependence of φ_x on X throughout this proof so as to simplify notation. By rescaling and recentering arguments sketched out in Exercise 3.6, we may assume without loss of generality that $\mathbb{E}[X]=0$ and $b=1$, in which case the inequality (2.8) simplifies to

$$\mathbb{H}\left(e^{\lambda X}\right) \leq \lambda^2\left\{\varphi'(\lambda)+\varphi(\lambda)\sigma^2\right\} \quad \text{for all } \lambda \in [0,1) \tag{2.10}$$

Recalling the function $G(\lambda)=\frac{1}{\lambda}\log\varphi(\lambda)$ from the proof of Proposition 2.0.1, a little bit of algebra shows that condition (2.10) is equivalent to the differential inequality .

$$\begin{aligned} \lambda\varphi'(\lambda)-\varphi(\lambda)\log\varphi(\lambda) &\leq \lambda^2\left\{\varphi'(\lambda)+\varphi(\lambda)\sigma^2\right\} \\ \frac{1}{\lambda}\frac{\varphi'(\lambda)}{\varphi(\lambda)}-\frac{1}{\lambda^2}\log\varphi(\lambda) &\leq \sigma^2+\frac{\varphi'}{\varphi} \\ G' &\leq \sigma^2+\frac{\varphi'}{\varphi} \end{aligned}$$

Letting $\lambda_0>0$ be arbitrary and integrating both sides of this inequality over the interval (λ_0,λ) , we obtain

$$G(\lambda)-G(\lambda_0) \leq \sigma^2(\lambda-\lambda_0)+\log\varphi(\lambda)-\log\varphi(\lambda_0)$$

Since this inequality holds for all $\lambda_0>0$, we may take the limit as $\lambda_0 \to 0^+$. Doing so and using the facts that $\lim_{\lambda_0\to 0^+} G(\lambda_0)=G(0)=\mathbb{E}[X]$ and $\log\varphi(0)=0$, we obtain the bound

$$G(\lambda)-\mathbb{E}[X] \leq \sigma^2\lambda+\log\varphi(\lambda)$$

Substituting the definition of G and rearranging yields the claim (2.9). ■

Separately convex functions and the entropic method

The real power of the entropic method-as we now will see-manifests itself in dealing with concentration for functions of many random variables.

As an illustration, we begin by stating a deep result that can be proven in a relatively direct manner using the entropy method.

Definition 2.0.1 — Separately convex. We say that a function $f:\mathbb{R}^n \to \mathbb{R}$ is separately convex if, for each index $k \in \{1,2,\ldots,n\}$, the univariate function

$$y_k \mapsto f(x_1,x_2,\ldots,x_{k-1},y_k,x_{k+1},\ldots,x_n)$$

is convex for each fixed vector $(x_1,x_2,\ldots,x_{k-1},x_{k+1},\ldots,x_n) \in \mathbb{R}^{n-1}$.

Definition 2.0.2 — L-Lipschitz. A function f is L Lipschitz with respect to the Euclidean norm if

$$|f(x)-f(x')| \leq L\|x-x'\|_2 \quad \text{for all } x,x' \in \mathbb{R}^n \tag{2.11}$$

Theorem 2.0.3 Let $\{X_i\}_{i=1}^n$ be independent random variables, each supported on the interval $[a,b]$, and let $f:\mathbb{R}^n \to \mathbb{R}$ be separately convex, and L-Lipschitz with respect to the Euclidean norm. Then, for all $\delta > 0$, we have

$$\mathbb{P}\left[f(X) \geq \mathbb{E}[f(X)] + \delta\right] \leq \exp\left(-\frac{\delta^2}{4L^2(b-a)^2}\right) \tag{2.12}$$

R This result is the analog of the upper tail bound for Lipschitz functions of Gaussian variables (cf. Theorem 1.0.12 in Chapter 2), but applicable to independent and bounded variables instead. In contrast to the Gaussian case, the additional assumption of separate convexity cannot be eliminated in general; see the bibliographic section for further discussion. When f is jointly convex, other techniques can be used to obtain the lower tail bound as well; see Theorem 2.0.10 in the sequel for one such example.

Theorem 2.0.3 can be used to obtain order-optimal bounds for a number of interesting problems. As one illustration, we return to the Rademacher complexity, first introduced in Example 1.13 of Chapter 2

■ **Example 2.2 — Sharp bounds on Rademacher complexity.** Given a bounded subset $\mathcal{A} \subset \mathbb{R}^n$, consider the random variable $Z = \sup_{a\in\mathcal{A}} \sum_{k=1}^n a_k\varepsilon_k$, where $\varepsilon_k \in \{-1,+1\}$ are i.i.d. Rademacher variables. Let us view Z as a function of the random signs, and use Theorem 2.0.3to bound the probability of the tail event $\{Z \geq \mathbb{E}[Z]+t\}$

It suffices to verify the convexity and Lipschitz conditions of the theorem. First, since $Z = Z(\varepsilon_1,\ldots,\varepsilon_n)$ is the maximum of a collection of linear functions, it is jointly (and hence separately) convex. Let $Z' = Z(\varepsilon_1',\ldots,\varepsilon_n')$ where $\varepsilon' \in \{-1,+1\}^n$ is a second vector of sign variables. For any $a \in \mathcal{A}$, we have

$$\underbrace{\langle a,\varepsilon\rangle}_{\sum_{k=1}^n a_k\varepsilon_k} - Z' = \langle a,\varepsilon\rangle - \sup_{a'\in\mathcal{A}}\langle a',\varepsilon'\rangle \leq \langle a,\varepsilon-\varepsilon'\rangle \leq \|a\|_2\left\|\varepsilon-\varepsilon'\right\|_2$$

Taking suprema over $a \in \mathcal{A}$ yields that $Z - Z' \leq \left(\sup_{a\in\mathcal{A}}\|a\|_2\right)\|\varepsilon-\varepsilon'\|_2$. Since the same argument may be applied with the roles of ε and ε' reversed, we conclude that Z is lipschitz with parameter $\mathcal{W}(\mathcal{A}) := \sup_{a\in\mathcal{A}}\|a\|_2$, corresponding to the Euclidean width of the set. Putting together the pieces, Theorem 2.0.3 implies that

$$\mathbb{P}[Z \geq \mathbb{E}[Z]+t] \leq \exp\left(-\frac{t^2}{16\mathcal{W}^2(\mathcal{A})}\right)$$

Note that parameter $\mathcal{W}^2(\mathcal{A})$ may be substantially smaller than the quantity $\sum_{k=1}^n \sup_{a\in\mathcal{H}} a_k^2$ -indeed, possibly as much as a factor of n smaller! In such cases, Theorem 2.0.3 yields a much sharper tail bound than our earlier tail bound from Example 1.13, which was obtained by applying the bounded differences inequality.

Another use of Theorem 2.0.3 is in random matrix theory.

■ **Example 2.3 — Operator norm of a random matrix.** Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a random matrix, say with X_{ij} drawn i.i.d. from some zero-mean distribution supported on the unit interval $[-1,+1]$. The spectral or ℓ_2 -operator norm of X, denoted by $\|\mathbf{X}\|_2$, is its maximum singular value, given by

$$\|\mathbf{X}\|_2 = \max_{\substack{v\in\mathbb{R}^d \\ \|v\|_2=1}} \|\mathbf{X}v\|_2 = \max_{\substack{v\in\mathbb{R}^d \\ \|v\|_2=1}} \max_{\substack{u\in\mathbb{R}^d \\ \|u\|_2=1}} u^{\mathrm{T}}\mathbf{X}v \tag{2.13}$$

Theorem 2.0.3, we need to show that f is both lipschitz and convex. From its definition (2.13) , the operator norm is the supremum of a collection of functions that are linear in the entries $\mathbf{X}$; any such supremum is a convex function. Moreover, we have

$$\left|\|\mathbf{X}\|_2 - \|\mathbf{X}'\|_2\right| \leq^{(i)} \|\mathbf{X}-\mathbf{X}'\|_2 \leq^{(ii)} \|\mathbf{X}-\mathbf{X}'\|_{\mathrm{F}}$$

where step (i) follows from the triangle inequality, and step (ii) follows since the Frobenius norm of a matrix always upper bounds the operator norm. (The Frobenius norm $\|\mathbf{M}\|_{\mathrm{F}}$ of a matrix $\mathbf{M} \in \mathbb{R}^{n\times d}$ is simply the Euclidean norm of all its entries; see equation (**??**).) Consequently, the operator norm is Lipschitz with parameter $L = 1$, and thus Theorem 2.0.3 implies that

$$\mathbb{P}\left[\|\mathbf{X}\|_2 \geq \mathbb{E}\left[\|\mathbf{X}\|_2\right] + \delta\right] \leq e^{-\frac{\delta^2}{16}}$$

It is worth observing that this bound is the analog of our earlier bound (**??**) on the operator norm of a Gaussian random matrix, albeit with a worse constant. See Example 1.18 in Chapter 2 for further details on this Gaussian case.

Tensorization and separately convex functions

We now return to prove Theorem 2.0.3. The proof is based on two lemmas, both of which are of independent interest. Here we state these results and discuss some of their consequences, deferring their proofs to the end of this section. Our first lemma establishes an entropy bound for univariate functions:

Lemma 2.1 — Entropy bound for univariate functions. Let $X, Y \sim \mathbb{P}$ be a pair of i.i.d. variates. Then for any function $g : \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{H}\left(e^{\lambda g(X)}\right) \leq \lambda^2 \mathbb{E}\left[(g(X)-g(Y))^2 e^{\lambda g(X)} \mathbb{I}[g(X) \geq g(Y)]\right] \quad \text{for all } \lambda > 0 \tag{2.14}$$

If in addition X is supported on $[a,b]$, and g is convex and Lipschitz, then

$$\mathbb{H}\left(e^{\lambda g(X)}\right) \leq \lambda^2 (b-a)^2 \mathbb{E}\left[\left(g'(X)\right)^2 e^{\lambda g(X)}\right] \quad \text{for all } \lambda > 0 \tag{2.15}$$

where g' is the derivative.

In stating this lemma, we have used the fact that any convex and Lipschitz function has a derivative defined almost everywhere, a result known as Rademacher's theorem. Moreover, note that if g is lipschitz with parameter L, then we are guaranteed that $\|g'\|_\infty \le L$, so that inequality (2.15) implies an entropy bound of the form

$$\mathbb{H}\left(e^{\lambda g(X)}\right) \le \lambda^2 L^2 (b-a)^2 \mathbb{E}\left[e^{\lambda g(X)}\right] \quad \text{for all } \lambda > 0$$

In turn, by an application of Proposition 2.0.1, such an entropy inequality implies the upper tail bound

$$\mathbb{P}[g(X) \ge \mathbb{E}[g(X)] + \delta] \le e^{-\frac{\delta^2}{4L^2(b-a)^2}}$$

Thus, Lemma 2.1 implies the univariate version of Theorem 2.0.3. However, the inequality (2.15) is sharper, in that it involves $g'(X)$ as opposed to the worst-case bound L, and this distinction will be important in deriving the sharp result of Theorem 2.0.3. The more general inequality (2.15) will be useful in deriving functional versions of the Hoeffding and Bernstein inequalities (see Section 3.4).

Returning to the main thread, it remains to extend this univariate result to the multivariate setting, and the so-called tensorization property of entropy plays a key role here.

Definition 2.0.3 Given a function $f : \mathbb{R}^n \to \mathbb{R}$, an index $k \in \{1,2,\dots,n\}$ and a vector $X^{\backslash k} = (x_i, i \neq k) \in \mathbb{R}^{n-1}$, we define the **conditional entropy in coordinate** k via

$$\mathbb{H}\left(e^{\lambda f_k(X_k)} \mid X^{\backslash k}\right) := \mathbb{H}\left(e^{\lambda f(x_1,\dots,x_{k-1},X_k,x_{k+1},\dots,x_n)}\right)$$

where $f_k : \mathbb{R} \to \mathbb{R}$ is the coordinate function $x_k \mapsto f(x_1,\dots,x_k,\dots,x_n)$. To be clear, for a random vector $X^{\backslash k} \in \mathbb{R}^{n-1}$, the entropy $\mathbb{H}\left(e^{\lambda f_k(X_k)} \mid X^{\backslash k}\right)$ is a random variable, and its expectation is often referred to as the conditional entropy.)

The following result shows that the joint entropy can be upper bounded by a sum of univariate entropies, suitably defined.

Lemma 2.2 — Tensorization of entropy. Let $f : \mathbb{R}^n \to \mathbb{R}$, and let $\{X_k\}_{k=1}^n$ be independent random variables. Then

$$\mathbb{H}\left(e^{\lambda f(X_1,\dots,X_n)}\right) \le \mathbb{E}\left[\sum_{k=1}^n \mathbb{H}\left(e^{\lambda f_k(X_k)} \mid X^{\backslash k}\right)\right] \quad \text{for all } \lambda > 0$$

Equipped with these two results, we are now ready to prove Theorem 2.0.3
Proof of Theorem 2.0.3.

Proof. For any $k \in \{1,2,\dots,n\}$ and fixed vector $X^{\backslash k} \in \mathbb{R}^{n-1}$, our assumptions imply that the coordinate function f_k is convex, and hence Lemma 2.1 implies that, for all $\lambda > 0$, we have

$$\begin{aligned} \mathbb{H}\left(e^{\lambda f_k(X_k)} \mid X^{\backslash k}\right) &\leq \lambda^2(b-a)^2 \mathbb{E}_{X_k}\left[\left(f_k'(X_k)\right)^2 e^{\lambda f_k(X_k)} \mid X^{\backslash k}\right] \\ &= \lambda^2(b-a)^2 \mathbb{E}_{X_k}\left[\left(\frac{\partial f(x_1,\ldots,X_k,\ldots,x_n)}{\partial x_k}\right)^2 e^{\lambda f(x_1,\ldots,X_k,\ldots,x_n)}\right] \end{aligned}$$

where the second line involves unpacking the definition of the conditional entropy. Combined with Lemma 2.2, we find that the unconditional entropy is upper bounded as

$$\mathbb{H}\left(e^{\lambda f(X)}\right) \leq \lambda^2(b-a)^2 \mathbb{E}\left[\sum_{k=1}^{n}\left(\frac{\partial f(X)}{\partial x_k}\right)^2 e^{\lambda f(X)}\right] \leq^{(i)} \lambda^2(b-a)^2 L^2 \mathbb{E}\left[e^{\lambda f(X)}\right]$$

Here step (i) follows from the Lipschitz condition, which guarantees that

$$\|\nabla f(x)\|_2^2 = \sum_{k=1}^{n}\left(\frac{\partial f(x)}{\partial x_k}\right)^2 \leq L^2$$

almost surely. Thus, the tail bound (2.12) follows from an application of Proposition 2.0.1. ■

It remains to prove the two auxiliary lemmas used in the preceding proof-namely, Lemma 2.1 on entropy bounds for univariate Lipschitz functions, and Lemma 2.2 on the tensorization of entropy. We begin with the former property.

Proof of Lemma 2.1

Proof. By the definition of entropy, we can write

$$\begin{aligned} \mathbb{H}\left(e^{\lambda g(X)}\right) &= \mathbb{E}_X\left[\lambda g(X) e^{\lambda g(X)}\right] - \mathbb{E}_X\left[e^{\lambda g(X)}\right] \log\left(\mathbb{E}_Y\left[e^{\lambda g(Y)}\right]\right) \\ &\leq^{(i)} \mathbb{E}_X\left[\lambda g(X) e^{\lambda g(X)}\right] - \mathbb{E}_{X,Y}\left[e^{\lambda g(X)} \lambda g(Y)\right] \\ &= \frac{1}{2}\mathbb{E}_{X,Y}\left[\lambda\{g(X)-g(Y)\}\left\{e^{\lambda g(X)} - e^{\lambda g(Y)}\right\}\right] \\ &\overset{\text{(ii)}}{=} \lambda \mathbb{E}\left[\{g(X)-g(Y)\}\left\{e^{\lambda g(X)} - e^{\lambda g(Y)}\right\} \mathbb{I}[g(X) \geq g(Y)]\right] \end{aligned} \tag{2.16}$$

where step (i) follows from Jensen's inequality, and step (ii) follows from symmetry of X and Y

By convexity of the exponential, we have $e^s - e^t \leq e^s(s-t)$ for all $s, t \in \mathbb{R}$. For $s \geq t$, we can multiply both sides by $(s-t) \geq 0$, thereby obtaining

$$(s-t)\left(e^s - e^t\right) \mathbb{I}[s \geq t] \leq (s-t)^2 e^s \mathbb{I}[s \geq t]$$

Applying this bound with $s = \lambda g(X)$ and $t = \lambda g(Y)$ to the inequality (2.16) yields

$$\mathbb{H}\left(e^{\lambda g(X)}\right) \leq \lambda^2 \mathbb{E}\left[(g(X)-g(Y))^2 e^{\lambda g(X)} \mathbb{I}[g(X) \geq g(Y)]\right]$$

where we have recalled the assumption that $\lambda > 0$. If in addition g is convex, then we have the upper bound $g(x) - g(y) \leq g'(x)(x-y)$, and hence, for $g(x) \geq g(y)$

$$(g(x) - g(y))^2 \leq \left(g'(x)\right)^2 (x-y)^2 \leq \left(g'(x)\right)^2 (b-a)^2$$

where the final step uses the assumption that $x, y \in [a,b]$. Combining the pieces yields the claim. ■

We now turn to the tensorization property of entropy.

Proof of Lemma 2.2

Proof. The proof makes use of the following variational representation for entropy:

$$\mathbb{H}\left(e^{\lambda f(X)}\right) = \sup_g \left\{\mathbb{E}\left[g(X)e^{\lambda f(X)}\right] \mid \mathbb{E}\left[e^{g(X)}\right] \leq 1\right\} \tag{2.17}$$

This equivalence follows by a duality argument that we explore in Exercise 3.9.

For each $j \in \{1, 2, \ldots, n\}$, define $X_j^n = \left(X_j, \ldots, X_n\right)$. Let g be any function that satisfies $\mathbb{E}\left[e^{g(X)}\right] \leq 1$. We can then define an auxiliary sequence of functions $\left\{g^1, \ldots, g^n\right\}$ via and

$$g^1\left(X_1, \ldots, X_n\right) := g(X) - \log \mathbb{E}\left[e^{g(X)} \mid X_2^n\right]$$

$$g^k\left(X_k, \ldots, X_n\right) := \log \frac{\mathbb{E}\left[e^{g(X)} \mid X_k^n\right]}{\mathbb{E}\left[e^{g(X)} \mid X_{k+1}^n\right]} \quad \text{for } k = 2, \ldots, n$$

By construction, we have

$$\sum_{k=1}^n g^k\left(X_k, \ldots, X_n\right) = g(X) - \log \mathbb{E}\left[e^{g(X)}\right] \geq g(X) \tag{2.18}$$

and moreover $\mathbb{E}\left[\exp\left(g^k\left(X_k, X_{k+1}, \ldots, X_n\right)\right) \mid X_{k+1}^n\right] = 1$. We now use this decomposition within the variational representation (2.17), thereby obtaining the chain of upper bounds

$$\begin{aligned}
\mathbb{E}\left[g(X)e^{\lambda f(X)}\right] &\leq^{(i)} \sum_{k=1}^n \mathbb{E}\left[g^k\left(X_k, \ldots, X_n\right) e^{\lambda f(X)}\right] \\
&= \sum_{k=1}^n \mathbb{E}_{X^{\backslash k}}\left[\mathbb{E}_{X_k}\left[g^k\left(X_k, \ldots, X_n\right) e^{\lambda f(X)} \mid X^{\backslash k}\right]\right] \\
&\leq^{(ii)} \sum_{k=1}^n \mathbb{E}_{X^{\backslash k}}\left[\mathbb{H}\left(e^{\lambda f_k(X_k)} \mid X^{\backslash k}\right)\right]
\end{aligned}$$

where inequality (i) uses the bound (2.18), and inequality (ii) applies the variational representation (2.17) to the univariate functions, and also makes use of the fact that $\mathbb{E}\left[g^k\left(X_k, \ldots, X_n\right) \mid X^{\backslash k}\right] = 1$. Since this argument applies to any function g such that $\mathbb{E}\left[e^{g(X)}\right] \leq 1$, we may take the supremum over the left-hand side, and combined with the variational representation (2.17), we conclude that

$$\mathbb{H}\left(e^{\lambda f(X)}\right) \leq \sum_{k=1}^n \mathbb{E}_{X_{|k}}\left[\mathbb{H}\left(e^{\lambda f_k(X_k)} \mid X_{|k}\right)\right]$$

as claimed. ■

2.0.2 A geometric perspective on concentration

Historically, this geometric viewpoint is among the oldest, dating back to the classical result of Lévy on concentration of measure for Lipschitz functions of Gaussians. It also establishes deep links between probabilistic concepts and high-dimensional geometry.

The results of this section are most conveniently stated in terms of a metric measure space-namely, a metric space $(\mathcal{X},\rho)$ endowed with a probability measure $\mathbb{P}$ on its Borel sets. Some canonical examples of metric spaces for the reader to keep in mind are the set $\mathcal{X} = \mathbb{R}^n$ equipped with the usual Euclidean metric $\rho(x,y) := \|x-y\|_2$, and the discrete cube $\mathcal{X} = \{0,1\}^n$ equipped with the Hamming metric $\rho(x,y) = \sum_{j=1}^n \mathbb{I}\left[x_j \neq y_j\right]$. Associated with any metric measure space is an object known as its concentration function, which is defined in a geometric manner via the ϵ -enlargements of sets. The concentration function specifies **how rapidly**, as a function of ϵ, the probability of any ϵ -enlargement increases towards one. As we will see, this function is intimately related to the concentration properties of Lipschitz functions on the metric space.

Concentration functions

Given a set $A \subseteq X$ and a point $x \in X$, define the quantity

$$\rho(x,A) := \inf_{y\in A} \rho(x,y) \tag{2.19}$$

which measures the distance between the point x and the closest point in the set A. Given a parameter $\epsilon > 0$, the ϵ **-enlargement of** A is given by

$$A^\epsilon := \{x \in \mathcal{X} \mid \rho(x,A) < \epsilon\}$$

In words, the set A^ϵ corresponds to the open neighborhood of points lying at distance less than ϵ from A. With this notation, the concentration function of the metric measure space $(\mathcal{X},\rho,\mathbb{P})$ is defined as follows:

Definition 2.0.4 — concentration function. The **concentration function** $\alpha : [0,\infty) \to \mathbb{R}_+$ associated with metric measure space $(\mathbb{P},\mathcal{X},\rho)$ is given by

$$\alpha_{\mathbb{P},(X,\rho)}(\epsilon) := \sup_{A\subseteq\mathcal{X}} \left\{1 - \mathbb{P}\left[A^\epsilon\right] \mid \mathbb{P}[A] \geq \frac{1}{2}\right\} \tag{2.20}$$

where the supremum is taken over all measurable subsets A.

When the underlying metric space (X,ρ) is clear from the context, we frequently use the abbreviated notation $\alpha_{\mathbb{P}}$. It follows immediately from the definition (2.20) that $\alpha_{\mathbb{P}}(\epsilon) \in \left[0,\frac{1}{2}\right]$ for all $\epsilon \geq 0$. Of primary interest is the behavior of the concentration function as ϵ increases, and, more precisely, how rapidly it approaches zero. Let us consider some examples to illustrate.

■ **Example 2.4 — Concentration function for sphere.** Consider the metric measure space defined by the uniform distribution over the n -dimensional Euclidean sphere

$$S^{n-1} := \{x \in \mathbb{R}^n \mid \|x\|_2 = 1\}$$

equipped with the geodesic distance $\rho(x,y) := \arccos\langle x, y\rangle$. Let us upper bound the concentration function $\alpha_{S^{n-1}}$ defined by the triplet $(\mathbb{P}, S^{n-1}, \rho)$, where $\mathbb{P}$ is the uniform distribution over the sphere. For each $y \in S^{n-1}$, we can define the hemisphere

$$H_y := \left\{x \in S^{n-1} \mid \rho(x,y) \geq \pi/2\right\} = \left\{x \in S^{n-1} \mid \langle x, y\rangle \leq 0\right\} \tag{2.21}$$

as illustrated in Figure 3.1(a). With some simple geometry, it can be shown that its ϵ enlargement corresponds to the set

$$H_y^\epsilon = \left\{z \in S^{n-1} \mid \langle z, y\rangle < \sin(\epsilon)\right\}$$

as illustrated in Figure 3.1(b). Note that $\mathbb{P}\left[H_y\right] = 1/2$, so that the hemisphere (2.21) is a candidate set for the supremum defining the concentration function (2.20). The classical isoperimetric theorem of Lévy asserts that these hemispheres are extremal, meaning that they achieve the supremum, viz.

$$\alpha_{S^{n-1}}(\epsilon) = 1 - \mathbb{P}\left[H_y^\epsilon\right]$$

Let us take this fact as given, and use it to compute an upper bound on the concentration function. In order to do so, we need to lower bound the probability $\mathbb{P}\left[H_y^\epsilon\right]$. Since $\sin(\epsilon) \geq \epsilon/2$ for all $\epsilon \in (0, \pi/2]$, the enlargement contains the set

$$\widetilde{H}_y^\epsilon := \left\{z \in S^{n-1} \mid \langle z, y\rangle \leq \frac{1}{2}\epsilon\right\}$$

and hence $\mathbb{P}\left[H_y^\epsilon\right] \geq \mathbb{P}\left[\widetilde{H}_y^\epsilon\right]$. Finally, a geometric calculation, left as an exercise for the reader, yields that, for all $\epsilon \in (0, \sqrt{2})$, we have

$$\mathbb{P}\left[\widetilde{H}_y^\epsilon\right] \geq 1 - \left(1 - \left(\frac{\epsilon}{2}\right)^2\right)^{n/2} \geq 1 - e^{-n\epsilon^2/8}$$

where we have used the inequality $(1-t) \leq e^{-t}$ with $t = \epsilon^2/4$. We thus obtain that the concentration function is upper bounded as $\alpha_{S^{n-1}}(\epsilon) \leq e^{-n\epsilon^2/8}$.

A similar but more careful approach to bounding $\mathbb{P}\left[H_y\right]$ can be used to establish the sharper upper bound

$$\alpha_{S^{n-1}}(\epsilon) \leq \sqrt{\frac{\pi}{2}} e^{-\frac{n\epsilon^2}{2}} \tag{2.22}$$

The bound (2.22) is an extraordinary conclusion, originally due to Lévy, and it is worth pausing to think about it in more depth. Among other consequences, it implies that, if we consider a central slice of the sphere of width ϵ, say a set of the form

$$T_y(\epsilon) := \left\{z \in S^{n-1} || \langle z, y\rangle | \leq \epsilon/2\right\}$$

as illustrated in Figure 2.1(c), then it occupies a huge fraction of the total volume: in particular, we have $\mathbb{P}\left[T_y(\epsilon)\right] \geq 1-\sqrt{2\pi}\exp\left(-\frac{n\epsilon^2}{2}\right)$. Moreover, this conclusion holds for any such slice. To be clear, the two-dimensional instance shown in Figure 2.1(c)$-$ like any lowdimensional example -fails to capture the behavior of high-dimensional spheres. In general, our low-dimensional intuition can be very misleading when applied to high-dimensional settings.

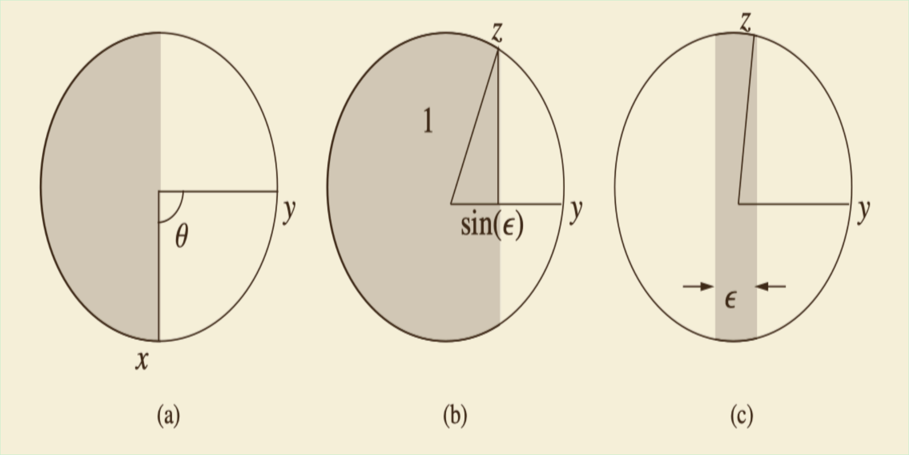


Figure 2.1: (a) Idealized illustration of the sphere S^{n-1}. Any vector $y \in S^{n-1}$ defines a hemisphere $H_y=\left\{x \in S^{n-1} \mid \langle x,y\rangle \leq 0\right\}$, corresponding to those vectors whose angle $\theta=\arccos\langle x,y\rangle$ with y is at least $\pi/2$ radians. (b) The ϵ -enlargement of the hemisphere H_y. (c) A central slice $T_y(\epsilon)$ of the sphere of width ϵ.

Connection to Lipschitz functions

In Chapter 2 and the preceding section of this chapter, we explored some methods for obtaining deviation and concentration inequalities for various types of Lipschitz functions. The concentration function $\alpha_{\mathbb{P},(X,\rho)}$ turns out to be intimately related to such results on the tail behavior of Lipschitz functions. In particular, suppose that a function $f: X \rightarrow \mathbb{R}$ is L -Lipschitz with respect to the metric $\rho--$ that is,

$$|f(x)-f(y)| \leq L\rho(x,y) \quad \text{for all } x,y \in \mathcal{X}$$

Given a random variable $X \sim \mathbb{P}$, let m_f be any median of $f(X)$, meaning a number such that

$$\mathbb{P}\left[f(X) \geq m_f\right] \geq 1/2 \quad \text{and} \quad \mathbb{P}\left[f(X) \leq m_f\right] \geq 1/2$$

Define the set $A=\left\{x \in \mathcal{X} \mid f(x) \leq m_f\right\}$, and consider its $\frac{\epsilon}{L}$ -enlargement $A^{\epsilon/L}$. For any $x \in A^{\epsilon/L}$, there exists some $y \in A$ such that $\rho(x,y)<\epsilon/L$. Combined with the Lipschitz property, we conclude that $|f(y)-f(x)| \leq L\rho(x,y)<\epsilon$, and hence that

$$A^{\epsilon/L} \subseteq \left\{x \in \mathcal{X} \mid f(x)<m_f+\epsilon\right\} \tag{2.23}$$

Consequently, we have

$$\mathbb{P}\left[f(X) \geq m_f + \epsilon\right] \leq^{(i)} 1 - \mathbb{P}\left[A^{\epsilon/L}\right] \leq^{(ii)} \alpha_{\mathbb{P}}(\epsilon/L)$$

where inequality (i) follows from the inclusion (2.23), and inequality (ii) uses the fact $\mathbb{P}[A] \geq 1/2$, and the definition (2.20). Applying a similar argument to $-f$ yields an analogous left-sided deviation inequality $\mathbb{P}\left[f(X) \leq m_f - \epsilon\right] \leq \alpha_{\mathbb{P}}(\epsilon/L)$, and putting together the pieces yields the concentration inequality

$$\mathbb{P}\left[\left|f(X) - m_f\right| \geq \epsilon\right] \leq 2\alpha_{\mathbb{P}}(\epsilon/L)$$

As shown in Exercise 2.14 from Chapter 2, such sharp concentration around the median is equivalent (up to constant factors) to concentration around the mean. Consequently, we have shown that bounds on the concentration function (2.20) imply concentration inequalities for any Lipschitz function. This argument can also be reversed, yielding the following equivalence between control on the concentration function, and the behavior of Lipschitz functions.

Proposition 2.0.4 Given a random variable $X \sim \mathbb{P}$ and concentration function $\alpha_{\mathbb{P}}$, any 1 -Lipschitz function on $(\mathcal{X}, \rho)$ satisfies

$$\mathbb{P}\left[\left|f(X) - m_f\right| \geq \epsilon\right] \leq 2\alpha_{\mathbb{P}}(\epsilon) \tag{2.24}$$

where m_f is any median of f. Conversely, suppose that there is a function $\beta : \mathbb{R}_+ \to \mathbb{R}_+$ such that, for any 1 - Lipschitz function on $(\mathcal{X}, \rho)$,

$$\mathbb{P}[f(X) \geq \mathbb{E}[f(X)] + \epsilon] \leq \beta(\epsilon) \quad \text{for all } \epsilon \geq 0 \tag{2.25}$$

Then the concentration function satisfies the bound $\alpha_{\mathbb{P}}(\epsilon) \leq \beta(\epsilon/2)$

Proof. It remains to prove the converse claim. Fix some $\epsilon \geq 0$, and let A be an arbitrary measurable set with $\mathbb{P}[A] \geq 1/2$. Recalling the definition of $\rho(x, A)$ from equation (2.19), let us consider the function $f(x) := \min\{\rho(x, A), \epsilon\}$. It can be seen that f is 1 -Lipschitz, and moreover that $1 - \mathbb{P}\left[A^\epsilon\right] = \mathbb{P}[f(X) \geq \epsilon]$. On the other hand, our construction guarantees that

$$\mathbb{E}[f(X)] \leq (1 - \mathbb{P}[A])\epsilon \leq \epsilon/2$$

(the value of f in A is 0) whence we have

$$\mathbb{P}[f(X) \geq \epsilon] \leq \mathbb{P}[f(X) \geq \mathbb{E}[f(X)] + \epsilon/2] \leq \beta(\epsilon/2)$$

where the final inequality uses the assumed condition $(2.25b)$. ■

Proposition 2.0.4 has a number of concrete interpretations in specific settings.

■ **Example 2.5 — Lévy concentration on S^{n-1}.** From our earlier discussion in Example 2.4 , the concentration function for the uniform distribution over the sphere S^{n-1} can be upper bounded as

$$\alpha_{S^{n-1}}(\epsilon) \leq \sqrt{\frac{\pi}{2}} e^{-\frac{n^2\epsilon^2}{2}}$$

Consequently, for any 1 -Lipschitz function f defined on the sphere S^{n-1}, we have the two-sided bound

$$\mathbb{P}\left[\left|f(X) - m_f\right| \geq \epsilon\right] \leq \sqrt{2\pi} e^{-\frac{n^2\epsilon^2}{2}}$$

where m_f is any median of f. Moreover, by the result of Exercise 2.14(d), we also have

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq \epsilon] \leq 2\sqrt{2\pi} e^{-\frac{n\epsilon^2}{8}}$$

■ **Example 2.6 — Concentration for Boolean hypercube.**

Consider the Boolean hypercube $\mathcal{X} = \{0,1\}^n$ equipped with the usual Hamming metric

$$\rho_H(x,y) := \sum_{j=1}^{n} \mathbb{I}\left[x_j \neq y_j\right]$$

Given this metric, we can define the Hamming ball

$$\mathbb{B}_H(r;x) = \{y \in \{0,1\}^n \mid \rho_H(y,x) \leq r\}$$

of radius r centered at some $x \in \{0,1\}^n$. Of interest here are the Hamming balls centered at the all-zeros vector 0 and all-ones vector 1, respectively. In particular, in this example, we show how a classical combinatorial result due to Harper can be used to bound the concentration function of the metric measure space consisting of the Hamming metric along with the uniform distribution $\mathbb{P}$.

Given two non-empty subsets A and B of the binary hypercube, one consequence of Harper's theorem is that we can always find two positive integers r_A and r_B, and associated subsets A' and B', with the following properties:

1. the sets A' and B' are sandwiched as

$$\mathbb{B}_H\left(r_A - 1;0\right) \subseteq A' \subseteq \mathbb{B}_H\left(r_A;0\right) \quad \text{and} \quad \mathbb{B}_H\left(r_B - 1;1\right) \subseteq B' \subseteq \mathbb{B}_H\left(r_B;1\right)$$

2. the cardinalities are matched as card $(A) = \operatorname{card}(A')$ and $\operatorname{card}(B) = \operatorname{card}(B')$;
3. we have the lower bound $\rho_H(A',B') \geq \rho_H(A,B)$.

Let us now show that this combinatorial theorem implies that

$$\alpha_{\mathbb{P}}(\epsilon) \leq e^{-\frac{2\epsilon^2}{n}} \quad \text{for all } n \geq 3 \tag{2.26}$$

Consider any subset such that $\mathbb{P}[A] = \frac{\text{card}(A)}{2^n} \geq \frac{1}{2}$. For any $\epsilon > 0$, define the set $B = \{0,1\}^n \backslash A^\epsilon$. In order to prove the bound (2.26), it suffices to show that $\mathbb{P}[B] \leq e^{-\frac{2\epsilon^2}{n}}$. Since we always have $\mathbb{P}[B] \leq \frac{1}{2} \leq e^{-\frac{2}{n}}$ for $n \geq 3$, it suffices to restrict our attention to $\epsilon > 1$. By construction, we have

$$\rho_H(A,B) = \min_{a \in A, b \in B} \rho_H(a,b) \geq \epsilon$$

Let A' and B' denote the subsets guaranteed by Harper's theorem. Since A has cardinality at least 2^{n-1}, the set A', which has the same cardinality as A, must contain all vectors with at most $n/2$ ones. Moreover, by the cardinality matching condition and our choice of the uniform distribution, we have $\mathbb{P}[B] = \mathbb{P}[B']$. On the other hand, the set B' is contained within a Hamming ball centered at the all-ones vector, and we have $\rho_H(A', B') \geq \epsilon > 1$.

Consequently, any vector $b \in B'$ must contain at least $\frac{n}{2} + \epsilon$ ones. Thus, if we let $\{X_i\}_{i=1}^n$ be a sequence of i.i.d. Bernoulli variables, we have

$$\mathbb{P}[B'] \leq \mathbb{P}\left[\sum_{i=1}^n X_i \geq \frac{n}{2} + \epsilon\right] \leq e^{-\frac{2\epsilon^2}{n}}$$

where the final inequality follows from the Hoeffding bound. Since A was an arbitrary set with $\mathbb{P}[A] \geq \frac{1}{2}$, we have shown that the concentration function satisfies the bound (2.26). Applying Proposition 2.0.4, we conclude that any 1-Lipschitz function on the Boolean hypercube satisfies the concentration bound

$$\mathbb{P}\left[|f(X) - m_f| \geq \epsilon\right] \leq 2e^{-\frac{2\epsilon^2}{n}}$$

Thus, modulo the negligible difference between the mean and median (see Exercise 2.14), we have recovered the bounded differences inequality (1.21) for Lipschitz functions on the Boolean hypercube.

From geometry to concentration

The geometric perspective suggests the possibility of a variety of connections between convex geometry and the concentration of measure. Consider, for instance, the Brunn-Minkowski inequality: in one of its formulations, it asserts that, for any two convex bodies C and D in $\mathbb{R}^n$, we have

$$[\text{vol}(\lambda C + (1-\lambda)D)]^{1/n} \geq \lambda[\text{vol}(C)]^{1/n} + (1-\lambda)[\text{vol}(D)]^{1/n} \quad \text{for all } \lambda \in [0,1] \tag{2.27}$$

Here we use

$$\lambda C + (1-\lambda)D := \{\lambda c + (1-\lambda)d \mid c \in C, d \in D\}$$

to denote the Minkowski sum of the two sets. The Brunn-Minkowski inequality and its variants are intimately connected to concentration of measure. To appreciate the connection, observe that the concentration function (2.20) defines a notion of extremal sets-namely, those that minimize the measure $\mathbb{P}[A^\epsilon]$ subject to a constraint on the size of

$\mathbb{P}[A]$. Viewing the volume as a type of unnormalized probability measure, the Brunn-Minkowski inequality (2.27) can be used to prove a classical result of this type:

■ **Example 2.7 — Classical isoperimetric inequality in $\mathbb{R}^n$.** Consider the Euclidean sphere $\mathbb{B}_2^n := \{x \in \mathbb{R}^n \mid \|x\|_2 \leq 1\}$ in $\mathbb{R}^n$. The classical isoperimetric inequality asserts that, for any set $A \subset \mathbb{R}^n$ such that $\operatorname{vol}(A) = \operatorname{vol}(\mathbb{B}_2^n)$, the volume of its ϵ -enlargement A^ϵ is lower bounded as

$$\operatorname{vol}(A^\epsilon) \geq \operatorname{vol}\left([\mathbb{B}_2^n]^\epsilon\right)$$

showing that the ball $\mathbb{B}_2^n$ is extremal. In order to verify this bound, we note that

$$[\operatorname{vol}(A^\epsilon)]^{1/n} = [\operatorname{vol}(A + \epsilon\mathbb{B}_2^n)]^{1/n} \geq [\operatorname{vol}(A)]^{1/n} + [\operatorname{vol}(\epsilon\mathbb{B}_2^n)]^{1/n}$$

where the lower bound follows by applying the Brunn-Minkowski inequality (2.27) with appropriate choices of (λ, C, D); see Exercise 3.10 for the details. Since $\operatorname{vol}(A) = \operatorname{vol}(\mathbb{B}_2^n)$ and $[\operatorname{vol}(\epsilon\mathbb{B}_2^n)]^{1/n} = \epsilon \operatorname{vol}(\mathbb{B}_2^n)^{\frac{1}{n}}$, we see that

$$\operatorname{vol}(A^\epsilon)^{1/n} \geq (1+\epsilon)\operatorname{vol}(\mathbb{B}_2^n)^{1/n} = \left[\operatorname{vol}\left((B_2^n)^\epsilon\right)\right]^{1/n}$$

which establishes the claim.

The Brunn-Minkowski inequality has various equivalent formulations. For instance, it can also be stated as

$$\operatorname{vol}(\lambda C + (1-\lambda)D) \geq [\operatorname{vol}(C)]^\lambda [\operatorname{vol}(D)]^{1-\lambda} \quad \text{for all } \lambda \in [0,1] \tag{2.28}$$

This form of the Brunn-Minkowski inequality can be used to establish Lévy-type concentration for the uniform measure on the sphere, albeit with slightly weaker constants than the derivation in Example 2.4. In Exercise 2.4, we explore the equivalence between inequality (2.28) and our original statement (2.27) of the Brunn-Minkowski inequality.

The modified form (2.28) of the Brunn-Minkowski inequality also leads naturally to a functional-analytic generalization, due to Prékopa and Leindler. In turn, this generalized inequality can be used to derive concentration inequalities for strongly log-concave measures.

Theorem 2.0.5 — Prékopa-Leindler inequality. Let u, v, w be non-negative integrable functions such that, for some $\lambda \in [0,1]$, we have

$$w(\lambda x + (1-\lambda)y) \geq [u(x)]^\lambda [v(y)]^{1-\lambda} \quad \text{for all } x, y \in \mathbb{R}^n \tag{2.29}$$

Then

$$\int w(x)dx \geq \left(\int u(x)dx\right)^\lambda \left(\int v(x)dx\right)^{1-\lambda}$$

In order to see how this claim implies the classical Brunn-Minkowski inequality (2.28), consider the choices

$$u(x) = \mathbb{I}_C(x), \quad v(x) = \mathbb{I}_D(x) \quad \text{and} \quad w(x) = \mathbb{I}_{\lambda C+(1-\lambda)D}(x)$$

respectively. Here $\mathbb{I}_C$ denotes the binary-valued indicator function for the event $\{x \in C\}$, with the other indicators defined in an analogous way. In order to show that the classical inequality (2.28) follows as a consequence of Theorem 2.0.5, we need to verify that

$$\mathbb{I}_{\lambda C+(1-\lambda)D}(\lambda x + (1-\lambda)y) \geq [\mathbb{I}_C(x)]^{\lambda} [\mathbb{I}_D(y)]^{1-\lambda} \quad \text{for all } x, y \in \mathbb{R}^n$$

For $\lambda = 0$ or $\lambda = 1$, the claim is immediate. For any $\lambda \in (0,1)$, if either $x \notin C$ or $y \notin D$, the right-hand side is zero, so the statement is trivial. Otherwise, if $x \in C$ and $y \in D$, then both sides are equal to one.

The Prékopa-Leindler inequality can be used to establish some interesting concentration inequalities of Lipschitz functions for a particular subclass of distributions, one which allows for some dependence.

Definition 2.0.5 — Strongly long-concave distribution. In particular, we say that a distribution $\mathbb{P}$ with a density p (with respect to the Lebesgue measure) is a **strongly log-concave distribution** if the function $\log p$ is strongly concave. Equivalently stated, this condition means that the density can be written in the form $p(x) = \exp(-\psi(x))$, where the function $\psi : \mathbb{R}^n \to \mathbb{R}$ is strongly convex, meaning that there is some $\gamma > 0$ such that

$$\lambda\psi(x) + (1-\lambda)\psi(y) - \psi(\lambda x + (1-\lambda)y) \geq \frac{\gamma}{2}\lambda(1-\lambda)\|x-y\|_2^2$$

for all $\lambda \in [0,1]$, and $x, y \in \mathbb{R}^n$.

For instance, it is easy to verify that the distribution of a standard Gaussian vector in n dimensions is strongly log-concave with parameter $\gamma = 1$. More generally, any Gaussian distribution with covariance matrix $\Sigma > 0$ is strongly log-concave with parameter $\gamma = \gamma_{\min}\left(\Sigma^{-1}\right) = \left(\gamma_{\max}(\Sigma)\right)^{-1}$. In addition, there are a variety of non-Gaussian distributions that are also strongly log-concave. For any such distribution, Lipschitz functions are guaranteed to concentrate, as summarized in the following:

Theorem 2.0.6 Let $\mathbb{P}$ be any strongly log-concave distribution with parameter $\gamma > 0$. Then for any function $f : \mathbb{R}^n \to \mathbb{R}$ that is L -Lipschitz with respect to Euclidean norm, we have

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2e^{-\frac{\gamma t^2}{4L^2}}$$

R Since the standard Gaussian distribution is log concave with parameter $\gamma = 1$, this theorem implies our earlier result (Theorem 1.0.12), albeit with a sub-optimal constant in the exponent.

Proof. Let h be an arbitrary zero-mean function with Lipschitz constant L with respect to the Euclidean norm. It suffices to show that $\mathbb{E}\left[e^{h(X)}\right] \leq e^{\frac{L^2}{\gamma}}$. Indeed, if this inequality holds, then, given an arbitrary function f with Lipschitz constant K and $\lambda \in \mathbb{R}$, we can apply this inequality to the zero-mean function $h := \lambda(f - \mathbb{E}[f(X)])$, which has Lipschitz constant $L = \lambda K$. Doing so yields the bound

$$\mathbb{E}\left[e^{\lambda(f(X)-\mathbb{E}[f(X)]}\right] \leq e^{\frac{\lambda^2 K^2}{\gamma}} \quad \text{for all } \lambda \in \mathbb{R}$$

which shows that $f(X) - \mathbb{E}[f(X)]$ is a sub-Gaussian random variable. As shown in Chapter 2 , this type of uniform control on the moment generating function implies the claimed tail bound.

Accordingly, for a given zero-mean function h that is L -Lipschitz and for given $\lambda \in (0,1)$ and $x, y \in \mathbb{R}^n$, define the function

$$g(y) := \inf_{x \in \mathbb{R}^n} \left\{h(x) + \frac{\gamma}{4}\|x-y\|_2^2\right\}$$

known as the inf-convolution of h with the rescaled Euclidean norm. With this definition, the proof is based on applying the Prékopa-Leindler inequality with $\lambda = 1/2$ to the triplet of functions $w(z) \equiv p(z) = \exp(-\psi(z))$, the density of $\mathbb{P}$, and the pair of functions

$$u(x) := \exp(-h(x) - \psi(x)) \quad \text{and} \quad v(y) := \exp(g(y) - \psi(y))$$

We first need to verify that the inequality (2.29) holds with $\lambda = 1/2$. By the definitions of u and v, the logarithm of the right-hand side of inequality (2.29) $-$ call it R for short- -is given by

$$R = \frac{1}{2}\{g(y) - h(x)\} - \frac{1}{2}\psi(x) - \frac{1}{2}\psi(y) = \frac{1}{2}\{g(y) - h(x) - 2E(x,y)\} - \psi(x/2 + y/2)$$

where $E(x,y) := \frac{1}{2}\psi(x) + \frac{1}{2}\psi(y) - \psi(x/2 + y/2)$. Since $\mathbb{P}$ is a γ -log-concave distribution, the function ψ is γ strongly convex, and hence $2E(x,y) \geq \frac{\gamma}{4}\|x-y\|_2^2$. Substituting into the earlier representation of R, we find that

$$R \leq \frac{1}{2}\left\{g(y) - h(x) - \frac{\gamma}{4}\|x-y\|_2^2\right\} - \psi(x/2 + y/2) \leq -\psi(x/2 + y/2)$$

where the final inequality follows from the definition of the inf-convolution g. We have thus verified condition (2.29) with $\lambda = 1/2$.

Now since $\int w(x)dx = \int p(x)dx = 1$ by construction, the Prékopa-Leindler inequality implies that

$$0 \geq \frac{1}{2}\log\int e^{-h(x)-\psi(x)}dx + \frac{1}{2}\log\int e^{g(y)-\psi(y)}dy$$

Rewriting the integrals as expectations and rearranging yields

$$1 \geq \int e^{-h(x)-\psi(x)}dx + \int e^{g(y)-\psi(y)}dy \tag{2.30}$$

and $w(z) \equiv p(z) = \exp(-\psi(z))$ is the density of $\mathbb{P}$,

$$\mathbb{E}\left[e^{g(Y)}\right] \leq \frac{1}{\mathbb{E}\left[e^{-h(X)}\right]} \leq^{(i)} \frac{1}{e^{\mathbb{E}[-h(X)]}} \overset{(ii)}{=} 1 \tag{2.31}$$

where step (i) follows from Jensen's inequality, and convexity of the function $t \mapsto \exp(-t)$, and step (ii) uses the fact that $\mathbb{E}[-h(X)] = 0$ by assumption. Finally, since h is an L -Lipschitz function, we have $|h(x) - h(y)| \leq L\|x-y\|_2(i)$, and hence

$$\begin{aligned} g(y) = \inf_{x \in \mathbb{R}^n} \left\{h(x) + \frac{\gamma}{4}\|x-y\|_2^2\right\} &\geq^{(i)} h(y) + \inf_{x \in \mathbb{R}^n} \left\{-L\|x-y\|_2 + \frac{\gamma}{4}\|x-y\|_2^2\right\} \\ &= h(y) - \frac{L^2}{\gamma} \end{aligned}$$

Combined with the bound (2.31), we conclude that $\mathbb{E}\left[e^{h(Y)}\right] \leq \exp\left(\frac{L^2}{\gamma}\right)$, as claimed. ■

2.0.3 Wasserstein distances and information inequalities

We now turn to the topic of Wasserstein distances and information inequalities, also known as transportation cost inequalities. On one hand, the transportation cost approach can be used to obtain some sharp results for Lipschitz functions of independent random variables. Perhaps more importantly, it is especially well suited to certain types of dependent random variables, such as those arising in Markov chains and other types of mixing processes.

Wasserstein distances

We begin by defining the notion of a Wasserstein distance.

Definition 2.0.6 — Smallest-Lipschitz. Given a metric space $(\mathcal{X}, \rho)$, a function $f : \mathcal{X} \to \mathbb{R}$ is L -Lipschitz with respect to the metric ρ if

$$|f(x) - f(x')| \leq L\rho(x, x') \qquad \text{for all } x, x' \in \mathcal{X}$$

and we use $\|f\|_{\text{Lip}}$ to denote the smallest L for which this inequality holds.

Definition 2.0.7 — Wasserstein metric induced by ρ. Given two probability distributions $\mathbb{Q}$ and $\mathbb{P}$ on $\mathcal{X}$, we can then measure the distance between them via

$$W_\rho(\mathbb{Q}, \mathbb{P}) = \sup_{\|f\|_{\text{Lip}} \leq 1} \left[\int f\, d\mathbb{Q} - \int f\, d\mathbb{P}\right] \tag{2.32}$$

where the supremum ranges over all 1-Lipschitz functions. This distance measure is referred to as the Wasserstein metric induced by ρ ($\|f\|_{\text{Lip}}$ is associated with ρ). It can be verified that, for each choice of the metric ρ, this definition defines a distance on the space of probability measures.

■ **Example 2.8 — Hamming metric and total variation distance.** Consider the Hamming metric $\rho(x, x') = \mathbb{I}[x \neq x']$. We claim that, in this case, the associated Wasserstein distance is equivalent to the total variation distance

$$\|\mathbb{Q}-\mathbb{P}\|_{\mathrm{TV}} := \sup_{A\subseteq\mathcal{X}} |\mathbb{Q}(A)-\mathbb{P}(A)| \tag{2.33}$$

where the supremum ranges over all measurable subsets A. To see this equivalence, note that any function that is 1 -Lipschitz with respect to the Hamming distance satisfies the bound $|f(x)-f(x')|\leq 1$. Since the supremum (2.32) is invariant to constant offsets of the function, we may restrict the supremum to functions such that $f(x)\in[0,1]$ for all $x\in X$, thereby obtaining

$$W_{\mathrm{Ham}}(\mathbb{Q},\mathbb{P}) = \sup_{f:\mathcal{X}\to[0,1]} \int f(d\mathbb{Q}-d\mathbb{P}) \overset{\text{(i)}}{=} \|\mathbb{Q}-\mathbb{P}\|_{\mathrm{TV}}$$

where equality (i) follows from Exercise 3.13. In terms of the underlying densities ^3p and q taken with respect to a base measure v (This assumption entails no loss of generality, since $\mathbb{P}$ and $\mathbb{Q}$ both have densities with respect to $v=\frac{1}{2}(\mathbb{P}+\mathbb{Q})$.), we can write

$$W_{\mathrm{Ham}}(\mathbb{Q},\mathbb{P}) = \|\mathbb{Q}-\mathbb{P}\|_{\mathrm{TV}} = \frac{1}{2}\int |p(x)-q(x)|v(dx)$$

corresponding to (one half) the $L^1(v)$ -norm between the densities. Again, see Exercise 3.13 for further details on this equivalence.

By a classical and deep result in duality theory (see the bibliographic section for details), any Wasserstein distance has an equivalent definition as a type of **coupling-based distance**.

Definition 2.0.8 — coupling-based distance. A distribution $\mathbb{M}$ on the product space $\mathcal{X}\otimes\mathcal{X}$ is a coupling of the pair $(\mathbb{Q},\mathbb{P})$ if its marginal distributions in the first and second coordinates coincide with $\mathbb{Q}$ and $\mathbb{P}$, respectively.

In order to see the relation to the Wasserstein distance, let $f:\mathcal{X}\to\mathbb{R}$ be any 1 -Lipschitz function, and let $\mathbb{M}$ be any coupling. We then have

$$\int \rho(x,x')\,d\mathbb{M}(x,x') \geq^{(i)} \int (f(x)-f(x'))\,d\mathbb{M}(x,x') \overset{\text{(ii)}}{=} \int f(d\mathbb{P}-d\mathbb{Q}) \tag{2.34}$$

where the inequality (i) follows from the 1 -Lipschitz nature of f, and the equality (ii) follows since $\mathbb{M}$ is a coupling. The Kantorovich-Rubinstein duality guarantees the following important fact: if we minimize over all possible couplings, then this argument can be reversed, and in fact we have the equivalence

$$\underbrace{\sup_{\|f\|_{\mathrm{Lip}}\leq 1} \int f(d\mathbb{Q}-d\mathbb{P})}_{W_\rho(\mathbb{P},\mathbb{Q})} = \inf_{\mathbb{M}} \int_{\mathcal{X}\times\mathcal{X}} \rho(x,x')\,d\mathbb{M}(x,x') = \inf_{\mathbb{M}} \mathbb{E}_{\mathbb{M}}\left[\rho(X,X')\right] \tag{2.35}$$

where the infimum ranges over all couplings $\mathbb{M}$ of the pair $(\mathbb{P},\mathbb{Q})$. This coupling-based representation of the Wasserstein distance plays an important role in many of the proofs to follow.

The term "transportation cost" arises from the following interpretation of coupling-based representation (2.35). For concreteness, let us consider the case where $\mathbb{P}$ and $\mathbb{Q}$ have densities p and q with respect to Lebesgue measure on $\mathcal{X}$, and the coupling $\mathbb{M}$ has density m with respect to Lebesgue measure on the product space. The density p can be viewed as describing some initial distribution of mass over the space $\mathcal{X}$, whereas the density q can be interpreted as some desired distribution of the mass. Our goal is to shift mass so as to transform the initial distribution p to the desired distribution q. The quantity $\rho\left(x, x^{\prime}\right) d x d x^{\prime}$ can be interpreted as the cost of transporting a small increment of mass $d x$ to the new increment $d x^{\prime}$. The joint distribution $m\left(x, x^{\prime}\right)$ is known as a transportation plan, meaning a scheme for shifting mass so that p is transformed to q. Combining these ingredients, we conclude that the transportation cost associated with the plan m is given by

$$\int_{\mathcal{X} \times \mathcal{X}} \rho\left(x, x^{\prime}\right) m\left(x, x^{\prime}\right) d x d x^{\prime}$$

and minimizing over all admissible plans- -that is, those that marginalize down to p and q,

respectively - yields the Wasserstein distance.

Transportation cost and concentration inequalities

Let us now turn to the notion of a transportation cost inequality, and its implications for the concentration of measure. Transportation cost inequalities are based on upper bounding the Wasserstein distance $W_\rho(\mathbb{Q}, \mathbb{P})$ in terms of the Kullback-Leibler (KL) divergence.

Definition 2.0.9 — Kullback-Leibler (KL) divergence. Given two distributions $\mathbb{Q}$ and $\mathbb{P}$, the KL divergence between them is given by

$$D(\mathbb{Q} \| \mathbb{P}):= \begin{cases}\mathbb{E}_{\mathbb{Q}}\left[\log \frac{d \mathbb{Q}}{d \mathbb{P}}\right] & \text{when } \mathbb{Q} \text{ is absolutely continuous with respect to } \mathbb{P} \\ +\infty & \text{otherwise}\end{cases} \tag{2.36}$$

If the measures have densities with respect to some underlying measure $v-$ say q and $p-$ then the Kullback-Leibler divergence can be written in the form

$$D(\mathbb{Q} \| \mathbb{P})=\int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x)} v(d x) \tag{2.37}$$

Although the KL divergence provides a measure of distance between distributions, it is not actually a metric (since, for instance, it is not symmetric in general).

We say that a transportation cost inequality is satisfied when the Wasserstein distance is upper bounded by a multiple of the square-root KL divergence.

Definition 2.0.10 — ρ -transportation cost inequality. For a given metric ρ, the probability measure $\mathbb{P}$ is said to satisfy a ρ -transportation cost inequality with parameter $\gamma > 0$ if

$$W_\rho(\mathbb{Q},\mathbb{P}) \leq \sqrt{2\gamma D(\mathbb{Q}\|\mathbb{P})} \tag{2.38}$$

for all probability measures $\mathbb{Q}$.

Such results are also known as information inequalities, due to the role of the KullbackLeibler divergence in information theory.

■ **Example 2.9** A classical example of an information inequality is the Pinsker-Csiszár-Kullback inequality, which relates the total variation distance with the KL divergence. More precisely, for all probability distributions $\mathbb{P}$ and $\mathbb{Q}$, we have

$$\|\mathbb{P}-\mathbb{Q}\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2}D(\mathbb{Q}\|\mathbb{P})} \tag{2.39}$$

From our development in Example 2.8, this inequality corresponds to a transportation cost inequality, in which $\gamma = 1/4$ and the Wasserstein distance is based on the Hamming norm $\rho(x,x') = \mathbb{I}[x \neq x']$. As will be seen shortly, this inequality can be used to recover the bounded differences inequality, corresponding to a concentration statement for functions that are Lipschitz with respect to the Hamming norm. See Exercise 15.6 in Chapter 15 for the proof of this bound.

By the definition (2.32) of the Wasserstein distance, the transportation cost inequality (2.38) can be used to upper bound the deviation $\int f d\mathbb{Q} - \int f d\mathbb{P}$ in terms of the Kullback-Leibler divergence $D(\mathbb{Q}\|\mathbb{P})$. As shown by the following result, a particular choice of distribution $\mathbb{Q}$ can be used to derive a concentration bound for f under $\mathbb{P}$. In this way, a transportation cost inequality leads to concentration bounds for Lipschitz functions:

Theorem 2.0.7 — From transportation cost to concentration. Consider a metric measure space $(\mathbb{P},\mathcal{X},\rho)$, and suppose that $\mathbb{P}$ satisfies the ρ -transportation cost inequality (2.38). Then its concentration function satisfies the bound

$$\alpha_{\mathbb{P},(\mathcal{X},\rho)}(t) \leq 2\exp\left(-\frac{t^2}{2\gamma}\right) \tag{2.40}$$

Moreover, for any $X \sim \mathbb{P}$ and any L-Lipschitz function $f : \mathcal{X} \to \mathbb{R}$, we have the concentration inequality

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2\exp\left(-\frac{t^2}{2\gamma L^2}\right) \tag{2.41}$$

By Proposition 2.0.4, the bound (2.40) implies that

$$\mathbb{P}\left[\left|f(X)-m_f\right|\geq t\right]\leq 2\exp\left(-\frac{t^2}{2\gamma L^2}\right)$$

where m_f is any median of f. In turn, this bound can be used to establish concentration around the mean, albeit with worse constants than the bound (2.41). (See Exercise 2.14 for details on this equivalence.) In our proof, we make use of separate arguments for the median and mean, so as to obtain sharp constants.

Proof. We begin by proving the bound (2.40). For any set A with $\mathbb{P}[A]\geq 1/2$ and a given $\epsilon>0$, consider the set

$$B:=(A^{\epsilon})^c=\{y\in\mathcal{X}\mid\rho(x,y)\geq\epsilon\quad\forall x\in A\}$$

If $\mathbb{P}(A^{\epsilon})=1$, then the proof is complete, so that we may assume that $\mathbb{P}(B)>0$. By construction, we have $\rho(A,B):=\inf_{x\in A}\inf{}_{y\in B}\rho(x,y)\geq\epsilon$. On the other hand, let $\mathbb{P}_A$ and $\mathbb{P}_B$ denote the distributions of $\mathbb{P}$ conditioned on A and B, and let $\mathbb{M}$ denote any coupling of this pair. Since the marginals of $\mathbb{M}$ are supported on A and B, respectively, we have $\rho(A,B)\leq\int\rho(x,x')\,d\mathbb{M}(x,x')$. Taking the infimum over all couplings, we conclude that $\epsilon\leq\rho(A,B)\leq W_\rho(\mathbb{P}_A,\mathbb{P}_B)$. Now applying the triangle inequality, we have

$$\begin{aligned}\epsilon\leq W_\rho(\mathbb{P}_A,\mathbb{P}_B)\leq W_\rho(\mathbb{P},\mathbb{P}_A)+W_\rho(\mathbb{P},\mathbb{P}_B)&\leq^{(ii)}\sqrt{\gamma D(\mathbb{P}_A\|\mathbb{P})}+\sqrt{\gamma D(\mathbb{P}_B\|\mathbb{P})}\\&\leq^{(iii)}\sqrt{2\gamma}\{D(\mathbb{P}_A\|\mathbb{P})+D(\mathbb{P}_B\|\mathbb{P})\}^{1/2}\end{aligned}$$

where step (ii) follows from the transportation cost inequality, and step (iii) follows from the inequality $(a+b)^2\leq 2a^2+2b^2$. It remains to compute the Kullback-Leibler divergences. For any measurable set C, we have $\mathbb{P}_A(C)=\mathbb{P}(C\cap A)/\mathbb{P}(A)$, so that $D(\mathbb{P}_A\|\mathbb{P})=\log\frac{1}{\mathbb{P}(A)}$. Similarly, we have $D(\mathbb{P}_B\|\mathbb{P})=\log\frac{1}{P(B)}$. Combining the pieces, we conclude that

$$\epsilon^2\leq 2\gamma\{\log(1/\mathbb{P}(A))+\log(1/\mathbb{P}(B))\}=2\gamma\log\left(\frac{1}{\mathbb{P}(A)\mathbb{P}(B)}\right)$$

or equivalently $\mathbb{P}(A)\mathbb{P}(B)\leq\exp\left(-\frac{\epsilon^2}{2\gamma}\right)$. Since $\mathbb{P}(A)\geq 1/2$ and $B=(A^{\epsilon})^c$, we conclude that $\mathbb{P}(A^{\epsilon})\geq 1-2\exp\left(-\frac{\epsilon^2}{2\gamma}\right)$. Since A was an arbitrary set with $\mathbb{P}(A)\geq 1/2$, the bound (2.40)follows.

We now turn to the proof of the concentration statement (2.41) for the mean. If one is not concerned about constants, such a bound follows immediately by combining claim (2.40) with the result of Exercise 2.14. Here we present an alternative proof with the dual goals of obtaining the sharp result and illustrating a different proof technique. Throughout this proof, we use $\mathbb{E}_{\mathbb{Q}}[f]$ and $\mathbb{E}_{\mathbb{P}}[f]$ to denote the mean of the random variable $f(X)$ when $X\sim\mathbb{Q}$ and $X\sim\mathbb{P}$, respectively. We begin by observing that

$$\int f(d\mathbb{Q}-d\mathbb{P})\leq^{(i)}LW_\rho(\mathbb{Q},\mathbb{P})\leq^{(ii)}\sqrt{2L^2\gamma D(\mathbb{Q}\|\mathbb{P})}$$

where step (i) follows from the L -Lipschitz condition on f and the definition (2.32); and step (ii) follows from the information inequality (2.38). For any positive numbers (u,v,λ), we have $\sqrt{2uv} \leq \frac{u}{2}\lambda + \frac{v}{\lambda}$. Applying this inequality with $u = L^2\gamma$ and $v = D(\mathbb{Q}\|\mathbb{P})$ yields

$$\int f(d\mathbb{Q} - d\mathbb{P}) \leq \frac{\lambda\gamma L^2}{2} + \frac{1}{\lambda}D(\mathbb{Q}\|\mathbb{P}) \tag{2.42}$$

valid for all $\lambda > 0$. Now define a distribution $\mathbb{Q}$ with Radon-Nikodym derivative $\frac{d\mathbb{Q}}{d\mathbb{P}}(x) = e^{g(x)}/\mathbb{E}_{\mathbb{P}}\left[e^{g(X)}\right]$ where $g(x) := \lambda\left(f(x) - \mathbb{E}_P(f)\right) - \frac{L^2\gamma\lambda^2}{2}$. (Note that our proof of the bound (2.41) ensures that $\mathbb{E}_{\mathbb{P}}\left[e^{g(X)}\right]$ exists.) With this choice, we have

$$D(\mathbb{Q}\|\mathbb{P}) = \mathbb{E}_{\mathbb{Q}}\log\left(\frac{e^{g(X)}}{\mathbb{E}_{\mathbb{P}}\left[e^{g(X)}\right]}\right) = \lambda\left\{\mathbb{E}_{\mathbb{Q}}(f(X)) - \mathbb{E}_{\mathbb{P}}(f(X))\right\} - \frac{\gamma L^2\lambda^2}{2} - \log\mathbb{E}_{\mathbb{P}}\left[e^{g(X)}\right]$$

Combining with inequality (2.42) and performing some algebra (during which the reader should recall that $\lambda > 0$), we find that $\log\mathbb{E}_{\mathbb{P}}\left[e^{g(X)}\right] \leq 0$, or equivalently

$$\mathbb{E}_{\mathbb{P}}\left[e^{\lambda(f(X) - \mathbb{E}_{\mathbb{P}}[f(X')])}\right] \leq e^{\frac{\lambda^2\gamma L^2}{2}}$$

The upper tail bound thus follows by the Chernoff bound. The same argument can be applied to $-f$, which yields the lower tail bound. ■

Tensorization for transportation cost

Based on Theorem 2.0.7, we see that transportation cost inequalities can be translated into concentration inequalities. Like entropy, transportation cost inequalities behave nicely for product measures, and can be combined in an additive manner. Doing so yields concentration inequalities for Lipschitz functions in the higher-dimensional space. We summarize in the following:

Proposition 2.0.8 Suppose that, for each $k = 1,2,\ldots,n$, the univariate distribution $\mathbb{P}_k$ satisfies a ρ_k -transportation cost inequality with parameter γ_k. Then the product distribution $\mathbb{P} = \bigotimes_{k=1}^n \mathbb{P}_k$ satisfies the transportation cost inequality

$$W_\rho(\mathbb{Q},\mathbb{P}) \leq \sqrt{2\left(\sum_{k=1}^n \gamma_k\right) D(\mathbb{Q}\|\mathbb{P})} \quad \text{for all distributions } \mathbb{Q} \tag{2.43}$$

where the Wasserstein metric is defined using the distance $\rho(x,y) := \sum_{k=1}^n \rho_k(x_k,y_k)$

Before turning to the proof of Proposition 2.0.8, it is instructive to see how, in conjunction with Theorem 2.0.7, it can be used to recover the bounded differences inequality.

■ **Example 2.10 — Bounded differences inequality.** Suppose that f satisfies the bounded differences inequality with parameter L_k in coordinate k. Then using the triangle inequality and the bounded differences property, it can be verified that f is a 1 -Lipschitz function with respect to the rescaled Hamming metric

$$\rho(x,y) := \sum_{k=1}^{n} \rho_k(x_k,y_k), \quad \text{where } \rho_k(x_k,y_k) := L_k \mathbb{I}[x_k \neq y_k]$$

By the Pinsker-Csiszár-Kullback inequality (2.39), each univariate distribution $\mathbb{P}_k$ satisfies a ρ_k- transportation cost inequality with parameter $\gamma_k = \frac{L_k^2}{4}$, so that Proposition 2.0.8 implies that $\mathbb{P} = \bigotimes_{k=1}^{n} \mathbb{P}_k$ satisfies a ρ -transportation cost inequality with parameter $\gamma := \frac{1}{4}\sum_{k=1}^{n} L_k^2$. Since f is 1 -Lipschitz with respect to the metric ρ, Theorem 2.0.7 implies that

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2\exp\left(-\frac{2t^2}{\sum_{k=1}^{n} L_k^2}\right) \tag{2.44}$$

In this way, we recover the bounded differences inequality from Chapter 2 from a transportation cost argument.

Our proof of Proposition 2.0.8 is based on the coupling-based characterization (2.35) of Wasserstein distances.

Proof. Letting $\mathbb{Q}$ be an arbitrary distribution over the product space $\mathcal{X}^n$, we construct a coupling $\mathbb{M}$ of the pair $(\mathbb{P},\mathbb{Q})$. For each $j = 2,\ldots,n$, let $\mathbb{M}_1^j$ denote the joint distribution over the pair $\left(X_1^j, Y_1^j\right) = (X_1,\ldots,X_j,Y_1,\ldots,Y_j)$, and let $\mathbb{M}_{j|j-1}$ denote the conditional distribution of (X_j,Y_j) given $\left(X_1^{j-1}, Y_1^{j-1}\right)$. By the dual representation (2.35), we have

$$W_\rho(\mathbb{Q},\mathbb{P}) \leq \mathbb{E}_{\mathbb{M}_1}[\rho_1(X_1,Y_1)] + \sum_{j=2}^{n} \mathbb{E}_{\mathbb{M}_1^{j-1}}\left[\mathbb{E}_{\mathbb{M}_{j|j-1}}[\rho_j(X_j,Y_j)]\right]$$

where $\mathbb{M}_j$ denotes the marginal distribution over the pair (X_j,Y_j). We now define our coupling $\mathbb{M}$ in an inductive manner as follows. First, choose $\mathbb{M}_1$ to be an optimal coupling of the pair $(\mathbb{P}_1,\mathbb{Q}_1)$, thereby ensuring that

$$\mathbb{E}_{\mathbb{M}_1}[\rho_1(X_1,Y_1)] \overset{\text{(i)}}{=} W_\rho(\mathbb{Q}_1,\mathbb{P}_1) \leq^{(ii)} \sqrt{2\gamma_1 D(\mathbb{Q}_1 \| \mathbb{P}_1)}$$

where equality (i) follows by the optimality of the coupling, and inequality (ii) follows from the assumed transportation cost inequality for $\mathbb{P}_1$. Now assume that the joint distribution over $\left(X_1^{j-1}, Y_1^{j-1}\right)$ has been defined. We choose conditional distribution $\mathbb{M}_{j|j-1}(\cdot \mid x_1^{j-1}, y_1^{j-1})$ to be an optimal coupling for the pair $\left(\mathbb{P}_j, \mathbb{Q}_{j|j-1}\left(\cdot \mid y_1^{j-1}\right)\right)$, thereby ensuring that

$$\mathbb{E}_{\mathbb{M}_{j|j-1}}[\rho_j(X_j,Y_j)] \leq \sqrt{2\gamma_j D\left(\mathbb{Q}_{j|j-1}\left(\cdot \mid y_1^{j-1}\right) \| \mathbb{P}_j\right)}$$

valid for each y_1^{j-1}. Taking averages over Y_1^{j-1} with respect to the marginal distribution $\mathbb{M}_1^{j-1}-$ or, equivalently, the marginal $\mathbb{Q}_1^{j-1}-$ the concavity of the square-root function and Jensen's inequality implies that

$$\mathbb{E}_{\mathbb{M}_1^{j-1}}\left[\mathbb{E}_{M_{j|j-1}}[\rho_j(X_j,Y_j)]\right] \leq \sqrt{2\gamma_j \mathbb{E}_{\mathbb{Q}_1^{j-1}} D\left(\mathbb{Q}_{j|j-1}\left(\cdot \mid Y_1^{j-1}\right) \| \mathbb{P}_j\right)}$$

Combining the ingredients, we obtain

$$
\begin{aligned}
W_\rho(\mathbb{Q},\mathbb{P}) &\leq \sqrt{2\gamma_1 D\left(\mathbb{Q}_1 \| \mathbb{P}_1\right)} + \sum_{j=2}^{n} \sqrt{2\gamma_j \mathbb{E}_{\mathbb{Q}_1^{j-1}}\left[D\left(\mathbb{Q}_{j|j-1}\left(\cdot \mid Y_1^{j-1}\right) \| \mathbb{P}_j\right)\right]} \\
&\leq^{(i)} \sqrt{2\left(\sum_{j=1}^{n}\gamma_j\right)} \sqrt{D\left(\mathbb{Q}_1 \| P_1\right) + \sum_{j=2}^{n} \mathbb{E}_{\mathbb{Q}_1^{j-1}}\left[D\left(\mathbb{Q}_{j|j-1}\left(\cdot \mid Y_1^{j-1}\right) \| \mathbb{P}_j\right)\right]} \\
&\overset{(ii)}{=} \sqrt{2\left(\sum_{j=1}^{n}\gamma_j\right) D(\mathbb{Q}\|\mathbb{P})}
\end{aligned}
$$

where step (i) by follows the Cauchy-Schwarz inequality, and equality (ii) uses the chain rule for Kullback-Leibler divergence from Exercise 3.2. ■

In Exercise 3.14, we sketch out an alternative proof of Proposition 2.0.8, one which makes direct use of the Lipschitz characterization of the Wasserstein distance.

Transportation cost inequalities for Markov chains

As mentioned previously, the transportation cost approach has some desirable features in application to Lipschitz functions involving certain types of dependent random variables. Here we illustrate this type of argument for the case of a Markov chain. (See the bibliographic section for references to more general results on concentration for dependent random variables.)

More concretely, let $(X_1,\ldots,X_n)$ be a random vector generated by a Markov chain, where each X_i takes values in a countable space $\mathcal{X}$. Its distribution $\mathbb{P}$ over $\mathcal{X}^n$ is defined by an initial distribution $X_1 \sim \mathbb{P}_1$, and the transition kernels

$$
\mathbb{K}_{i+1}\left(x_{i+1} \mid x_i\right) = \mathbb{P}_{i+1}\left(X_{i+1} = x_{i+1} \mid X_i = x_i\right) \tag{2.45}
$$

Definition 2.0.11 — β -contractive. Here we focus on discrete state Markov chains that are β -contractive, meaning that there exists some $\beta \in [0,1)$ such that

$$
\max_{i=1,\ldots,n-1} \sup_{x_i,x_i'} \left\| \mathbb{K}_{i+1}\left(\cdot \mid x_i\right) - \mathbb{K}_{i+1}\left(\cdot \mid x_i'\right) \right\|_{\mathrm{TV}} \leq \beta \tag{2.46}
$$

where the total variation norm (2.33) was previously defined.

Theorem 2.0.9 Let $\mathbb{P}$ be the distribution of a β -contractive Markov chain (2.46) over the discrete space $\mathcal{X}^n$. Then for any other distribution $\mathbb{Q}$ over $\mathcal{X}^n$, we have

$$
W_\rho(\mathbb{Q},\mathbb{P}) \leq \frac{1}{1-\beta}\sqrt{\frac{n}{2} D(\mathbb{Q}\|\mathbb{P})} \tag{2.47}
$$

where the Wasserstein distance is defined with respect to the Hamming norm $\rho(x,y) =$

$\sum_{i=1}^{n} \mathbb{I}\left[x_i \neq y_i\right]$

R See the bibliography section for references to proofs of this result. Using Theorem 2.0.7, an immediate corollary of the bound (2.47) is that for any function $f : X^n \to \mathbb{R}$ that is L -Lipschitz with respect to the Hamming norm, we have

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2\exp\left(-\frac{2(1-\beta)^2 t^2}{nL^2}\right) \tag{2.48}$$

Note that this result is a strict generalization of the bounded difference inequality for independent random variables, to which it reduces when $\beta = 0$.

■ **Example 2.11 — Parameter estimation for a binary Markov chain.** Consider a Markov chain over binary variables $X_i \in \{0,1\}^2$ specified by an initial distribution $\mathbb{P}_1$ that is uniform, and the transition kernel

$$\mathbb{K}_{i+1}\left(x_{i+1} \mid x_i\right) = \begin{cases} \frac{1}{2}(1+\delta) & \text{if } x_{i+1} = x_i \\ \frac{1}{2}(1-\delta) & \text{if } x_{i+1} \neq x_i \end{cases}$$

where $\delta \in [0,1]$ is a "stickiness" parameter. Suppose that our goal is to estimate the parameter δ based on an n -length vector $(X_1, \ldots, X_n)$ drawn according to this chain. An unbiased estimate of $\frac{1}{2}(1+\delta)$ is given by the function

$$f\left(X_1, \ldots, X_n\right) := \frac{1}{n-1} \sum_{i=1}^{n-1} \mathbb{I}\left[X_i = X_{i+1}\right]$$

corresponding to the fraction of times that successive samples take the same value. We claim that f satisfies the concentration inequality

$$\mathbb{P}\left[\left|f(X) - \frac{1}{2}(1+\delta)\right| \geq t\right] \leq 2e^{-\frac{(n-1)^2(1-\delta)^2 t^2}{2n}} \leq 2e^{-\frac{(n-1)(1-\theta)^2 t^2}{4}} \tag{2.49}$$

Following some calculation, we find that the chain is β -contractive with $\beta = \delta$. Moreover, the function f is $\frac{2}{n-1}$ -lipschitz with respect to the Hamming norm. Consequently, the bound (2.49) follows as a consequence of our earlier general result (2.48).

Asymmetric coupling cost

Thus far, we have considered various types of Wasserstein distances, which can be used to obtain concentration for Lipschitz functions. However, this approach-as with most methods that involve Lipschitz conditions with respect to ℓ_1 -type norms - typically does not yield **dimension-independent bounds**. By contrast, as we have seen previously, Lipschitz conditions based on the ℓ_2 -norm often do lead to dimension-independent results.

With this motivation in mind, this section is devoted to consideration of another type of coupling-based distance between probability distributions, but one that is asymmetric in its two arguments, and of a quadratic nature. In particular, we define

$$C(\mathbb{Q},\mathbb{P}) := \inf_{\mathbb{M}} \sqrt{\int \sum_{i=1}^{n} (\mathbb{M}[Y_i \neq x_i \mid X_i = x_i])^2 \, d\mathbb{P}(x)} \tag{2.50}$$

where once again the infimum ranges over all couplings $\mathbb{M}$ of the pair $(\mathbb{P},\mathbb{Q})$. This distance is relatively closely related to the total variation distance; in particular, it can be shown that an equivalent representation for this asymmetric distance is

$$C(\mathbb{Q},\mathbb{P}) = \sqrt{\int \left|1 - \frac{d\mathbb{Q}}{d\mathbb{P}}(x)\right|_+^2 d\mathbb{P}(x)} \tag{2.51}$$

where $t_+ := \max\{0,t\}$. We leave this equivalence as an exercise for the reader. This representation reveals the close link to the total variation distance, for which

$$\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}} = \int \left|1 - \frac{d\mathbb{Q}}{d\mathbb{P}}\right| d\mathbb{P}(x) = 2\int \left|1 - \frac{d\mathbb{Q}}{d\mathbb{P}}\right|_+ d\mathbb{P}(x)$$

An especially interesting aspect of the asymmetric coupling distance is that it satisfies a Pinsker-type inequality for product distributions. In particular, given any product distribution $\mathbb{P}$ in n variables, we have

$$\max\{C(\mathbb{Q},\mathbb{P}), C(\mathbb{P},\mathbb{Q})\} \leq \sqrt{2D(\mathbb{Q}\|\mathbb{P})} \tag{2.52}$$

for all distributions Q in n dimensions. This deep result is due to Samson; see the bibliographic section for further discussion. While simple to state, it is non-trivial to prove, and has some very powerful consequences for the concentration of convex and Lipschitz functions, as summarized in the following:

Theorem 2.0.10 Consider a vector of independent random variables $(X_1,\ldots,X_n)$, each taking values in $[0,1]$, and let $f : \mathbb{R}^n \to \mathbb{R}$ be convex, and L-Lipschitz with respect to the Euclidean norm. Then for all $t \geq 0$, we have

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2e^{-\frac{t^2}{2L^2}} \tag{2.53}$$

R Note that this is the analog of Theorem 1.0.12— namely, a dimension-independent form of concentration for Lipschitz functions of independent Gaussian variables, but formulated for Lipschitz and convex functions of bounded random variables.

Of course, the same bound also applies to a concave and Lipschitz function. Earlier, we saw that upper tail bounds can obtained under a slightly milder condition, namely that of separate convexity (see Theorem 2.0.3). However, two-sided tail bounds (or concentration inequalities) require these stronger convexity or concavity conditions, as imposed here.

■ **Example 2.12 — Rademacher revisited.** As previously introduced in Example 2.2, the Rademacher complexity of a set $\mathcal{A} \subseteq \mathbb{R}^n$ is defined in terms of the random variable

$$Z \equiv Z(\varepsilon_1, \ldots, \varepsilon_n) := \sup_{a \in \mathcal{A}} \sum_{k=1}^{n} a_k \varepsilon_k$$

where $\{\varepsilon_k\}_{k=1}^n$ is an i.i.d. sequence of Rademacher variables. As shown in Example 2.2, the function $(\varepsilon_1, \ldots, \varepsilon_n) \mapsto Z(\varepsilon_1, \ldots, \varepsilon_n)$ is jointly convex, and Lipschitz with respect to the Euclidean norm with parameter $\mathcal{W}(\mathcal{A}) := \sup_{a \in \mathcal{A}} \|a\|_2$. Consequently, Theorem 2.0.10 implies that

$$\mathbb{P}[|Z - \mathbb{E}[Z]| \geq t] \leq 2 \exp\left(-\frac{t^2}{2\mathcal{W}^2(\mathcal{A})}\right) \tag{2.54}$$

Note that this bound sharpens our earlier inequality (**??**), both in terms of the exponent and in providing a two-sided result.

Let us now prove Theorem 2.0.10

Proof. As defined, any Wasserstein distance immediately yields an upper bound on a quantity of the form $\int f(d\mathbb{Q} - d\mathbb{P})$, where f is a Lipschitz function. Although the asymmetric coupling-based distance is not a Wasserstein distance, the key fact is that it can be used to upper bound such differences when $f : [0,1]^n \to \mathbb{R}$ is Lipschitz and convex. Indeed, for a convex f, we have the lower bound $f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$, which implies that

$$f(y) - f(x) \leq \sum_{j=1}^{n} \left| \frac{\partial f}{\partial y_j}(y) \right| \mathbb{I}\left[x_j \neq y_j\right]$$

Here we have also used the fact that $|x_j - y_j| \leq \mathbb{I}\left[x_j \neq y_j\right]$ for variables taking values in the unit interval [0, 1]. Consequently, for any coupling $\mathbb{M}$ of the pair $(\mathbb{P}, \mathbb{Q})$, we have

$$\begin{aligned} \int f(y) d\mathbb{Q}(y) - \int f(x) d\mathbb{P}(x) &\leq \sum_{j=1}^{n} \left| \frac{\partial f}{\partial y_j}(y) \right| \mathbb{I}\left[x_j \neq y_j\right] d\mathbb{M}(x,y) \\ &= \int \sum_{j=1}^{n} \left| \frac{\partial f}{\partial y_j}(y) \right| \mathbb{M}\left[X_j \neq y_j \mid Y_j = y_j\right] d\mathbb{Q}(y) \\ &\leq \int \|\nabla f(y)\|_2 \sqrt{\sum_{j=1}^{n} \mathbb{M}^2\left[X_j \neq y_j \mid Y_j = y_j\right] d\mathbb{Q}(y)} \end{aligned}$$

where we have applied the Cauchy-Schwarz inequality. By the Lipschitz condition and convexity, we have $\|\nabla f(y)\|_2 \leq L$ almost everywhere, and hence

$$\begin{aligned} \int f(y)d\mathbb{Q}(y) - \int f(x)d\mathbb{P}(x) &\leq L \int \left\{ \sum_{j=1}^{n} \mathbb{M}^2 \left[X_j \neq y_j \mid Y_j = y_j \right] \right\}^{1/2} d\mathbb{Q}(y) \\ &\leq L \left[\int \sum_{j=1}^{n} \mathbb{M}^2 \left[X_j \neq y_j \mid Y_j = y_j \right] d\mathbb{Q}(y) \right]^{1/2} \\ &= LC(\mathbb{P},\mathbb{Q}) \end{aligned}$$

Consequently, the upper tail bound follows by a combination of the information inequality (2.52) and Theorem 2.0.7.

To obtain the lower bound for a convex lipschitz function, it suffices to establish an upper bound for a concave Lipschitz function, say $g : [0,1]^n \to \mathbb{R}$. In this case, we have the upper bound

$$g(y) \leq g(x) + \langle \nabla g(x), y - x \rangle \leq g(x) + \sum_{j=1}^{n} \left| \frac{\partial g(x)}{\partial x_j} \right| \mathbb{I} \left[x_j \neq y_j \right]$$

and consequently

$$\int g d\mathbb{Q}(y) - \int g d\mathbb{P}(x) \leq \sum_{j=1}^{n} \left| \frac{\partial g(x)}{\partial x_j} \right| \mathbb{I} \left[x_j \neq y_j \right] d\mathbb{M}(x,y)$$

The same line of reasoning then shows that $\int g d\mathbb{Q}(y) - \int g d\mathbb{P}(x) \leq LC(\mathbb{Q},\mathbb{P})$, from which the claim then follows as before. ■

We have stated Theorem 2.0.10 for the familiar case of independent random variables. However, a version of the underlying information inequality (2.52) holds for many collections of random variables. In particular, consider an n -dimensional distribution $\mathbb{P}$ for which there exists some $\gamma > 0$ such that the following inequality holds:

$$\max\{C(\mathbb{Q},\mathbb{P}), C(\mathbb{P},\mathbb{Q})\} \leq \sqrt{2\gamma D(\mathbb{Q}\|\mathbb{P})} \quad \text{for all distributions } \mathbb{Q} \tag{2.55}$$

The same proof then shows that any L -Lipschitz function satisfies the concentration inequality

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2\exp\left(-\frac{t^2}{2\gamma L^2}\right) \tag{2.56}$$

For example, for a Markov chain that satisfies the β -contraction condition (2.46), it can be shown that the information inequality (2.55) holds with $\gamma = \left(\frac{1}{1-\sqrt{\beta}}\right)^2$. Consequently, any L-Lipschitz function (with respect to the Euclidean norm) of a β -contractive Markov chain satisfies the concentration inequality

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2\exp\left(-\frac{(1-\sqrt{\beta})^2 t^2}{2L^2}\right) \tag{2.57}$$

This bound is a dimension-independent analog of our earlier bound (2.48) for a contractive Markov chain. We refer the reader to the bibliographic section for further discussion of results of this type.

2.1 Tail bounds for empirical processes

In this section, we illustrate the use of concentration inequalities in application to empirical processes. We encourage the interested reader to look ahead to Chapter 4 so as to acquire the statistical motivation for the classes of problems studied in this section. Here we use the entropy method to derive various tail bounds on the suprema of empirical processes - in particular, for random variables that are generated by taking suprema of sample averages over function classes. More precisely let $\mathscr{F}$ be a class of functions (each of the form $f : \mathcal{X} \rightarrow \mathbb{R}$), and let $(X_1, \ldots, X_n)$ be drawn from a product distribution $\mathbb{P} = \bigotimes_{i=1}^{n} \mathbb{P}_i$, where each $\mathbb{P}_i$ is supported on some set $\mathcal{X}_i \subseteq \mathcal{X}$. We then consider the random variable

$$Z = \sup_{f \in \mathscr{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right\} \tag{2.58}$$

The primary goal of this section is to derive a number of upper bounds on the tail event $\{Z \geq \mathbb{E}[Z] + \delta\}$

As a passing remark, we note that, if the goal is to obtain bounds on the random variable $\sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right|$, then it can be reduced to an instance of the variable (2.58) by considering the augmented function class $\widetilde{\mathscr{F}} = \mathscr{F} \cup \{-\mathscr{F}\}$.

2.1.1 A functional Hoeffding inequality

We begin with the simplest type of tail bound for the random variable Z, namely one of the Hoeffding type. The following result is a generalization of the classical Hoeffding theorem for sums of bounded random variables.

Theorem 2.1.1 — Functional Hoeffding theorem. For each $f \in \mathscr{F}$ and $i = 1, \ldots, n$ assume that there are real numbers $a_{i,f} \leq b_{i,f}$ such that $f(x) \in [a_{i,f}, b_{i,f}]$ for all $x \in \mathcal{X}_i$ Then for all $\delta \geq 0$, we have*e*

$$\mathbb{P}[Z \geq \mathbb{E}[Z] + \delta] \leq \exp\left(-\frac{n\delta^2}{4L^2}\right) \tag{2.59}$$

where $L^2 := \sup_{f \in \mathscr{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (b_{i,f} - a_{i,f})^2 \right\}$

R In a very special case, Theorem 2.1.1 can be used to recover the classical Hoeffding inequality in the case of bounded random variables, albeit with a slightly worse constant. Indeed, if we let $\mathscr{F}$ be a singleton consisting of the identity function $f(x) =$

x, then we have $Z = \frac{1}{n}\sum_{i=1}^{n} X_i$. Consequently, as long as $x_i \in [a_i, b_i]$, Theorem 2.1.1 implies that

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mathbb{E}\left[X_i\right]\right) \geq \delta\right] \leq e^{-\frac{n\delta^2}{4L^2}}$$

where $L^2 = \frac{1}{n}\sum_{i=1}^{n}\left(b_i - a_i\right)^2$. We thus recover the classical Hoeffding theorem, although the constant $1/4$ in the exponent is not optimal.

More substantive implications of Theorem 2.1.1 arise when it is applied to a larger function class $\mathscr{F}$. In order to appreciate its power, let us compare the upper tail bound (2.59) to the corresponding bound that can be derived from the bounded differences inequality, as applied to the function $\left(x_1, \ldots, x_n\right) \mapsto Z\left(x_1, \ldots, x_n\right)$. With some calculation, it can be seen that this function satisfies the bounded difference inequality with constant $L_i := \sup_{f \in \mathscr{F}}\left|b_{i,f} - a_{i,f}\right|$ in coordinate i. Consequently, the bounded differences method (Corollary 1.0.11) yields a sub-Gaussian tail bound, analogous to the bound (2.59), but with the parameter

$$\widetilde{L}^2 = \frac{1}{n}\sum_{i=1}^{n}\sup_{f \in \mathscr{F}}\left(b_{i,f} - a_{i,f}\right)^2$$

Note that the quantity $\widetilde{L}-$ since it is defined by applying the supremum separately to each coordinate - can be substantially larger than the constant L defined in the theorem statement.

It suffices to prove the result for a finite class of functions $\mathscr{F}$; the general result can be recovered by taking limits over an increasing sequence of such finite classes. Let us view Z as a function of the random variables $\left(X_1, \ldots, X_n\right)$. For each index $j = 1, \ldots, n$, define the random function

$$x_j \mapsto Z_j\left(x_j\right) = Z\left(X_1, \ldots, X_{j-1}, x_j, X_{j+1}, \ldots, X_n\right)$$

In order to avoid notational clutter, we work throughout this proof with the unrescaled version of Z, namely $Z = \sup_{f \in \mathscr{F}}\sum_{i=1}^{n} f\left(X_i\right)$. Combining the tensorization Lemma 2.2 with the bound (2.14a) from Lemma 2.1, we obtain

$$\mathbb{H}\left(e^{\lambda Z(X)}\right) \leq \lambda^2 \mathbb{E}\left[\sum_{j=1}^{n}\mathbb{E}\left[\left(Z_j\left(X_j\right) - Z_j\left(Y_j\right)\right)^2 \mathbb{I}\left[Z_j\left(X_j\right) \geq Z_j\left(Y_j\right)\right] e^{\lambda Z(X)} \mid X^{\backslash j}\right]\right] \tag{2.60}$$

For each $f \in \mathscr{F}$, define the set $\mathcal{A}(f) := \left\{\left(x_1, \ldots, x_n\right) \in \mathbb{R}^n \mid Z = \sum_{i=1}^{n} f\left(x_i\right)\right\}$, corresponding to the set of realizations for which the maximum defining Z is achieved by f. (If there are ties, then we resolve them arbitrarily so as to make the sets $\mathcal{A}(f)$ disjoint.) For any $x \in \mathcal{A}(f)$, we have

$$Z_j\left(x_j\right) - Z_j\left(y_j\right) = f\left(x_j\right) + \sum_{i \neq j}^{n} f\left(x_i\right) - \max_{\tilde{f} \in \mathscr{F}} \left\{ \tilde{f}\left(y_j\right) + \sum_{i \neq j}^{n} \tilde{f}\left(x_i\right) \right\} \leq f\left(x_j\right) - f\left(y_j\right)$$

As long as $Z_j\left(x_j\right) \geq Z_j\left(y_j\right)$, this inequality still holds after squaring both sides. Considering all possible sets $\mathcal{A}(f)$, we arrive at the upper bound

$$\left(Z_j\left(x_j\right) - Z_j\left(y_j\right)\right)^2 \mathbb{I}\left[Z_j\left(x_j\right) \geq Z_j\left(y_j\right)\right] \leq \sum_{f \in \mathscr{F}} \mathbb{I}[x \in \mathcal{A}(f)] \left(f\left(x_j\right) - f\left(y_j\right)\right)^2 \tag{2.61}$$

Since $\left(f\left(x_j\right) - f\left(y_j\right)\right)^2 \leq \left(b_{j,f} - a_{j,f}\right)^2$ by assumption, summing over the indices j yields

$$\begin{aligned}
\sum_{j=1}^{n} \left(Z_j\left(x_j\right) - Z_j\left(y_j\right)\right)^2 \mathbb{I}\left[Z_k\left(x_k\right) \geq Z_k\left(y_k\right)\right] e^{\lambda Z(x)} &\leq \sum_{h \in \mathscr{F}} \mathbb{I}[x \in \mathcal{A}(h)] \sum_{k=1}^{n} \left(b_{k,h} - a_{k,h}\right)^2 e^{\lambda Z(x)} \\
&\leq \sup_{f \in \mathscr{F}} \sum_{j=1}^{n} \left(b_{j,f} - a_{j,f}\right)^2 e^{\lambda Z(x)} \\
&= nL^2 e^{\lambda Z(x)}
\end{aligned}$$

Substituting back into our earlier inequality (3.81), we find that

$$\mathbb{H}\left(e^{\lambda Z(X)}\right) \leq nL^2 \lambda^2 \mathbb{E}\left[e^{\lambda Z(X)}\right]$$

This is a sub-Gaussian entropy bound (2.4) with $\sigma = \sqrt{2n}L$, so that Proposition 2.0.1 implies that the unrescaled version of Z satisfies the tail bound

$$\mathbb{P}[Z \geq \mathbb{E}[Z] + t] \leq e^{-\frac{t^2}{4nL^2}}$$

Setting $t = n\delta$ yields the claim (2.59) for the rescaled version of Z.

A functional Bernstein inequality

In this section, we turn to the Bernstein refinement of the functional Hoeffding inequality from Theorem 2.1.1. As opposed to control only in terms of bounds on the function values, it also brings a notion of variance into play. As will be discussed at length in later chapters, this type of variance control plays a key role in obtaining sharp bounds for various types of statistical estimators.

Theorem 2.1.2 — Talagrand concentration for empirical processes. Consider a countable class of functions $\mathscr{F}$ uniformly bounded by b. Then for all $\delta > 0$, the random variable (2.58) satisfies the upper tail bound

$$\mathbb{P}[Z \geq \mathbb{E}[Z] + \delta] \leq 2\exp\left(\frac{-n\delta^2}{8e\mathbb{E}\left[\Sigma^2\right] + 4b\delta}\right) \tag{2.62}$$

where $\Sigma^2 = \sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} f^2\left(X_i\right)$

In order to obtain a simpler bound, the expectation $\mathbb{E}\left[\Sigma^2\right]$ can be upper bounded. Using symmetrization techniques to be developed in Chapter 4, it can be shown that

$$\mathbb{E}\left[\Sigma^2\right] \leq \sigma^2 + 2b\mathbb{E}[Z] \tag{2.63}$$

where $\sigma^2 = \sup_{f\in\mathscr{F}} \mathbb{E}\left[f^2(X)\right]$. Using this upper bound on $\mathbb{E}\left[\Sigma^2\right]$ and performing some algebra, we obtain that there are universal positive constants (c_0, c_1) such that

$$\mathbb{P}\left[Z \geq \mathbb{E}[Z] + c_0\gamma\sqrt{t} + c_1 bt\right] \leq e^{-nt} \quad \text{for all } t > 0 \tag{2.64}$$

where $\gamma^2 = \sigma^2 + 2b\mathbb{E}[Z]$. See Exercise 3.16 for the derivation of this inequality from Theorem 2.1.2 and the upper bound (3.84). Although the proof outlined here leads to poor constants, the best known are $c_0 = \sqrt{2}$ and $c_1 = 1/3$; see the bibliographic section for further details.

In certain settings, it can be useful to exploit the bound (2.64) in an alternative form: in particular, for any $\epsilon > 0$, it implies the upper bound

$$\mathbb{P}\left[Z \geq (1+\epsilon)\mathbb{E}[Z] + c_0\sigma\sqrt{t} + \left(c_1 + c_0^2/\epsilon\right) bt\right] \leq e^{-nt} \tag{2.65}$$

Conversely, we can recover the tail bound (2.64) by optimizing over $\epsilon > 0$ in the family of bounds (2.65); see Exercise 3.16 for the details of this equivalence.

We assume without loss of generality that $b = 1$, since the general case can be reduced to this one. Moreover, as in the proof of Theorem 2.1.1, we work with the unrescaled version - namely, the variable $Z = \sup_{f\in\mathscr{F}} \sum_{i=1}^n f(X_i)$ -and then translate our results back. Recall the definition of the sets $\mathcal{A}(f)$, and the upper bound (2.61) from the previous proof; substituting it into the entropy bound (2.60) yields the upper bound

$$\mathbb{H}\left(e^{\lambda Z}\right) \leq \lambda^2\mathbb{E}\left[\sum_{j=1}^n \mathbb{E}\left[\sum_{f\in\mathscr{F}} \mathbb{I}[x \in \mathcal{A}(f)]\left(f(X_j) - f(Y_j)\right)^2 e^{\lambda Z} \mid X^{\backslash j}\right]\right]$$

Now we have

$$\begin{aligned}\sum_{i=1}^n \sum_{f\in\mathscr{F}} \mathbb{I}[X \in \mathcal{A}(f)]\left(f(X_j) - f(Y_j)\right)^2 &\leq 2\sup_{f\in\mathscr{F}}\sum_{i=1}^n f^2(X_i) + 2\sup_{f\in\mathscr{F}}\sum_{i=1}^n f^2(Y_i) \\ &= 2\{\Gamma(X) + \Gamma(Y)\}\end{aligned}$$

where $\Gamma(X) := \sup_{f\in\mathscr{F}}\sum_{i=1}^n f^2(X_i)$ is the unrescaled version of Σ^2. Combined with our earlier inequality, we see that the entropy satisfies the upper bound

$$\mathbb{H}\left(e^{\lambda Z}\right) \leq 2\lambda^2\left\{\mathbb{E}\left[\Gamma e^{\lambda Z}\right] + \mathbb{E}[\Gamma]\mathbb{E}\left[e^{\lambda Z}\right]\right\} \tag{2.66}$$

From the result of Exercise 3.4, we have $\mathbb{H}\left(e^{\lambda(Z+c)}\right) = e^{\lambda c}\mathbb{H}\left(e^{\lambda Z}\right)$ for any constant $c \in \mathbb{R}$. since the right-hand side also contains a term $e^{\lambda Z}$ in each component, we see that the same upper bound holds for $\mathbb{H}\left(e^{\lambda\widetilde{Z}}\right)$, where $\widetilde{Z} = Z - \mathbb{E}[Z]$ is the centered version. We now introduce a lemma to control the term $\mathbb{E}\left[\Gamma e^{\lambda\widetilde{Z}}\right]$.

Lemma 2.3 — Controlling the random variance. For all $\lambda > 0$, we have

$$\mathbb{E}\left[\Gamma e^{\lambda \widetilde{Z}}\right] \leq (e-1)\mathbb{E}[\Gamma]\mathbb{E}\left[e^{\lambda \widetilde{Z}}\right] + \mathbb{E}\left[\widetilde{Z} e^{\lambda \widetilde{Z}}\right] \tag{2.67}$$

Combining the upper bound (2.67) with the entropy upper bound (2.66) for $\widetilde{Z}$, we obtain

$$\mathbb{H}\left(e^{\lambda \widetilde{Z}}\right) \leq \lambda^2 \left\{2e\mathbb{E}[\Gamma]\varphi(\lambda) + 2\varphi'(\lambda)\right\} \quad \text{for all } \lambda > 0$$

where $\varphi(\lambda) := \mathbb{E}\left[e^{\lambda \widetilde{Z}}\right]$ is the moment generating function of $\widetilde{Z}$. Since $\mathbb{E}[\widetilde{Z}] = 0$, we recognize this as an entropy bound of the Bernstein form (2.8) with $b = 2$ and $\sigma^2 = 2e\mathbb{E}[\Gamma]$.

Consequently, by the consequence (**??**) stated following Proposition 3.3, we conclude that

$$\mathbb{P}[\widetilde{Z} \geq \mathbb{E}[\widetilde{Z}] + \delta] \leq \exp\left(-\frac{\delta^2}{8e\mathbb{E}[\Gamma] + 4\delta}\right) \quad \text{for all } \delta \geq 0$$

Recalling the definition of Γ and rescaling by $1/n$, we obtain the stated claim of the theorem with $b = 1$

It remains to prove Lemma 2.3. Consider the function $g(t) = e^t$ with conjugate dual $g^*(s) = s\log s - s$ for $s > 0$. By the definition of conjugate duality (also known as Young's inequality), we have $st \leq s\log s - s + e^t$ for all $s > 0$ and $t \in \mathbb{R}$. Applying this inequality with $s = e^{\lambda \widetilde{Z}}$ and $t = \Gamma - (e-1)\mathbb{E}[\Gamma]$ and then taking expectations, we find that

$$\mathbb{E}\left[\Gamma e^{\lambda \widetilde{Z}}\right] - (e-1)\mathbb{E}\left[e^{\lambda \widetilde{Z}}\right]\mathbb{E}[\Gamma] \leq \lambda \mathbb{E}\left[\widetilde{Z} e^{\lambda \widetilde{Z}}\right] - \mathbb{E}\left[e^{\lambda \widetilde{Z}}\right] + \mathbb{E}\left[e^{\Gamma - (e-1)\mathbb{E}[\Gamma]}\right]$$

Note that Γ is defined as a supremum of a class of functions taking values in [0,1] . Therefore, by the result of Exercise 3.15, we have $\mathbb{E}\left[e^{\Gamma-(e-1)\mathbb{E}[\Gamma]}\right] \leq 1$. Moreover, by Jensen's inequality, we have $\mathbb{E}\left[e^{\lambda \widetilde{Z}}\right] \geq e^{\lambda \mathbb{E}[\widetilde{Z}]} = 1$. Putting together the pieces yields the claim (2.67)

3. Uniform laws of large numbers

The focus of this chapter is a class of results known as uniform laws of large numbers. As suggested by their name, these results represent a strengthening of the usual law of large numbers, which applies to a fixed sequence of random variables, to related laws that hold uniformly over collections of random variables. On one hand, such uniform laws are of theoretical interest in their own right, and represent an entry point to a rich area of probability and statistics known as empirical process theory. On the other hand, uniform laws also play a key role in more applied settings, including understanding the behavior of different types of statistical estimators. The classical versions of uniform laws are of an asymptotic nature, whereas more recent work in the area has emphasized non-asymptotic results. Consistent with the overall goals of this book, this chapter will follow the non-asymptotic route, presenting results that apply to all sample sizes. In order to do so, we make use of the tail bounds and the notion of Rademacher complexity previously introduced in Chapter 2 .

Motivation

We begin with some statistical motivations for deriving laws of large numbers, first for the case of cumulative distribution functions and then for more general function classes.

Uniform convergence of cumulative distribution functions

The law of any scalar random variable X can be fully specified by its cumulative distribution function (CDF), whose value at any point $t \in \mathbb{R}$ is given by $F(t) := \mathbb{P}[X \leq t]$. Now suppose that we are given a collection $\{X_i\}_{i=1}^{n}$ of n i.i.d. samples, each drawn according to the law specified by F. A natural estimate of F is the empirical CDF given by

$$\widehat{F}_n(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(-\infty,t]}\left[X_i\right] \tag{3.1}$$

where $\mathbb{I}_{(-\infty,t]}[x]$ is a {0,1} -valued indicator function for the event $\{x \leq t\}$. Since the population CDF can be written as $F(t) = \mathbb{E}\left[\mathbb{I}_{(-\infty,t]}[X]\right]$, the empirical CDF is an unbiased estimate. Note that $\widehat{F}_n$ is a random function, with the value $\widehat{F}_n(t)$ corresponding to the fraction of samples that lie in the interval $(-\infty, t]$. As the sample size n grows, we see that $\widehat{F}_n$ approaches F. It is easy to see that $\widehat{F}_n$ converges to F in a pointwise sense. Indeed, for any fixed $t \in \mathbb{R}$, the random variable $\widehat{F}_n(t)$ has mean $F(t)$, and moments of all orders, so that the strong law of large numbers implies that $\widehat{F}_n(t) \xrightarrow{a.s.} F(t)$. A natural goal is to strengthen this point-wise convergence to a form of uniform convergence. Why are uniform convergence results interesting and important? In statistical settings, a typical use of the empirical CDF is to construct estimators of various quantities associated with the population CDF. Many such estimation problems can be formulated in a terms of functional γ that maps any CDFF to a real number $\gamma(F)-$ that is, $F \mapsto \gamma(F)$. Given a set of samples distributed according to F, the plug-in principle suggests replacing the unknown F with the empirical CDF $\widehat{F}_n$, thereby obtaining $\gamma\left(\widehat{F}_n\right)$ as an estimate of $\gamma(F)$. Let us illustrate this procedure via some examples.

■ **Example 3.1 — Expectation functionals.** Given some integrable function g, we may define the expectation functional γ_g via

$$\gamma_g(F) := \int g(x) dF(x) \tag{3.2}$$

For instance, for the function $g(x) = x$, the functional γ_g maps$F$ to $\mathbb{E}[X]$, where X is a random variable with CDF F. For any g, the plug-in estimate is given by $\gamma_g\left(\widehat{F}_n\right) = \frac{1}{n}\sum_{i=1}^{n} g\left(X_i\right)$ corresponding to the sample mean of $g(X)$. In the special case $g(x) = x$, we recover the usual sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ as an estimate for the mean $\mu = \mathbb{E}[X]$. A similar interpretation applies to other choices of the underlying function g.

■ **Example 3.2 — Quantile functionals.** For any $\alpha \in [0,1]$, the quantile functional Q_α is given by

$$Q_\alpha(F) := \inf\{t \in \mathbb{R} \mid F(t) \geq \alpha\} \tag{3.3}$$

The median corresponds to the special case $\alpha = 0.5$. The plug-in estimate is given by

$$Q_\alpha\left(\widehat{F}_n\right) := \inf\left\{t \in \mathbb{R} \mid \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(-\infty,t]}\left[X_i\right] \geq \alpha\right\} \tag{3.4}$$

and corresponds to estimating the α th quantile of the distribution by the α th sample quantile. In the special case $\alpha = 0.5$, this estimate corresponds to the sample median.

Again, it is of interest to determine in what sense (if any) the random variable $Q_\alpha\left(\widehat{F}_n\right)$ approaches $Q_\alpha(F)$ as n becomes large. In this case, $Q_\alpha\left(\widehat{F}_n\right)$ is a fairly complicated, non-linear function of all the variables, so that this convergence does not follow immediately by a classical result such as the law of large numbers.

■ **Example 3.3 — Goodness-of-fit functionals.** It is frequently of interest to test the hypothesis of whether or not a given set of data has been drawn from a known distribution F_0. For instance, we might be interested in assessing departures from uniformity, in which case F_0 would be a uniform distribution on some interval, or departures from Gaussianity, in which case F_0 would specify a Gaussian with a fixed mean and variance. Such tests can be performed using functionals that measure the distance between F and the target CDF F_0 including the sup-norm distance $\|F - F_0\|_\infty$, or other distances such as the Cramér-von Mises criterion based on the functional $\gamma(F) := \int_{-\infty}^{\infty} [F(x) - F_0(x)]^2 dF_0(x)$

For any plug-in estimator $\gamma\left(\widehat{F}_n\right)$, an important question is to understand when it is consistent- - that is, when does $\gamma\left(\widehat{F}_n\right)$ converge to $\gamma(F)$ in probability (or almost surely)? This question can be addressed in a unified manner for many functionals by defining a notion of continuity.

Definition 3.0.1 — Continuity. Given a pair of CDFs F and G, let us measure the distance between them using the sup-norm

$$\|G - F\|_\infty := \sup_{t \in \mathbb{R}} |G(t) - F(t)| \tag{3.5}$$

We can then define the continuity of a functional γ with respect to this norm: more precisely, we say that the functional γ is continuous at F in the sup-norm if, for all $\epsilon > 0$, there exists a $\delta > 0$ such that $\|G - F\|_\infty \leq \delta$ implies that $|\gamma(G) - \gamma(F)| \leq \epsilon$.

As we explore in Exercise 4.1, this notion is useful, because for any continuous functional, it reduces the consistency question for the plug-in estimator $\gamma\left(\widehat{F}_n\right)$ to the issue of whether or not the random variable $\left\|\widehat{F}_n - F\right\|_\infty$ converges to zero. A classical result, known as the Glivenko-Cantelli theorem, addresses the latter question:

Theorem 3.0.1 — Glivenko-Cantelli. For any distribution, the empirical CDF $\widehat{F}_n$ is a strongly consistent estimator of the population CDF in the uniform norm, meaning that

$$\left\|\widehat{F}_n - F\right\|_\infty \xrightarrow{a.s.} 0 \tag{3.6}$$

We provide a proof of this claim as a corollary of a more general result to follow (see Theorem 3.1.1). For statistical applications, an important consequence of Theorem 3.0.1 is that the plug-in estimate $\gamma\left(\widehat{F}_n\right)$ is almost surely consistent as an estimator of $\gamma(F)$ for any functional γ that is continuous with respect to the sup-norm. See Exercise 4.1 for further exploration of this connection.

Uniform laws for more general function classes

We now turn to more general consideration of uniform laws of large numbers. Let $\mathscr{F}$ be a class of integrable real-valued functions with domain $\mathcal{X}$, and let $\{X_i\}_{i=1}^n$ be a collection of i.i.d. samples from some distribution $\mathbb{P}$ over $\mathcal{X}$. Consider the random variable

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} := \sup_{f \in \mathscr{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right| \tag{3.7}$$

which measures the absolute deviation between the sample average $\frac{1}{n}\sum_{i=1}^n f(X_i)$ and the population average $\mathbb{E}[f(X)]$, uniformly over the class $\mathscr{F}$. Note that there can be measurability concerns associated with the definition (3.7); see the bibliographic section for discussion of different ways in which to resolve them.

Definition 3.0.2 — Glivenko-Cantelli class. We say that $\mathscr{F}$ is a Glivenko-Cantelli class for $\mathbb{P}$ if $\|\mathbb{P}_n - \mathbb{P}\| \mathscr{F}$ converges to zero in probability as $n \to \infty$

This notion can also be defined in a stronger sense, requiring almost sure convergence of $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$, in which case we say that $\mathscr{F}$ satisfies a strong Glivenko-Cantelli law. The classical result on the empirical CDF (Theorem 3.0.1) can be reformulated as a particular case of this notion:

■ **Example 3.4 — Empirical CDFs and indicator functions.** Consider the function class

$$\mathscr{F} = \left\{ \mathbb{I}_{(-\infty,t]}(\cdot) \mid t \in \mathbb{R} \right\} \tag{3.8}$$

where $\mathbb{I}_{(-\infty,t]}$ is the {0,1} -valued indicator function of the interval $(-\infty, t]$. For each fixed $t \in \mathbb{R}$, we have the equality $\mathbb{E}\left[\mathbb{I}_{(-\infty,t]}(X)\right] = \mathbb{P}[X \leq t] = F(t)$, so that the classical Glivenko-Cantelli theorem is equivalent to a strong uniform law for the class (3.8).

Not all classes of functions are Glivenko-Cantelli, as illustrated by the following example.

■ **Example 3.5 — Failure of uniform law.** Let $\mathcal{S}$ be the class of all subsets S of [0,1] such that the subset S has a finite number of elements, and consider the function class $\mathscr{F}_S = \{\mathbb{I}_S(\cdot) \mid S \in \mathcal{S}\}$ of $(\{0-1\}-$ valued $)$ indicator functions of such sets. Suppose that samples X_i are drawn from some distribution over [0,1] that has no atoms (i.e., $\mathbb{P}(\{x\}) = 0$ for all $x \in [0,1]$); this class includes any distribution that has a density with respect to Lebesgue measure. For any such distribution, we are guaranteed that $\mathbb{P}[S] = 0$ for all $S \in \mathcal{S}$. On the other hand, for any positive integer $n \in \mathbb{N}$, the discrete set $\{X_1, \ldots, X_n\}$ belongs to $\mathcal{S}$, and moreover, by definition of the empirical distribution, we have $\mathbb{P}_n[X_1^n] = 1$.

Putting together the pieces, we conclude that

$$\sup_{S\in\mathcal{S}}|\mathbb{P}_n[S]-\mathbb{P}[S]|=1-0=1$$

so that the function class $\mathscr{F}_S$ is not a Glivenko-Cantelli class for $\mathbb{P}$.

We have seen that the classical Glivenko-Cantelli law-which guarantees convergence of a special case of the variable $\|\mathbb{P}_n-\mathbb{P}\|_{\mathscr{F}}-$ is of interest in analyzing estimators based on "plugging in" the empirical CDF. It is natural to ask in what other statistical contexts do these quantities arise? In fact, variables of the form $\|\mathbb{P}_n-\mathbb{P}\|_{\mathscr{F}}$ are ubiquitous throughout statistics - in particular, they lie at the heart of methods based on empirical risk minimization. In order to describe this notion more concretely, let us consider an indexed family of probability distributions $\{\mathbb{P}_\theta \mid \theta\in\Omega\}$, and suppose that we are given n samples $\{X_i\}_{i=1}^n$, each sample lying in some space $\mathcal{X}$. Suppose that the samples are drawn i.i.d. according to a distribution $\mathbb{P}_{\theta^*}$, for some fixed but unknown $\theta^*\in\Omega$. Here the index θ^* could lie within a finite-dimensional space, such as $\Omega=\mathbb{R}^d$ in a vector estimation problem, or could lie within some function class $\Omega=\mathscr{G}$, in which case the problem is of the nonparametric variety. In either case, a standard decision-theoretic approach to estimating θ^* is based on minimizing a cost function of the form $\theta\mapsto\mathcal{L}_\theta(X)$, which measures the "fit" between a parameter $\theta\in\Omega$ and the sample $X\in X$. Given the collection of n samples $\{X_i\}_{i=1}^n$, the principle of empirical risk minimization is based on the objective function

$$\widehat{R}_n\left(\theta,\theta^*\right):=\frac{1}{n}\sum_{i=1}^n\mathcal{L}_\theta\left(X_i\right)$$

This quantity is known as the empirical risk, since it is defined by the samples X_1^n, and our notation reflects the fact that these samples depend-in turn-on the unknown distribution $\mathbb{P}_{\theta^*}$. This empirical risk should be contrasted with the population risk,

$$R\left(\theta,\theta^*\right):=\mathbb{E}_{\theta^*}\left[\mathcal{L}_\theta(X)\right]$$

where the expectation $\mathbb{E}_{\theta^*}$ is taken over a sample $X\sim\mathbb{P}_{\theta^*}$. In practice, one minimizes the empirical risk over some subset Ω_0 of the full space Ω, thereby obtaining some estimate $\widehat{\theta}$. The statistical question is how to bound the excess risk, measured in terms of the population quantities- -namely the difference

$$E\left(\widehat{\theta},\theta^*\right):=R\left(\widehat{\theta},\theta^*\right)-\inf_{\theta\in\Omega_0}R\left(\theta,\theta^*\right)$$

Let us consider some examples to illustrate.

■ **Example 3.6 — Maximum likelihood.** Consider a parameterized family of distributions-say $\{\mathbb{P}_\theta,\theta\in\Omega\}$ -each with a strictly positive density p_θ defined with respect to a common underlying measure. Now suppose that we are given n i.i.d. samples from an unknown distribution $\mathbb{P}_{\theta^*}$, and we would like to estimate the unknown parameter θ^*. In order to do so, we consider the cost function

$$\mathcal{L}_\theta(x) := \log\left[\frac{p_{\theta^*}(x)}{p_\theta(x)}\right]$$

The term $p_{\theta^*}(x)$, which we have included for later theoretical convenience, has no effect on the minimization over θ. Indeed, the maximum likelihood estimate is obtained by minimizing the empirical risk defined by this cost function - that is

$$\widehat{\theta} \in \arg\min_{\theta\in\Omega_0} \underbrace{\left\{\frac{1}{n}\sum_{i=1}^n \log\frac{p_{\theta^*}(X_i)}{p_\theta(X_i)}\right\}}_{\widehat{R}_n(\theta,\theta^*)} = \arg\min_{\theta\in\Omega_0}\left\{\frac{1}{n}\sum_{i=1}^n \log\frac{1}{p_\theta(X_i)}\right\}$$

The population risk is given by $R(\theta,\theta^*) = \mathbb{E}_{\theta^*}\left[\log\frac{p_{\theta^*}(X)}{p_\theta(X)}\right]$, a quantity known as the Kullback-Leibler divergence between p_{θ^*} and p_θ. In the special case that $\theta^* \in \Omega_0$, the excess risk is simply the Kullback-Leibler divergence between the true density p_{θ^*} and the fitted model $p_{\widehat{\theta}}$. See Exercise 4.3 for some concrete examples.

■ **Example 3.7 — Binary classification.** Suppose that we observe n pairs of samples, each of the form $(X_i,Y_i) \in \mathbb{R}^d \times \{-1,+1\}$, where the vector X_i corresponds to a set of d predictors or features, and the binary variable Y_i corresponds to a label. We can view such data as being generated by some distribution $\mathbb{P}_X$ over the features, and a conditional distribution $\mathbb{P}_{Y|X}$ since Y takes binary values, the conditional distribution is fully specified by the likelihood ratio $\psi(x) = \frac{\mathbb{P}[Y=+1|X=x]}{\mathbb{P}[Y=-1|X=x]}$

The goal of binary classification is to estimate a function $f : \mathbb{R}^d \to \{-1,+1\}$ that minimizes the probability of misclassification $\mathbb{P}[f(X) \neq Y]$, for an independently drawn pair (X,Y). Note that this probability of error corresponds to the population risk for the cost function

$$\mathcal{L}_f(X,Y) := \begin{cases} 1 & \text{if } f(X) \neq Y \\ 0 & \text{otherwise} \end{cases}$$

A function that minimizes this probability of error is known as a Bayes classifier f^*; in the special case of equally probable classes- - that is, when $\mathbb{P}[Y=+1] = \mathbb{P}[Y=-1] = \frac{1}{2}$ – a Bayes classifier is given by

$$f^*(x) = \begin{cases} +1 & \text{if } \psi(x) \geq 1 \\ -1 & \text{otherwise} \end{cases}$$

Since the likelihood ratio ψ (and hence f^*) is unknown, a natural approach to approximating the Bayes rule is based on choosing $\widehat{f}$ to minimize the empirical risk

$$\widehat{R}_n(f,f^*) := \frac{1}{n}\sum_{i=1}^n \underbrace{\mathbb{I}[f(X_i) \neq Y_i]}_{\mathcal{L}_f(X_i,Y_i)}$$

corresponding to the fraction of training samples that are misclassified. Typically, the minimization over f is restricted to some subset of all possible decision rules. See Chapter 14 for some further discussion of how to analyze such methods for binary classification.

Returning to the main thread, our goal is to develop methods for controlling the excess risk. For simplicity, let us assume [1] that there exists some $\theta_0 \in \Omega_0$ such that $R\left(\theta_0, \theta^*\right) = \inf_{\theta \in \Omega_0} R\left(\theta, \theta^*\right)$. With this notation, the excess risk can be decomposed as

$$E\left(\widehat{\theta}, \theta^*\right) = \underbrace{\left\{R\left(\widehat{\theta}, \theta^*\right) - \widehat{R}_n\left(\widehat{\theta}, \theta^*\right)\right\}}_{T_1} + \underbrace{\left\{\widehat{R}_n\left(\widehat{\theta}, \theta^*\right) - \widehat{R}_n\left(\theta_0, \theta^*\right)\right\}}_{T_2 \leq 0} + \underbrace{\left\{\widehat{R}_n\left(\theta_0, \theta^*\right) - R\left(\theta_0, \theta^*\right)\right\}}_{T_3}$$

Note that T_2 is non-positive, since $\widehat{\theta}$ minimizes the empirical risk over Ω_0. The third term T_3 can be dealt with in a relatively straightforward manner, because θ_0 is an unknown but non-random quantity. Indeed, recalling the definition of the empirical risk, we have

$$T_3 = \left[\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_{\theta_0}\left(X_i\right)\right] - \mathbb{E}_X\left[\mathcal{L}_{\theta_0}(X)\right]$$

corresponding to the deviation of a sample mean from its expectation for the random variable $\mathcal{L}_{\theta_0}(X)$. This quantity can be controlled using the techniques introduced in Chapter 2$-$ for instance, via the Hoeffding bound when the samples are independent and the cost function is bounded.

Finally, returning to the first term, it can be written in a similar way, namely as the difference

$$T_1 = \mathbb{E}_X\left[\mathcal{L}_{\hat{\theta}}(X)\right] - \left[\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_{\hat{\theta}}\left(X_i\right)\right]$$

This quantity is more challenging to control, because the parameter $\widehat{\theta}$ -in contrast to the deterministic quantity θ_0- is now random, and moreover depends on the samples $\{X_i\}_{i=1}^n$, since it was obtained by minimizing the empirical risk. For this reason, controlling the first term requires a stronger result, such as a uniform law of large numbers over the cost function class $\mathfrak{L}\left(\Omega_0\right) := \left\{x \mapsto \mathcal{L}_\theta(x), \theta \in \Omega_0\right\}$. With this notation, we have

$$T_1 \leq \sup_{\theta \in \Omega_0}\left|\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_\theta\left(X_i\right) - \mathbb{E}_X\left[\mathcal{L}_\theta(X)\right]\right| = \|\mathbb{P}_n - \mathbb{P}\|_{\mathfrak{L}(\Omega_0)}$$

Since T_3 is also dominated by this same quantity, we conclude that the excess risk is at most $2\|\mathbb{P}_n - \mathbb{P}\|_{\mathfrak{L}(\Omega_0)}$. This derivation demonstrates that the central challenge in analyzing estimators based on empirical risk minimization is to establish a uniform law of large numbers for the loss class $\mathfrak{L}\left(\Omega_0\right)$. We explore various concrete examples of this procedure in the exercises.

3.1 A uniform law via Rademacher complexity

Having developed various motivations for studying uniform laws, let us now turn to the technical details of deriving such results. An important quantity that underlies the study of uniform laws is the Rademacher complexity of the function class $\mathscr{F}$.

Definition 3.1.1 — Rademacher complexity. For any fixed collection $x_1^n := (x_1, \ldots, x_n)$ of points, consider the subset of $\mathbb{R}^n$ given by

$$\mathscr{F}\left(x_1^n\right) := \left\{\left(f\left(x_1\right), \ldots, f\left(x_n\right)\right) \mid f \in \mathscr{F}\right\} \tag{3.9}$$

The set $\mathscr{F}\left(x_1^n\right)$ corresponds to all those vectors in $\mathbb{R}^n$ that can be realized by applying a function $f \in \mathscr{F}$ to the collection $(x_1, \ldots, x_n)$, and the empirical Rademacher complexity is given by

$$\mathcal{R}\left(\mathscr{F}\left(x_1^n\right)/n\right) := \mathbb{E}_{\varepsilon}\left[\sup_{f \in \mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f\left(x_i\right)\right|\right]$$

Note that this definition coincides with our earlier definition of the Rademacher complexity of a set (see Example 2.25).

Given a collection $X_1^n := \{X_i\}_{i=1}^n$ of random samples, then the empirical Rademacher complexity $\mathcal{R}\left(\mathscr{F}\left(X_1^n\right)/n\right)$ is a random variable. Taking its expectation yields the Rademacher complexity of the function class $\mathscr{F}$ – namely, the deterministic quantity

$$\mathcal{R}_n(\mathscr{F}) := \mathbb{E}_X\left[\mathcal{R}\left(\mathscr{F}\left(X_1^n\right)/n\right)\right] = \mathbb{E}_{X,\varepsilon}\left[\sup_{f \in \mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f\left(X_i\right)\right|\right]$$

Note that the Rademacher complexity is the average of the maximum correlation between the vector $\left(f\left(X_1\right), \ldots, f\left(X_n\right)\right)$ and the "noise vector" $(\varepsilon_1, \ldots, \varepsilon_n)$, where the maximum is taken over all functions $f \in \mathscr{F}$.

The intuition is a natural one: a function class is extremely large- -and, in fact, "too large" for statistical purposes-if we can always find a function that has a high correlation with a randomly drawn noise vector. Conversely, when the Rademacher complexity decays as a function of sample size, then it is impossible to find a function that correlates very highly in expectation with a randomly drawn noise vector. We now make precise the connection between Rademacher complexity and the GlivenkoCantelli property, in particular by showing that, for any bounded function class $\mathscr{F}$, the condition $\mathcal{R}_n(\mathscr{F}) = o(1)$ implies the Glivenko-Cantelli property. More precisely, we prove a non-asymptotic statement, in terms of a tail bound for the probability that the random variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ deviates substantially above a multiple of the Rademacher complexity. It applies to a function class $\mathscr{F}$ that is b -uniformly bounded, meaning that $\|f\|_{\infty} \leq b$ for all $f \in \mathscr{F}$

Theorem 3.1.1 For any b -uniformly bounded class of functions $\mathscr{F}$, any positive integer $n \geq 1$ and any scalar $\delta \geq 0$, we have

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \leq 2\mathcal{R}_n(\mathscr{F}) + \delta$$

with $\mathbb{P}$ -probability at least $1 - \exp\left(-\frac{n\delta^2}{2b^2}\right)$. Consequently, as long as $\mathcal{R}_n(\mathscr{F}) = o(1)$, we have$\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} \xrightarrow{\text{a.s.}} 0$

In order for Theorem 3.1.1 to be useful, we need to obtain upper bounds on the Rademacher complexity. There are a variety of methods for doing so, ranging from direct calculations to alternative complexity measures. In Section 4.3, we develop some techniques for upper bounding the Rademacher complexity for indicator functions of half-intervals, as required for the classical Glivenko-Cantelli theorem (see Example 3.4); we also discuss the notion of Vapnik-Chervonenkis dimension, which can be used to upper bound the Rademacher complexity for other function classes. In Chapter 5, we introduce more advanced techniques based on metric entropy and chaining for controlling Rademacher complexity and related sub-Gaussian processes. In the meantime, let us turn to the proof of Theorem 3.1.1

Proof. We first note that if $\mathcal{R}_n(\mathscr{F}) = o(1)$, then the almost-sure convergence follows from the tail bound (**??**) and the Borel-Cantelli lemma. Accordingly, the remainder of the argument is devoted to proving the tail bound (**??**).

Concentration around mean: We first claim that, when $\mathscr{F}$ is uniformly bounded, then the random variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ is sharply concentrated around its mean. In order to simplify notation, it is convenient to define the recentered functions $\bar{f}(x) := f(x) - \mathbb{E}[f(X)]$, and to write $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} = \sup_{f\in\mathscr{F}} \left|\frac{1}{n}\sum_{i=1}^n \bar{f}(X_i)\right|$. Thinking of the samples as fixed for the moment, consider the function

$$G(x_1, \ldots, x_n) := \sup_{f\in\mathscr{F}} \left|\frac{1}{n}\sum_{i=1}^n \bar{f}(x_i)\right|$$

We claim that G satisfies the Lipschitz property required to apply the bounded differences method (recall Corollary 1.0.11). since the function G is invariant to permutation of its coordinates, it suffices to bound the difference when the first coordinate x_1 is perturbed. Accordingly, we define the vector $y \in \mathbb{R}^n$ with $y_i = x_i$ for all $i \neq 1$, and seek to bound the difference $|G(x) - G(y)|$. For any function $\bar{f} = f - \mathbb{E}[f]$, we have

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n \bar{f}(x_i)\right| - \sup_{h\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \bar{h}(y_i)\right| &\leq \left|\frac{1}{n}\sum_{i=1}^n \bar{f}(x_i)\right| - \left|\frac{1}{n}\sum_{i=1}^n \bar{f}(y_i)\right| \\ &\leq \frac{1}{n}\left|\bar{f}(x_1) - \bar{f}(y_1)\right| \\ &\leq \frac{2b}{n}\end{aligned} \tag{3.10}$$

where the final inequality uses the fact that

$$\left|\bar{f}(x_1)-\bar{f}(y_1)\right| = |f(x_1)-f(y_1)| \leq 2b$$

which follows from the uniform boundedness condition $\|f\|_\infty \leq b$. since the inequality (3.10) holds for any function f, we may take the supremum over $f \in \mathscr{F}$ on both sides; doing so yields the inequality $G(x)-G(y) \leq \frac{2b}{n}$. Since the same argument may be applied with the roles of x and y reversed, we conclude that $|G(x)-G(y)| \leq \frac{2b}{n}$. Therefore, by the bounded differences method (see Corollary 1.0.11), we have

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}} - \mathbb{E}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}\right] \leq t \quad \text{with } \mathbb{P}\text{-prob. at least } 1-\exp\left(-\frac{nt^2}{2b^2}\right) \tag{3.11}$$

valid for all $t \geq 0$.

Upper bound on mean: It remains to show that $\mathbb{E}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}\right]$ is upper bounded by $2\mathcal{R}_n(\mathscr{F})$ and we do so using a classical symmetrization argument. Letting $(Y_1,\ldots,Y_n)$ be a second i.i.d. sequence, independent of $(X_1,\ldots,X_n)$, we have

$$\begin{aligned}
\mathbb{E}\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}\right] &= \mathbb{E}_X\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \{f(X_i) - \mathbb{E}_{Y_i}[f(Y_i)]\}\right|\right] \\
&= \mathbb{E}_X\left[\sup_{f\in\mathscr{F}}\left|\mathbb{E}_Y\left[\frac{1}{n}\sum_{i=1}^n \{f(X_i)-f(Y_i)\}\right]\right|\right] \\
&\leq^{(i)} \mathbb{E}_{X,Y}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \{f(X_i)-f(Y_i)\}\right|\right]
\end{aligned} \tag{3.12}$$

where the upper bound (i) follows from the calculation of Exercise 4.4. Now let $(\varepsilon_1,\ldots,\varepsilon_n)$ be an i.i.d. sequence of Rademacher variables, independent of X and Y. Given our independence assumptions, for any function $f \in \mathscr{F}$, the random vector with components $\varepsilon_i(f(X_i)-f(Y_i))$ has the same joint distribution as the random vector with components $f(X_i)-f(Y_i)$, whence

$$\begin{aligned}
\mathbb{E}_{X,Y}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \{f(X_i)-f(Y_i)\}\right|\right] &= \mathbb{E}_{X,Y,\varepsilon}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(f(X_i)-f(Y_i))\right|\right] \\
&\leq 2\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] = 2\mathcal{R}_n(\mathscr{F})
\end{aligned}$$

Combining the upper bound (**??**) with the tail bound (3.11) yields the claim. ■

3.1.1 Necessary conditions with Rademacher complexity

The proof of Theorem 3.1.1illustrates an important technique known as symmetrization, which relates the random variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ to its symmetrized version

$$\|\mathbb{S}_n\|_{\mathscr{F}} := \sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|$$

Note that the expectation of $\|\mathbb{S}_n\|_{\mathscr{F}}$ corresponds to the Rademacher complexity, which plays a central role in Theorem 3.1.1. It is natural to wonder whether much was lost in moving from the variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}$ to its symmetrized version. The following "sandwich" result relates these quantities.

Proposition 3.1.2 For any convex non-decreasing function $\Phi : \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{E}_{X,\varepsilon}\left[\Phi\left(\frac{1}{2}\|\mathbb{S}_n\|_{\mathscr{F}}\right)\right] \leq^{(a)} \mathbb{E}_X\left[\Phi\left(\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}\right)\right] \leq^{(b)} \mathbb{E}_{X,\varepsilon}\left[\Phi\left(2\|\mathbb{S}_n\|_{\mathscr{F}}\right)\right] \tag{3.13}$$

where $\overline{\mathscr{F}} = \{f - \mathbb{E}[f], f \in \mathscr{F}\}$ is the recentered function class.

When applied with the convex non-decreasing function $\Phi(t) = t$, Proposition 3.1.2 yields the inequalities

$$\frac{1}{2}\mathbb{E}_{X,\varepsilon}\|\mathbb{S}_n\|_{\mathscr{F}} \leq \mathbb{E}_X\left[\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}\right] \leq 2\mathbb{E}_{X,\varepsilon}\|\mathbb{S}_n\|_{\mathscr{F}} \tag{3.14}$$

with the only differences being the constant pre-factors, and the use of $\mathscr{F}$ in the upper bound, and the recentered class $\overline{\mathscr{F}}$ in the lower bound.

Other choices of interest include $\Phi(t) = e^{\lambda t}$ for some $\lambda > 0$, which can be used to control the moment generating function.

Proof. Beginning with bound (b), we have

$$\begin{aligned}
\mathbb{E}_X\left[\Phi\left(\|\mathbb{P}_n - \mathbb{P}\|_{\mathscr{F}}\right)\right] &= \mathbb{E}_X\left[\Phi\left(\sup_{f\in\mathscr{F}} | \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}_Y[f(Y_i)]\right)\right] \\
&\leq^{(i)} \mathbb{E}_{X,Y}\left[\Phi\left(\sup_{f\in\mathscr{F}} | \frac{1}{n}\sum_{i=1}^{n} f(X_i) - f(Y_i)\right)\right] \\
&\underbrace{\overset{\text{(ii)}}{=} \mathbb{E}_{X,Y,\varepsilon}\left[\Phi\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\{f(X_i) - f(Y_i)\}\right|\right)\right]}_{:=T_1}
\end{aligned}$$

where inequality (i) follows from Exercise 4.4, using the convexity and non-decreasing properties of Φ, and equality (ii) follows since the random vector with components $\varepsilon_i\left(f(X_i)\right.$ $\left.f(Y_i)\right)$ has the same joint distribution as the random vector with components $f(X_i) - f(Y_i)$ By the triangle inequality, we have

$$\begin{aligned}
T_1 \leq& \mathbb{E}_{X,Y,\varepsilon}\left[\Phi\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right| + \left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(Y_i)\right|\right)\right] \\
\leq^{(iii)}& \frac{1}{2}\mathbb{E}_{X,\varepsilon}\left[\Phi\left(2\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right)\right] + \frac{1}{2}\mathbb{E}_{Y,\varepsilon}\left[\Phi\left(2\sup_{f\in\mathscr{F}} | \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(Y_i)\right)\right] \\
\overset{\text{(iv)}}{=}& \mathbb{E}_{X,\varepsilon}\left[\Phi\left(2\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right)\right]
\end{aligned}$$

where step (iii) follows from Jensen's inequality and the convexity of Φ, and step (iv) follows since X and Y are i.i.d. samples. Turning to the bound (a), we have

$$\begin{aligned}\mathbb{E}_{X,\varepsilon}\left[\Phi\left(\frac{1}{2}\|\mathbb{S}_n\|_{\mathscr{F}}\right)\right] &= \mathbb{E}_{X,\varepsilon}\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left\{f(X_i)-\mathbb{E}_{Y_i}[f(Y_i)]\right\}\right|\right)\right]\\ &\leq \mathbb{E}_{X,Y,\varepsilon}\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left\{f(X_i)-f(Y_i)\right\}\right|\right)\right]\\ &\overset{\text{(ii)}}{=} \mathbb{E}_{X,Y}\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{f(X_i)-f(Y_i)\right\}\right|\right)\right]\end{aligned}$$

where inequality (i) follows from Jensen's inequality and the convexity of Φ; and equality (ii) follows since for each $i=1,2,\ldots,n$ and $f\in\mathscr{F}$, the variables $\varepsilon_i\{f(X_i)-f(Y_i)\}$ and $f(X_i)-f(Y_i)$ have the same distribution. Now focusing on the quantity $T_2 := \frac{1}{2}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\{f(X_i)-f(Y_i)\}\right|$, we add and subtract a term of the form $\mathbb{E}[f]$, and then apply the triangle inequality, thereby obtaining the upper bound

$$T_2 \leq \frac{1}{2}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\{f(X_i)-\mathbb{E}[f]\}\right| + \frac{1}{2}\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\{f(Y_i)-\mathbb{E}[f]\}\right|$$

since Φ is convex and non-decreasing, we are guaranteed that

$$\Phi(T_2) \leq \frac{1}{2}\Phi\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\{f(X_i)-\mathbb{E}[f]\}\right|\right) + \frac{1}{2}\Phi\left(\sup_{f\in\mathscr{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\{f(Y_i)-\mathbb{E}[f]\}\right|\right)$$

The claim follows by taking expectations and using the fact that X and Y are identically distributed. ■

A consequence of Proposition 3.1.2 is that the random variable $\|\mathbb{P}_n-\mathbb{P}\|_{\mathscr{F}}$ can be lower bounded by a multiple of Rademacher complexity, and some fluctuation terms. This fact can be used to prove the following:

Proposition 3.1.3 For any b-uniformly bounded function class $\mathscr{F}$, any integer $n\geq 1$ and any scalar $\delta\geq 0$, we have

$$\|\mathbb{P}_n-\mathbb{P}\|\mathscr{F} \geq \frac{1}{2}\mathcal{R}_n(\mathscr{F}) - \frac{\sup_{f\in\mathscr{F}}|\mathbb{E}[f]|}{2\sqrt{n}} - \delta$$

with $\mathbb{P}$-probability at least $1-e^{-\frac{n n^2}{2b^2}}$

We leave the proof of this result for the reader (see Exercise 4.5). As a consequence, if the Rademacher complexity $\mathcal{R}_n(\mathscr{F})$ remains bounded away from zero, then $\|\mathbb{P}_n-\mathbb{P}\|_{\mathscr{F}}$ cannot converge to zero in probability. We have thus shown that, for a uniformly bounded function class $\mathscr{F}$, the Rademacher complexity provides a necessary and sufficient condition for it to be Glivenko-Cantelli.

3.2 Upper bounds on the Rademacher complexity

Obtaining concrete results using Theorem 3.1.1 requires methods for upper bounding the Rademacher complexity. There are a variety of such methods, ranging from simple union bound methods (suitable for finite function classes) to more advanced techniques involving the notion of metric entropy and chaining arguments. We explore the latter techniques in Chapter 5 to follow. This section is devoted to more elementary techniques, including those required to prove the classical Glivenko-Cantelli result, and, more generally, those that apply to function classes with polynomial discrimination, as well as associated VapnikChervonenkis classes.

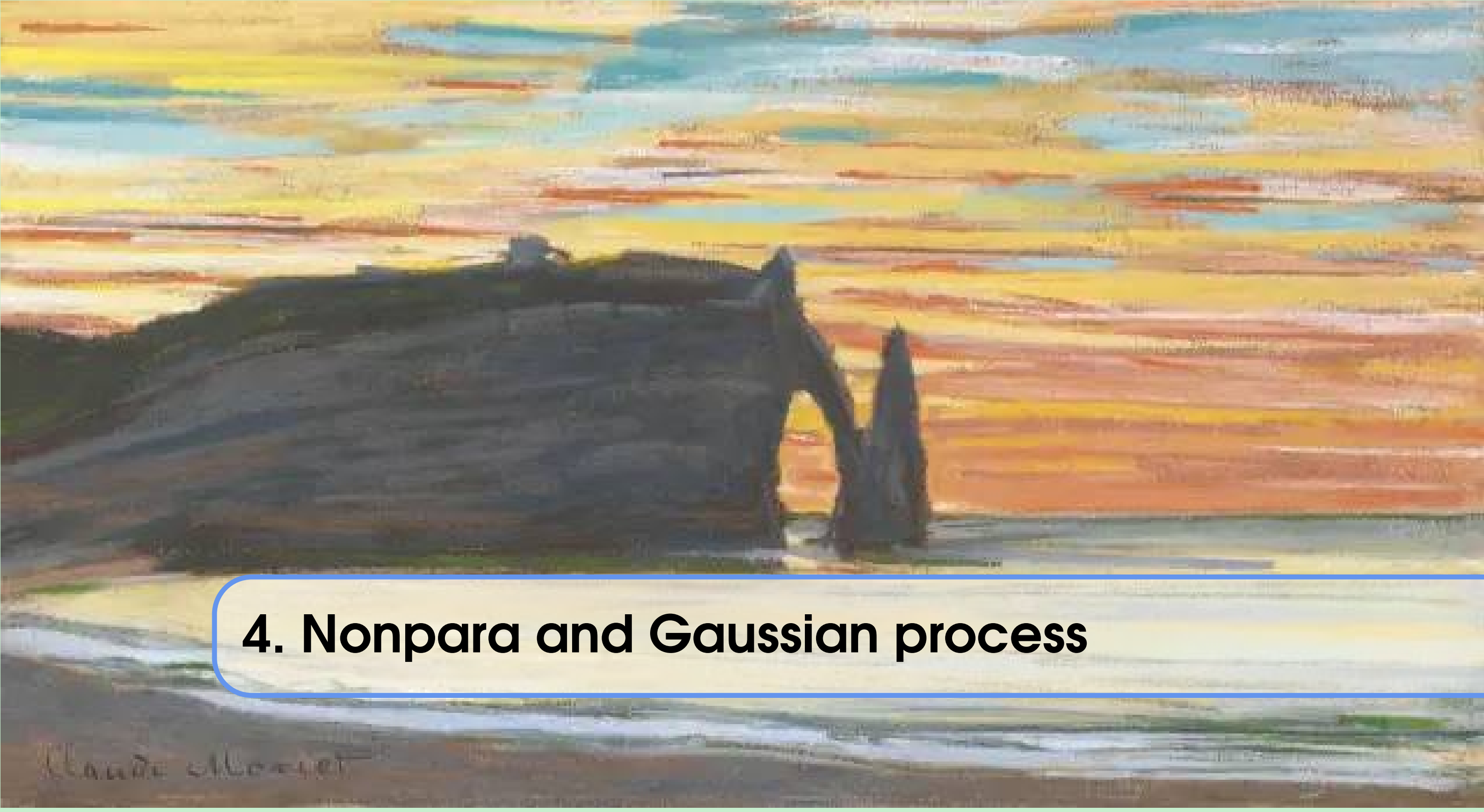
3.2.1 Classes with polynomial discrimination

For a given collection of points $x_1^n = (x_1, \ldots, x_n)$, the "size" of the set $\mathscr{F}(x_1^n)$ provides a sample-dependent measure of the complexity of $\mathscr{F}$. In the simplest case, the set $\mathscr{F}(x_1^n)$ contains only a finite number of vectors for all sample sizes, so that its "size" can be measured via its cardinality. For instance, if $\mathscr{F}$ consists of a family of decision rules taking binary values (as in Example 3.7), then $\mathscr{F}(x_1^n)$ can contain at most 2^n elements. Of interest to us are function classes for which this cardinality grows only as a polynomial function of n, as formalized in the following:

Definition 3.2.1 — Polynomial discrimination. A class $\mathscr{F}$ of functions with domain $\mathcal{X}$ has polynomial discrimination of order $v \geq 1$ if, for each positive integer n and collection $x_1^n = \{x_1, \ldots, x_n\}$ of n points in $\mathcal{X}$, the set $\mathscr{F}(x_1^n)$ has cardinality upper bounded as

$$\operatorname{card}(\mathscr{F}(x_1^n)) \leq (n+1)^v$$

The significance of this property is that it provides a straightforward approach to controlling the Rademacher complexity. For any set $\mathbb{S} \subset \mathbb{R}^n$, we use $D(\mathbb{S}) := \sup_{x \in \mathbb{S}} \|x\|_2$ to denote its maximal width in the ℓ_2 -norm.



4. Nonpara and Gaussian process

Very generally speaking, a statistical model for a random observation Y is a family

$$\{P_f : f \in \mathcal{F}\}$$

of probability distributions P_f, each of which is a candidate for having generated the observation Y. The parameter f belongs to the parameter space $\mathcal{F}$. The problem of statistical inference on f, broadly speaking, can be divided into three intimately connected problems of using the observation Y to

1. Estimate the parameter f by an estimator $T(Y)$,
2. Test hypotheses on f based on test functions $\Psi(Y)$ and/or
3. Construct confidence sets $C(Y)$ that contain f with high probability.

To interpret inferential results of these kinds, we will typically need to specify a distance, or loss function on $\mathcal{F}$, and for a given model, different loss functions may or may not lead to very different conclusions.

On the other hand, these models are naturally divided by the different probabilistic frameworks in which they occur - which will be either a Gaussian noise model or an independent sampling model. These frameworks are asymptotically related in a fundamental way (see the discussion after Theorem 4.0.1). However, the most effective probabilistic techniques available are based on a direct, nonasymptotic analysis of the Gaussian or product probability measures that arise in the relevant sampling context and hence require a separate treatment.

Thus, while many of the statistical intuitions are common to both the sampling and the Gaussian noise models and in fact inform each other, the probabilistic foundations of these models will be laid out independently.

The perhaps most basic problem of statistics is the following: consider repeated outcomes of the experiment X, that is, a random sample of independent and identically distributed (i.i.d.) copies $X_1, \ldots, X_n$ from X. The joint distribution of the X_i equals the product probability measure $P^n = \otimes_{i=1}^n P$ on $(\mathcal{X}^n, \mathcal{A}^n)$. The goal is to recover P from the n observations. Recovering P then simply means to use the observations to make inferences on the unknown parameter, and the fact that this parameter is finite dimensional is crucial for this traditional paradigm of statistical inference, in particular, for the famous likelihood principle of R. A. Fisher. In this book, we will follow the often more realistic assumption that no such parametric assumptions are made on P. For most sample spaces $\mathcal{X}$ of interest, this will naturally lead to models that are infinite dimensional, and we will investigate how the theory of statistical inference needs to be developed in this situation.

Nonparametric Models for Probability Measures

In its most elementary form, without imposing any parameterisations on P, we can simply consider the problem of making inferences on the unknown probability measure P based on the sample. Natural loss functions arise from the usual metrics on the space of probability measures on $\mathcal{X}$, such as the **total variation metric**

$$\begin{aligned}\|P-Q\|_{TV} &= \sup_{A\in\mathcal{A}} |P(A)-Q(A)| \\ &\left(=^{(i)} \sup_{A\in\mathcal{A}, f=1_A} \left| \int_{\mathcal{X}} f(\mathrm{d}P-\mathrm{d}Q) \right| =^{(ii)} \sup_{f\in\mathcal{A}, 0\le f\le 1} \left| \int_{\mathcal{X}} f(\mathrm{d}P-\mathrm{d}Q) \right| \right)\end{aligned}$$

(i) is because f is the indictor function of A. Consider that the sup is obtained at $A^* = \{x : p(x) \geq q(x)\}$, where $p(x), q(x)$ are the density of P, Q , so (ii) can be obtained by

$$\begin{aligned} P(A^*) - Q(A^*) &= \sup_{A\in\mathcal{A}, f=1_A} \left| \int_{\mathcal{X}} f(\mathrm{d}P-\mathrm{d}Q) \right| \\ &\le \sup_{f\in\mathcal{A}, 0\le f\le 1} \left| \int_{\mathcal{X}} f(\mathrm{d}P-\mathrm{d}Q) \right| \\ &= \sup_{f\in\mathcal{A}, 0\le f\le 1} \left| \int_{x\in A^*} f(\mathrm{d}P-\mathrm{d}Q) + \int_{x\notin A^*} f(\mathrm{d}P-\mathrm{d}Q) \right| \\ &\le^{(iii)} \int_{x\in A^*} (\mathrm{d}P-\mathrm{d}Q) \quad (f^* = 1_{A^*}) \\ &= P(A^*) - Q(A^*) \end{aligned}$$

(iii) can also be gotten by $\cdots \le \int_{x\notin A^*} (\mathrm{d}Q-\mathrm{d}P) = Q(A^{*c}) - P(A^{*c}) = P(A^*) - Q(A^*)$, and at this time the optimal is $f^* = 1_{A^{*c}}$ (The domain A^* make the first term to be positive and the second term to be nonpositive). We have $p(x), q(x)$ of P and Q because this assumption entails no loss of generality, since $\mathbb{P}$ and $\mathbb{Q}$ both have densities with respect to $v = \frac{1}{2}(\mathbb{P}+\mathbb{Q})$. (If there exists measure μ satisfies $P \ll \mu, Q \ll \mu$, then $p(x)$ and

$q(x)$ are P and Q to μ Radon-Nikodym deri. For example $\mathcal{X}=\mathbb{R}^d$, μ can be taken as Lebesgue measure. And for general measure space, we have $\mu=\frac{1}{2}(P+Q)$.) Actually, Total Variation Metric functions $\mathcal{D}$ can be written as $\mathcal{D}=\{f: f\in\mathcal{A}, |f|\leq 1\}$, which also is equal to $\|P-Q\|_{TV}=\frac{1}{2}\sup_{f\in\mathcal{A},|f|\leq 1}\left|\int_{\mathcal{X}} f(\mathrm{d}P-\mathrm{d}Q)\right| \quad (f^*=1_{A^*}-1_{A^*c})$. $\|P-Q\|_{TV}\triangleq 2\cdot\sup_{A\in\mathcal{A}}|P(A)-Q(A)|$, and the corrspondence integral form is $\sup_{f\in\mathcal{A},|f|\leq 1}\left|\int_{\mathcal{X}} f(\mathrm{d}P-\mathrm{d}Q)\right|$.

Or weaker metrics that generate the topology of weak convergence of probability measures on $\mathcal{X}$. For instance, if $\mathcal{X}$ itself is endowed with a metric d, we could take the bounded Lipschitz metric. Link: Wasserstein GAN and the Kantorovich-Rubinstein Duality

$$\beta_{(\mathcal{X},d)}(P,Q)=\sup_{f\in BL(1)}\left|\int_{\mathcal{X}} f(dP-dQ)\right|$$

General Wasserstein distence

$$W_p(P,Q)\triangleq\left(\inf_{\gamma\in\Gamma(P,Q)}\int_{\mathcal{X}\times\mathcal{X}} d(x,y)^p\mathrm{d}\gamma(x,y)\right)^{1/p}$$

can be as the metric of two measures,but when $p=1$,according to Kantorovich-Rubinstein Duality Theorem, W_1 has a dual represention

$$W_1(P,Q)=\sup_{f\in\mathrm{Lip}(1)}\left|\int_{\mathcal{X}} f(\mathrm{d}P-\mathrm{d}Q)\right|$$

where Lip (M) is the set of Lipschitz constant of M 's Lipschitz function.

For weak convergence of probability measures, where

$$BL(M)=\left\{f:\mathcal{X}\to\mathbb{R}, \sup_{x\in\mathcal{X}}|f(x)|+\sup_{x\neq y}\frac{|f(x)-f(y)|}{d(x,y)}\leq M\right\},\quad 0<M<\infty$$

If $\mathcal{X}$ has some geometric structure, we can consider more intuitive loss functions. For example, if $\mathcal{X}=\mathbb{R}$, we can consider the cumulative distribution function

$$F(x)=P(X\leq x),\quad x\in\mathbb{R}$$

or, if X takes values in $\mathbb{R}^d$, its multivariate analogue. A natural distance function on distribution functions is simply the **supremum-norm metric ('Kolmogorov distance')**

$$\|F_P-F_Q\|_\infty=\sup_{x\in\mathbb{R}}|F_P(x)-F_Q(x)|$$

Since the indicators $\left\{1_{(-\infty,x]}: x\in\mathbb{R}\right\}$ generate the Borel σ -field of $\mathbb{R}$, we see that, on $\mathbb{R}$, the statistical parameter P is characterised entirely by the functional parameter F, and vice versa. The parameter space is thus the infinite-dimensional space of all cumulative distribution functions on $\mathbb{R}$.

Often we will know that P has some more structure, such as that P possesses a probability-density function $f:\mathbb{R}\to[0,\infty)$, which itself may have further properties that will be seen to influence the complexity of the statistical problem at hand. For probability-density functions, a natural loss function is the L^1 -distance

$$\|f_P - f_Q\|_1 = \int_{\mathbb{R}} |f_P(x) - f_Q(x)|\, dx$$

and in some situations also other L^p -type and related loss functions. Although in some sense a subset of the other, the class of probability densities is more complex than the class of probability-distribution functions, as it is not described by monotonicity constraints and does not consist of functions bounded in absolute value by 1. In a heuristic way, we can anticipate that estimating a probability density is harder than estimating the distribution function, just as the preceding total variation metric is stronger than any metric for weak convergence of probability measures (on nontrivial sample spaces $\mathcal{X}$). In all these situations, we will see that the theory of statistical inference on the parameter f significantly departs from the usual finite-dimensional setting.

Instead of P, a particular functional $\Phi(P)$ may be the parameter of statistical interest, such as the moments of P or the quantile function F^{-1} of the distribution function $F-$ examples for this situation are abundant. The nonparametric theory is naturally compatible with such functional estimation problems because it provides the direct plug-in estimate $\Phi(T)$ based on an estimator T for P. Proving closeness of T to P in some strong loss function then gives access to 'many' continuous functionals Φ for which $\Phi(T)$ will be close to $\Phi(P)$, as we shall see later in this book.

Indirect Observations

A common problem in statistical sampling models is that some systematic measurement errors are present. A classical problem of this kind is the statistical regression problem, which will be introduced in the next section. Another problem, which is more closely related to the sampling model from earlier, is where one considers observations in $\mathbb{R}^d$ of the form

$$Y_i = X_i + \varepsilon_i, \quad i = 1, \ldots, n \tag{4.1}$$

where the X_i are i.i.d. with common law P_X, and the ε_i are random 'error' variables that are independent of the X_i and have law P_ε. The law P_ε is assumed to be known to the observer - the nature of this assumption is best understood by considering examples: the attempt is to model situations in which a scientist, for reasons of cost, complexity or lack of precision of the involved measurement device, is forced to observe Y_i instead of the realisations X_i of interest. The observer may, however, have very concrete knowledge of the source of the error, which could, for example, consist of light emissions of the Milky Way interfering with cosmic rays from deeper space, an erratic optical device through which images are observed (e.g., a space telescope which cannot be repaired except at very high cost) or transmissions of signals through a very busy communication channel. Such situations of implicit measurements are encountered frequently in the applied sciences and are often called inverse problems, as one wishes to 'undo' the errors inflicted on the signal in which one is interested. The model (4.1) gives a simple way to model the main aspects of such statistical inverse problems. It is also known as the deconvolution model because the law of the Y_i equals

$$P_Y = P_X * P_\varepsilon$$

the convolution of the two probability measures P_X, P_ε, and one wishes to 'deconvolve P_ε. As earlier, we will be interested in inference on the underlying distribution P_X of the signal X when the statistical model for P_X is infinite dimensional. The loss functions in this problem are thus typically the same as in the preceding subsection.

Gaussian Models

The randomness in the preceding sampling model was encoded in a general product measure P^n describing the joint law of the observations. Another paradigm of statistical modelling deals with situations in which the randomness in the model is described by a Gaussian (normal) distribution. This paradigm naturally encompasses a variety of nonparametric models, where the infinite-dimensional character of the problem does not necessarily derive from the probabilistic angle but from a functional relationship that one wishes to model.

Basic Ideas of Regression

Perhaps the most natural occurrence of a statistical model in the sciences is the one in which observations, modelled here as numerical values or vectors, say, (Y_i, x_i), arise according to a functional relationship

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n \tag{4.2}$$

where n is the number of observations (sample size), f is some function of the x_i and the ε_i are random noise. By 'random noise', we may mean here either a probabilistic model for certain measurement errors that we believe to be intrinsic to our method of making observations, or some innate stochastic nature of the way the Y_i are generated from the $f(x_i)$. In either case, we will model the ε_i as random variables in the sense of axiomatic probability theory . It is sometimes natural to assume also that the x_i are realisations of random variables X_i- we can either take this into account explicitly in our analysis or make statements conditional on the observed values $X_i = x_i$

The function f often will be unknown to the observer of observations (Y_i, x_i), and the goal is to recover f from the (Y_i, x_i). This may be of interest for various reasons, for instance, for predicting new values Y_{n+1} from $f(x_{n+1})$ or to gain quantitative and qualitative understanding of the functional relationship $Y_i = f(x_i)$ under consideration.

In the preceding context, a statistical model in the broad sense is an a priori specification of both a parameter space for the functions f that possibly could have generated (4.2) and a family of probability measures that describes the possible distributions of the random variables ε_i. By 'a priori', we mean here that this is done independently of (e.g., before) the observational process, reflecting the situation of an experimentalist.

A systematic use and study of such models was undertaken in the early nineteenth century by Carl Friedrich Gauss, who was mostly interested in predicting astronomical observations. When the model is translated into the preceding formalisation, Gauss effectively assumed that the x_i are vectors $\left(x_{i1},\ldots,x_{ip}\right)^T$ and thought of f as a linear function in that vector, more precisely,

$$f\left(x_i\right) = x_{i1}\theta_i + \ldots x_{ip}\theta_p, \quad i = 1,\ldots,n$$

for some real-valued parameters $\theta_j, j = 1,\ldots,p$. The parameter space for f is thus the Euclidean space $\mathbb{R}^p$ expressed through all such linear mappings. In Gauss's time, the assumption of linearity was almost a computational necessity.

Moreover, Gauss modelled the random noise ε_i as independent and identically distributed samples from a normal distribution $N\left(0,\sigma^2\right)$ with some variance σ^2. His motivation behind this assumption was twofold. First, it is reasonable to assume that $E\left(\varepsilon_i\right) = 0$ for every i. If this expectation were nonzero, then there would be some deterministic, or 'systematic', measurement error $e_i = E\left(\varepsilon_i\right)$ of the measurement device, and this could always be accommodated in the functional model by adding a constant $x_{10} = \cdots = x_{n0} = 1$ to the preceding linear relationship. The second assumption that ε_i has a normal distribution is deeper. If we think of each measurement error ε_i as the sum of many 'very small', or infinitesimal, independent measurement errors $\varepsilon_{ik}, k = 1,2,\ldots$, then, by the central limit theorem, $\varepsilon_i = \sum_k \varepsilon_{ik}$ should be approximately normally distributed, regardless of the actual distribution of the ε_{ik}. By the same reasoning, it is typically natural to assume that the ε_i are also independent among themselves. This leads to what is now called the standard Gaussian linear model

$$Y_i = f\left(x_i\right) + \varepsilon_i \equiv \sum_{j=1}^{p} x_{ij}\theta_j + \varepsilon_i, \quad \varepsilon_i \sim^{i.i.d.} N\left(0,\sigma^2\right), \quad i = 1,\ldots,n \tag{4.3}$$

which bears this name both because Gauss studied it and, since the $N\left(0,\sigma^2\right)$ distribution is often called the Gaussian distribution, because Gauss first made systematic use of it. The unknown parameter $\left(\theta,\sigma^2\right)$ varies in the $(p+1)$ -dimensional parameter space

$$\Theta \times \Sigma = \mathbb{R}^p \times (0,\infty)$$

This model constitutes perhaps the classical example of a finite-dimensional model, which has been studied extensively and for which a fairly complete theory is available. For instance, when p is smaller than n, the least-squares estimator of Gauss finds the value

$\hat{\theta} \in \mathbb{R}^p$ which solves the optimisation problem

$$\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} x_{ij}\theta_j \right)^2$$

and hence minimises the Euclidean distance of the vector $Y = (Y_1, \ldots, Y_n)^T$ to the p -dimensional subspace spanned by the p vectors $(x_{1j}, \ldots, x_{nj})^T, j = 1, \ldots, p$

Some Nonparametric Gaussian Models

We now give a variety of models that generalise Gauss's ideas to infinite-dimensional situations. In particular, we will introduce the Gaussian white noise model, which serves as a generic surrogate for a large class of nonparametric models, including even non-Gaussian ones, through the theory of equivalence of experiments (discussed in the next section).

Nonparametric Gaussian Regression

Gauss's model and its theory basically consist of two crucial assumptions: one is that the ε_i are normally distributed, and the other is that the function f is linear. The former assumption was argued to be in some sense natural, at least in a measurement-error model (see also the remarks after Theorem 4.0.1 for further justification). The latter assumption is in principle quite arbitrary, particularly in times when computational power does not constrain us as much any longer as it did in Gauss's time. A nonparametric approach therefore attempts to assume as little structure of f as possible. For instance, by the nonparametric regression model with fixed, equally spaced design on $[0,1]$, we shall understand here the model

$$Y_i = f(x_i) + \varepsilon_i, \quad x_i = \frac{i}{n}, \quad \varepsilon_i \sim^{i.i.d.} N(0, \sigma^2), \quad i = 1, \ldots, n \tag{4.4}$$

where f is any function defined on $[0,1]$. We are thus sampling the unknown function f at an equally spaced grid of $[0,1]$ that, as $n \to \infty$, grows dense in the interval $[0,1]$ as $n \to \infty$. The model immediately generalises to bounded intervals $[a,b]$, to 'approximately' equally spaced designs $\{x_i : i = 1, \ldots, n\} \subset [a,b]$ and to multivariate situations, where the x_i are equally spaced points in some hypercube. We note that the assumption that the x_i are equally spaced is important for the theory that will follow $-$ this is natural as we cannot hope to make inference on f in regions that contain no or too few observations x_i.

Other generalisations include the random design regression model, in which the x_i are viewed as i.i.d. copies of a random variable X. One can then either proceed to argue conditionally on the realisations $X_i = x_i$, or one takes this randomness explicitly into account by making probability statements under the law of X and ε simultaneously. For reasonable design distributions, this will lead to results that are comparable to the fixed-design model - one way of seeing this is through the equivalence theory for statistical experiments (see after Theorem 4.0.1).

A priori it may not be reasonable to assume that f has any specific properties other than that it is a continuous or perhaps a differentiable function of its argument. Even if we would assume that f has infinitely many continuous derivatives the set of all such f would be infinite dimensional and could never be fully captured by a p -dimensional parameter space. We thus have to expect that the theory of statistical inference in this nonparametric model will be different from the one in Gauss's classical linear model.

The Gaussian White Noise Model

For the mathematical development in this book we shall work with a mathematical idealisation of the regression model (4.4) in continuous time, known as the Gaussian white noise model, and with its infinite sequence space analogue. Consider the following stochastic differential equation:

$$dY(t) \equiv dY_f^{(n)}(t) = f(t)dt + \frac{\sigma}{\sqrt{n}}dW(t), \quad t \in [0,1], \quad n \in \mathbb{N} \tag{4.5}$$

where $f \in L^2 \equiv L^2([0,1])$ is a square integrable function on $[0,1]$,$\sigma > 0$ is a dispersion parameter and dW is a standard Gaussian white noise process. When we observe a realisation of (4.5), we shall say that we observe the function or signal f in Gaussian white noise, at the noise level, or a signal-to-noise ratio $\sigma/\sqrt{n}$. We typically think of n large, serving as a proxy for sample size, and of $\sigma > 0$ a fixed known value. If σ is unknown, one can usually replace it by a consistent estimate in the models we shall encounter in this book. The exact meaning of dW needs further explanation. Heuristically, we may think of dW as a weak derivative of a standard Brownian motion $\{W(t) : t \in [0,1]\}$, whose existence requires a suitable notion of stochastic derivative that we do not want to develop here explicitly. Instead, we take a 'stochastic process' approach to define this stochastic differential equation, which for statistical purposes is perfectly satisfactory. Let us thus agree that 'observing the trajectory (4.5)' will simply mean that we observe a realisation of the Gaussian process defined by the application

$$g \mapsto \int_0^1 g(t)dY^{(n)}(t) \equiv \mathbb{Y}_f^{(n)}(g) \sim N\left(\langle f,g\rangle, \frac{\|g\|_2^2}{n}\right) \tag{4.6}$$

where g is any element of the Hilbert space $L^2([0,1])$ with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|_2$. Even more explicitly, we observe all the $N\left(\langle f,g\rangle, \|g\|_2^2/n\right)$ variables, as g runs through $L^2([0,1])$. The randomness in the equation (4.5) comes entirely from the additive term dW so after translating by $\langle f,g\rangle$ and scaling by $1/\sqrt{n}$, this means that dW is defined through the Gaussian process obtained from the action

$$g \mapsto \int_0^1 g(t)dW(t) \equiv \mathbb{W}(g) \sim N\left(0, \|g\|_2^2\right), \quad g \in L^2([0,1]) \tag{4.7}$$

Note that this process has a diagonal covariance in the sense that for any finite set of orthonormal vectors $\{e_k\} \subset L^2$ we have that the family $\{\mathbb{W}(e_k)\}$ is a multivariate standard normal variable, and as a consequence of the Kolmogorov consistency theorem (Proposition 4.1.4), $\mathbb{W}$ and $\mathbb{Y}^{(n)}$ indeed define Gaussian processes on L^2.

The fact that the model (4.5) can be interpreted as a Gaussian process indexed by L^2 means that the natural sample space $\mathcal{Y}$ in which dY from (4.5) takes values is the 'path' space $\mathbb{R}^{L^2([0,1])}$. This space may be awkward to work with in practice. In Section 6.1 .1 we shall show that we can find more tractable choices for $\mathcal{Y}$ where dY concentrates with probability 1 .

Gaussian Sequence Space Model

Again, to observe the stochastic process $\left\{\mathbb{Y}_f^{(n)}(g) : g \in L^2\right\}$ just means that we observe $\mathbb{Y}_f^{(n)}(g)$ for all $g \in L^2$ simultaneously. In particular, we may pick any orthonormal basis $\{e_k : k \in \mathbb{Z}\}$ of L^2, giving rise to an observation in the Gaussian sequence space model

$$Y_k \equiv Y_{f,k}^{(n)} = \langle f, e_k \rangle + \frac{\sigma}{\sqrt{n}} g_k, \quad k \in \mathbb{Z}, \quad n \in \mathbb{N} \tag{4.8}$$

where the g_k are i.i.d. of law $\mathbb{W}(e_k) \sim N\left(0, \|e_k\|_2^2\right) = N(0,1)$. Here we observe all the basis coefficients of the unknown function f with additive Gaussian noise of variance σ^2/n. Note that since the $\{e_k : k \in \mathbb{Z}\}$ realise a sequence space isometry between L^2 and the sequence space ℓ_2 of all square-summable infinite sequences through the mapping $f \mapsto \langle f, e_k \rangle$, the law of $\left\{Y_{f,k}^{(n)} : k \in \mathbb{Z}\right\}$ completely characterises the finite-dimensional distributions, and thus the law, of the process $\mathbb{Y}_f^{(n)}$. Hence, models (4.5) and (4.8) are observationally equivalent to each other, and we can prefer to work in either one of them (see also Theorem 4.0.1).

We note that the random sequence $Y = (Y_k : k \in \mathbb{Z})$ itself does not take values in ℓ_2, but we can view it as a random variable in the 'path' space $\mathbb{R}^{\ell}$. A more tractable, separable sample space on which $(Y_k : k \in \mathbb{Z})$ can be realised is discussed in Section 6.1.1.

A special case of the Gaussian sequence model is obtained when the space is restricted to n coefficients

$$Y_k = \theta_k + \frac{\sigma}{\sqrt{n}} g_k, \quad k = 1, \ldots, n \tag{4.9}$$

where the θ_k are equal to the $\langle f, e_k \rangle$. This is known as the **normal means model**. While itself a finite-dimensional model, it cannot be compared to the standard Gaussian linear model from the preceding section as its dimension increases as fast as n. In fact, for most parameter spaces that we will encounter in this book, the difference between model (4.9) and model (4.8) is negligible, as follows, for instance, from inspection of the proof of Theorem 1.2.1.

Multivariate Gaussian Models

To define a Gaussian white noise model for functions of several variables on $[0,1]^d$ through the preceding construction is straightforward. We simply take, for $f \in L^2\left([0,1]^d\right)$,

$$dY(t) = f(t)dt + \frac{\sigma}{\sqrt{n}}dW(t), \quad t \in [0,1]^d, \quad n \in \mathbb{N}, \quad \sigma > 0 \tag{4.10}$$

where dW is defined through the action

$$g \mapsto \int_{[0,1]^d} g(t)dW(t) \equiv \mathbb{W}(g) \sim N\left(0, \|g\|_2^2\right) \tag{4.11}$$

on elements g of $L^2\left([0,1]^d\right)$, which corresponds to multivariate stochastic integrals with respect to independent Brownian motions $W_1(t_1), \ldots, W_d(t_d)$. Likewise, we can reduce to a sequence space model by taking an orthonormal basis $\left\{e_k : k \in \mathbb{Z}^d\right\}$ of $L^2\left([0,1]^d\right)$.

Equivalence of Statistical Experiments

It is time to build a bridge between the preceding abstract models and the statistically more intuitive nonparametric fixed-design regression model (4.4). Some experience with the preceding models reveals that a statistical inference procedure in any of these models constructively suggests a procedure in the others with comparable statistical properties. Using a suitable notion of distance between statistical experiments, this intuition can be turned into a theorem, as we show in this subsection. We present results for Gaussian regression models; the general approach, however, can be developed much further to show that even highly non-Gaussian models can be, in a certain sense, asymptotically equivalent to the standard Gaussian white noise model (4.5). This gives a general justification for a rigorous study of the Gaussian white noise model in itself. Some of the proofs in this subsection require material from subsequent chapters, but the main ideas can be grasped without difficulty.

The Le Cam Distance of Statistical Experiments

We employ a general notion of distance between statistical experiments $\mathcal{E}^{(i)}, i = 1,2$, due to Le Cam. Each experiment $\mathcal{E}^{(i)}$ consists of a sample space $\mathcal{Y}_i$ and a probability measure $P_f^{(i)}$ defined on it, indexed by a common parameter $f \in \mathcal{F}$. Let $\mathcal{T}$ be a measurable space of 'decision rules', and let

$$L : \mathcal{F} \times \mathcal{T} \to [0, \infty)$$

be a 'loss function' measuring the performance of a decision procedure $T^{(i)}\left(Y^{(i)}\right) \in \mathcal{T}$ based on observations $Y^{(i)}$ in experiment i. For instance, $T^{(i)}\left(Y^{(i)}\right)$ could be an estimator for f so that $\mathcal{T} = \mathcal{F}$ and $L(f,T) = d(f,T)$, where d is some metric on $\mathcal{F}$, but other scenarios are possible. The risk under $P_f^{(i)}$ for this loss is the $P_f^{(i)}$ -expectation of $L\left(f, T^{(i)}\left(Y^{(i)}\right)\right)$, denoted by $R^{(i)}\left(f,T^{(i)},L\right)$. Define also

$$|L| = \sup\{L(f,T) : f \in \mathcal{F}, T \in \mathcal{T})$$

The Le Cam distance between two experiments is defined as

$$\begin{aligned}\Delta_{\mathcal{F}}\left(\mathcal{E}^{(1)},\mathcal{E}^{(2)}\right) \equiv \max \Bigg[& \sup_{T^{(2)}} \inf_{T^{(1)}} \sup_{f,L:|L|=1} \left|R^{(1)}\left(f,T^{(1)},L\right) - R^{(2)}\left(f,T^{(2)},L\right)\right|, \\ & \sup_{T^{(1)}} \inf_{T^{(2)}} \sup_{f,L:|L|=1} \left|R^{(1)}\left(f,T^{(1)},L\right) - R^{(2)}\left(f,T^{(2)},L\right)\right|\Bigg]\end{aligned} \tag{4.12}$$

If this quantity equals zero, this means that any decision procedure $T^{(1)}$ in experiment $\mathcal{E}^{(1)}$ can be translated into a decision procedure $T^{(2)}$ in experiment $\mathcal{E}^{(2)}$, and vice versa, and that the statistical performance of these procedures in terms of the associated risk $R^{(i)}$ will be the same for any bounded loss function L. If the distance is not zero but small, then, likewise, the performance of the corresponding procedures in both experiments will differ by at most their Le Cam distance.

Some useful observations on the Le Cam distance are the following: if both experiments have a common sample space $\mathcal{Y}^{(1)} = \mathcal{Y}^{(2)} = \mathcal{Y}$ equal to a complete separable metric space, and if the probability measures $P_f^{(1)}, P_f^{(2)}$ have a common dominating measure μ on $\mathcal{Y}$, then

$$\Delta_{\mathcal{F}}\left(\mathcal{E}^{(1)},\mathcal{E}^{(2)}\right) \le \sup_{f\in\mathcal{F}} \int_{\mathcal{Y}} \left|\frac{dP_f^{(1)}}{d\mu} - \frac{dP_f^{(2)}}{d\mu}\right| d\mu \equiv \left\|P^{(1)} - P^{(2)}\right\|_{1,\mu,\mathcal{F}} \tag{4.13}$$

This follows from the fact that in this case we can always use the decision rule $T^{(2)}(Y)$ in experiment $\mathcal{E}^{(1)}$ and vice versa and from

$$\left|R^{(1)}(f,T,L) - R^{(2)}(f,T,L)\right| \le \int_{\mathcal{Y}} |L(f,T(Y))| \left|dP_f^{(1)}(Y) - dP_f^{(2)}(Y)\right| \le |L| \left\|P^{(1)} - P^{(2)}\right\|_{1,\mu,\mathcal{F}}$$

The situation in which the two experiments are not defined on the sample space needs some more thought. Suppose, in the simplest case, that we can find a bi-measurable isomorphism B of $\mathcal{Y}^{(1)}$ with $\mathcal{Y}^{(2)}$, independent of f, such that

$$P_f^{(2)} = P_f^{(1)} \circ B^{-1}, \quad P_f^{(1)} = P_f^{(2)} \circ B \quad \forall f \in \mathcal{F}$$

Then, given observations $Y^{(2)}$ in $\mathcal{Y}^{(2)}$, we can use the decision rule $T^{(2)}\left(Y^{(2)}\right) \equiv T^{(1)}\left(B^{-1}\left(Y^{(2)}\right)\right)$ in $\mathcal{E}^{(2)}$, and vice versa, and the risks $R^{(i)}$ in both experiments coincide by the image measure theorem. We can conclude in this case that

$$\Delta_{\mathcal{F}}\left(\mathcal{E}^{(1)},\mathcal{E}^{(2)}\right) = \Delta_{\mathcal{F}}\left(\mathcal{E}^{(1)},B^{-1}\left(\mathcal{E}^{(2)}\right)\right) = 0 \tag{4.14}$$

In the absence of such a bijection, the theory of sufficient statistics can come to our aid to bound the Le Cam distance. Let again $\mathcal{Y}^{(i)}, i = 1,2$, be two sample spaces that we assume to be complete separable metric spaces. Let $\mathcal{E}^{(1)}$ be the experiment giving rise to observations $Y^{(1)}$ of law $P_f^{(1)}$ on $\mathcal{Y}^{(1)}$, and suppose that there exists a mapping $S : \mathcal{Y}^{(1)} \rightarrow \mathcal{Y}^{(2)}$ independent of f such that

$$Y^{(2)} = S\left(Y^{(1)}\right), \quad Y^{(2)} \sim P_f^{(2)} \quad \text{on } \mathcal{Y}^{(2)}$$

Assume, moreover, that $S\left(Y^{(1)}\right)$ is a sufficient statistic for $Y^{(1)}$; that is, the conditional distribution of $Y^{(1)}$ given that we have observed $S\left(Y^{(1)}\right)$ is independent of $f \in \mathcal{F}$. Then

$$\Delta_{\mathcal{F}}\left(\mathcal{E}^{(1)}, \mathcal{E}^{(2)}\right) = 0 \tag{4.15}$$

The proof of this result, which is an application of the sufficiency principle from statistics, is left as Exercise 1.1.

Asymptotic Equivalence for Nonparametric Gaussian Regression Models

We can now give the main result of this subsection. We shall show that the experiments

$$Y_i = f(x_i) + \varepsilon_i, \quad x_i = \frac{i}{n}, \quad \varepsilon_i \sim^{i.i.d.} N\left(0, \sigma^2\right), \quad i = 1, \ldots, n \tag{4.16}$$

and

$$dY(t) = f(t)dt + \frac{\sigma}{\sqrt{n}} dW(t), \quad t \in [0,1], \quad n \in \mathbb{N} \tag{4.17}$$

are asymptotically $(n \rightarrow \infty)$ equivalent in the sense of Le Cam distance. In the course of the proofs, we shall show that any of these models is also asymptotically equivalent to the sequence space model (4.8). Further models that can be shown to be equivalent to (4.17) are discussed after the proof of the following theorem. We define classes

$$\mathcal{F}(\alpha, M) = \left\{ f : [0,1] \rightarrow \mathbb{R}, \sup_{x \in [0,1]} |f(x)| + \sup_{x \neq y} \frac{|f(x) - f(y)|}{|x-y|^{\alpha}} \leq M \right\}$$
$$0 < \alpha \leq 1, \quad 0 < M < \infty$$

of α -Hölderian functions. Moreover, for $(x_i)_{i=1}^n$ the design points of the fixed-design regression model (4.16) and for f any bounded function defined on $[0,1]$, let $\pi_n(f)$ be the unique function that interpolates f at the x_i and that is piecewise constant on each interval $(x_{i_1}, x_i] \subset [0,1]$.

Theorem 4.0.1 Let $\left(\mathcal{E}_n^{(i)} : n \in \mathbb{N}\right), i = 1,2,3$, equal the sequence of statistical experiments given by $i = 1$ the fixed-design nonparametric regression model (4.16); $i = 2$, the standard Gaussian white noise model (4.17); and $i = 3$, the Gaussian sequence space model (4.8), respectively. Then, for $\mathcal{F}$ any family of bounded functions on $[0,1]$, for $\pi_n(f)$ as earlier and for any $n \in \mathbb{N}$

$$\Delta_{\mathcal{F}}\left(\mathcal{E}_n^{(2)}, \mathcal{E}_n^{(3)}\right) = 0, \quad \Delta_{\mathcal{F}}\left(\mathcal{E}_n^{(1)}, \mathcal{E}_n^{(2)}\right) \leq \sqrt{\frac{n\sigma^2}{2}} \sup_{f \in \mathcal{F}} \|f - \pi_n(f)\|_2 \tag{4.18}$$

In particular, if $\mathcal{F}=\mathcal{F}(\alpha, M)$ for any $\alpha>1/2, M>0$, then all these experiments are asymptotically equivalent in the sense that their Le Cam distance satisfies, as $n \rightarrow \infty$

$$\Delta_{\mathcal{F}}\left(\mathcal{E}_n^{(i)}, \mathcal{E}_n^{(j)}\right) \rightarrow 0, \quad i, j \in\{1,2,3\} \tag{4.19}$$

Proof. In the proof we shall say that two experiments are equivalent if their Le Cam distance is exactly equal to zero. The first claim in (4.18) immediately follows from (4.14) and the isometry between $L^2([0,1])$ and ℓ_2 used in the definition of the sequence space model (4.8). Define $\mathcal{V}_n$ to equal the n -dimensional space of functions $f:[0,1] \rightarrow \mathbb{R}$ that are piecewise constant on the intervals

$$I_{in}=\left(x_{i-1}, x_i\right]=\left(\frac{i-1}{n}, \frac{i}{n}\right], \quad i=1, \ldots, n$$

The indicator functions $\phi_{in}=1_{I_{in}}$ of these intervals have disjoint support, and they form an orthonormal basis of $\mathcal{V}_n$ for the inner product

$$\langle f, g\rangle_n=\sum_{j=1}^n f\left(x_j\right) g\left(x_j\right)$$

noting that $\sum_{j=1}^n \phi_{in}^2\left(x_j\right)=1$ for every i. Given bounded $f:[0,1] \rightarrow \mathbb{R}$, let $\pi_n(f)$ be the $\langle\cdot, \cdot\rangle_n$ -projection of f onto $\mathcal{V}_n$. Since

$$\left\langle f, \phi_{in}\right\rangle=\sum_{j=1}^n f\left(x_j\right) \phi_{in}\left(x_j\right)=f\left(x_i\right) \ \forall\, i$$

we see

$$\pi_n(f)(t)=\sum_{i=1}^n f\left(x_i\right) \phi_{in}(t), \quad t \in[0,1]$$

so this projection interpolates f at the design points x_i, that is, $\pi_n(f)\left(x_j\right)=f\left(x_j\right)$ for all j . Note that the functions $\left\{\sqrt{n} \phi_{in}: i=1, \ldots, n\right\}$ also form a basis of $\mathcal{V}_n$ in the standard $L^2([0,1])$ inner product $\langle\cdot, \cdot\rangle$. This simultaneous orthogonality property will be useful in what follows. Observing $Y_i=f\left(x_i\right)+\varepsilon_i$ in $\mathbb{R}^n$ from model (4.16) with bounded f is, by (4.14) , equivalent to observations in the n -dimensional functional space $\mathcal{V}_n$ given by

$$\sum_{i=1}^n Y_i \phi_{in}(t)=\sum_{i=1}^n f\left(x_i\right) \phi_{in}(t)+\sum_{i=1}^n \varepsilon_i \phi_{in}(t), \quad t \in[0,1] \tag{4.20}$$

We immediately recognise that $\sum_{i=1}^n f\left(x_i\right) \phi_{in}$ is the interpolation $\pi_n(f)$ of f at the x_i . Moreover, the error process is a scaled white noise process restricted to the space $\mathcal{V}_n$: indeed, its $L^2([0,1])$ action on $h \in \mathcal{V}_n$ is given by

$$\int_0^1 \sum_{i=1}^n \varepsilon_i \phi_{in}(t) h(t) d t=\frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i\left\langle h, \sqrt{n} \phi_{in}\right\rangle \sim N\left(0, \frac{\sigma^2}{n} \sum_{i=1}^n\left\langle h, \sqrt{n} \phi_{in}\right\rangle^2\right)=N\left(0, \frac{\sigma^2}{n}\|h\|_2^2\right)$$

using Parseval's identity and that the $\sqrt{n}\phi_{in}$ form an $L^2([0,1])$ orthonormal basis of $\mathcal{V}_n$. If Π_n is the $L^2([0,1])$ projector onto $\mathcal{V}_n$ spanned by the $\{\sqrt{n}\phi_{in}\}$, then one shows, by the same arguments, that this process can be realised as a version of the Gaussian process defined on L^2 by the action $h \mapsto \mathbb{W}(\Pi_n(h))$, where $\mathbb{W}$ is as in (4.7). In other words, it equals the L^2 -projection of the standard white noise process dW onto the finite-dimensional space $\mathcal{V}_n$ justifying the notation

$$\frac{\sigma}{\sqrt{n}} dW_n(t) \equiv \sum_{i=1}^{n} \varepsilon_i \phi_{in}(t) dt$$

To summarise, (4.16) is equivalent to model (1.20), which itself can be rewritten as

$$d\widetilde{Y}(t) \equiv \pi_n(f)(t) + \frac{\sigma}{\sqrt{n}} dW_n(t), \quad t \in [0,1] \tag{4.21}$$

Next, consider the model

$$d\overline{Y}(t) = \pi_n(f)(t) + \frac{\sigma}{\sqrt{n}} dW(t), \quad t \in [0,1] \tag{4.22}$$

which is the standard white noise model (4.17) but with f replaced by its interpolation $\pi_n(f)$ at the design points x_i. since $\pi_n(f) \in \mathcal{V}_n$, we have $\Pi_n(\pi_n(f)) = \pi_n(f)$, and since $dW_n = \Pi_n(dW) \in \mathcal{V}_n$, the statistics

$$d\widetilde{Y} = \Pi_n(d\overline{Y}) = \left\{ \int_0^1 h(t) d\widetilde{Y}(t) : h \in \mathcal{V}_n \right\}$$

are sufficient for $d\bar{Y}$, so by (4.15) the models (4.21) and (4.22) are equivalent. [To use (4.15) rigorously, we interpret $d\tilde{Y}, d\bar{Y}$ as tight random variables in a large enough, separable Banach space (see Section 6.1 .1).] To prove the second claim in (1.18), we relate (4.22) to (1.17), that is, to

$$dY(t) = f(t) + \frac{\sigma}{\sqrt{n}} dW(t), \quad t \in [0,1]$$

Both experiments have the same sample space, which in view of Section 6.1.1 we can take to be, for instance, the space of continuous functions on $[0,1]$, and the standard white noise W gives a common dominating measure P_0^Y on that space for the corresponding probability measures $P_f^Y, P_{\pi_n(f)}^Y$. In view of (4.13) and using Proposition 6.1.7a) combined with (**??**), we see that the Le Cam distance is bounded by

$$\sup_{f \in \mathcal{F}} \left\| P_f^Y - P_{\pi_n(f)}^Y \right\|_{1,\mu,\mathcal{F}}^2 \leq \frac{n}{\sigma^2} \sup_{f \in \mathcal{F}} \| f - \pi_n(f) \|_2^2 \tag{4.23}$$

which gives (4.18). Finally, for (4.19), uniformly in $f \in \mathcal{F}(\alpha, M)$,

$$\begin{aligned} \| f - \pi_n(f) \|_2^2 &= \sum_{i=1}^{n} \int_{(i-1)/n}^{i/n} (f(x) - f(x_i))^2 dx \leq M^2 \sum_{i=1}^{n} \int_{(i-1)/n}^{i/n} |x - x_i|^{2\alpha} dx \\ &\leq M^2 n^{-2\alpha} \sum_{i=1}^{n} \int_{(i-1)/n}^{i/n} dx = O\left(n^{-2\alpha}\right) \end{aligned}$$

so for $\alpha > 1/2$, the quantity in (4.23) converges to zero, completing the proof. ■

In the preceding theorem the Hölder classes $\mathcal{F}(\alpha, M)$ could be replaced by balls in the larger Besov-Sobolev spaces $B^{\alpha}_{2\infty}$ (defined in Chapter 4) whenever $\alpha > 1/2$. The condition on α, however, cannot be relaxed, as we discuss in the notes.

The theory of asymptotic equivalence can be taken much further, to include results like the one preceding for random design regression experiments in possibly multivariate settings and with possibly non-Gaussian noise ε. The theory also extends to non-Gaussian settings that are not of regression type: one can show that nonparametric models for probability or spectral densities, or ergodic diffusions, are asymptotically equivalent to a suitable Gaussian white noise model. We discuss relevant references in the notes.

Asymptotic equivalence theory, which is a subject in its own, justifies that the Gaussian white noise model is, in the sense of the Le Cam distance, a canonical limit experiment in which one can develop some main theoretical ideas of nonparametric statistics. For Gaussian regression problems, the closeness of the experiments involved is in fact of a nonasymptotic nature, as shown by Theorem 4.0.1, and in this book we thus shall concentrate on the white noise model as the natural continuous surrogate for the standard fixed-design regression model. For other, non-Gaussian models, such as density estimation, asymptotic equivalence theory is, however, often overly simplistic in its account of the probabilistic structure of the problem at hand, and for the purposes of this book, we hence prefer to stay within the product-measure setting of Section 1.1, such that a nonasymptotic analysis is possible.

The modern understanding of statistical inference as consisting of the three related branches of estimation, testing and confidence statements probably goes back, in its most fundamental form, to the work of Fisher (1922;1925a,b), who considered mostly parametric (finite-dimensional) statistical models. The need to investigate nonparametric statistical models was realised not much later, roughly at the same time at which the axiomatic approach to probability theory was put forward by Kolmogorov (1933). Classic papers on fully nonparametric sampling models for the cumulative distribution function are, for instance, Glivenko (1933), Cantelli (1933), Kolmogorov (1933a), and Smirnov (1939). More recent developments will be reviewed in later chapters of this book.

The linear regression model with normally distributed errors was initiated by Gauss (1809), who used it successfully in the context of observational astronomy. Gauss most likely was the first to use the least-squares algorithm, although Legendre and even some others can claim priority as well. The history is reviewed, for example, in Plackett (1972) and Stigler (1981) .

Nonparametric regression models were apparently not studied systematically before the 1960 s; see Nadaraya (1964) and Watson (1964). The Gaussian white noise model and its sequence space analogue were systematically developed in the 1970 s and later by the Russian school - we refer to the seminal monograph by Ibragimov and Khasminskii

(1981). The asymptotic equivalence theory for statistical experiments was developed by Le Cam; we refer to his fundamental book Le Cam (1986) and also to Le Cam and Yang (1990). Landmark contributions in nonparametric asymptotic equivalence theory are the papers Brown and Low (1996) and Nussbaum (1996), who treated univariate regression models with fixed design and density estimation, respectively. The necessity of the assumption $\alpha \geq 1/2$ is the subject of the paper by Brown and Zhang (1998) Asymptotic equivalence for random design regression is somewhat more involved: the univariate case is considered in Brown et al. (2002), and the general, multivariate random design regression case is considered in ReiSS (2008) . Further important results include asymptotic equivalence for nonparametric regression with non-Gaussian error distributions in Grama and Nussbaum (2002), asymptotic equivalence for spectral density estimation in Golubev, Nussbaum and Zhou (2010), and asymptotic equivalence for ergodic diffusions in Dalalyan and ReiSS (2006)

4.1 Gaussian Processes

This chapter develops some classical theory and fundamental tools for Gaussian random processes. We start with the basic definitions of Gaussian processes indexed by abstract parameter spaces and, by way of introduction to the subject, derive some elementary yet powerful properties. We present the isoperimetric and log-Sobolev inequalities for Gaussian measures in $\mathbb{R}^n$ and apply them to establish concentration properties for the supremum of a Gaussian process about its median and mean, which are some of the deepest and most useful results on Gaussian processes. Then we introduce Dudley's metric entropy bounds for moments of suprema of (sub-) Gaussian processes as well as for their a.s. modulus of continuity. The chapter also contains a thorough discussion of convexity and comparison properties of Gaussian measures and of reproducing kernel Hilbert spaces and ends with an exposition of the limit theory for suprema of stationary Gaussian processes.

4.1.1 Definitions, Separability, 0-1 Law, Concentration

We start with some preliminaries about stochastic processes, mainly to fix notation and terminology. Then these concepts are specialised to Gaussian processes, and some first properties of Gaussian processes are developed. The fundamental observation is that a Gaussian process X indexed by a a set T induces an intrinsic distance d_X on T $(d_X(s,t)$ is the L^2 -distance between $X(s)$ and $X(t)$), and all the probabilistic information about X is contained in the metric or pseudo-metric space (T,d). This is tested on some of the first properties, such as the $0-1$ law and the existence of separable versions of X. One of the main properties of Gaussian processes, namely, their concentration about the mean, is introduced; this subject will be treated in the next section, but a first result on it, which is not sharp but that has been chosen for its simplicity, is given in this section.

Stochastic Processes: Preliminaries and Definitions

Let $(\Omega,\Sigma,\text{ Pr })$ be a probability space, and let T be a set. A stochastic process X indexed by T and defined on the probability space $(\Omega,\Sigma,\text{Pr})$ is a function $X : T \times \Omega \mapsto \mathbb{R},(t,\omega) \mapsto X(t,\omega)$ such that, for each $t \in T, X(t,\cdot)$ is a random variable. Then, for any finite set $F \subset T$, the maps $\Omega \mapsto \mathbb{R}^F$ given by $\omega \mapsto \{X(t,\omega) : t \in F\}$ are also measurable, and their probability laws $\mu_F = \text{Pro}\{X(t,\cdot) : t \in F\}^{-1}$ are the finite-dimensional distributions (or finite-dimensional marginal distributions or finite-dimensional marginals) of X. If $F \subset G \subset T$ and G is finite and π_{GF} is the natural projection from $\mathbb{R}^G$ onto $\mathbb{R}^F$, then, obviously, the consistency

conditions $\mu_F = \mu_G \circ \pi_{GF}^{-1}$ are satisfied $(\pi_{GF}(\{X(t) : t \in G\}) = \{X(t) : t \in F\})$. Conversely, the Kolmogorov consistency theorem shows that any collection of Borel probability measures μ_F on $\mathbb{R}^F$, indexed by the finite subsets $F \subset T$ and satisfying the consistency conditions, is the collection of finite-dimensional distributions of a stochastic process X indexed by T. In other words, a consistent family of probability measures $\mu_F, F \subset T, F$ finite, defines a unique probability measure μ on the cylindrical σ -algebra $\mathcal{C}$ of $\mathbb{R}^T$ such that $\mu_F = \mu \circ \pi_{TF}^{-1}$. (The cylindrical σ -algebra $\mathcal{C}$ is the σ -algebra generated by the cylindrical sets with finite-dimensional base, $\pi_{TF}^{-1}(A), A \in \mathcal{B}\left(\mathbb{R}^F\right), F \subset T, F$ finite.) Then the map $X : T \times \mathbb{R}^T \mapsto \mathbb{R},(t,x) \mapsto x(t)$, is a process defined on the probability space $\left(\mathbb{R}^T,\mathcal{C},\mu\right)$ If μ is the probability measure on $\left(\mathbb{R}^T,\mathcal{C}\right)$ defined by the finite-dimensional distributions of a process X, then we say that μ is the probability law of X (which can be thought of as a 'random variable' taking values on the measurable space $\left(\mathbb{R}^T,\mathcal{C}\right)$). See almost any probability textbook, for example, Dudley (2002).

Definition 4.1.1 Two processes X and Y of index set T are said to be a version of each other if both have the same finite-dimensional distributions $\mathcal{L}\left(X\left(t_1\right),\ldots,X\left(t_n\right)\right) = \mathcal{L}\left(Y\left(t_1\right),\ldots,Y\left(t_n\right)\right)$ for all $n \in \mathbb{N}$ and $t_i \in T$ or, what is the same, if both have the same probability law on $\left(\mathbb{R}^T,\mathcal{C}\right)$. They are said to be a strict version or a modification of each other if $\text{Pr}\{X(t) = Y(t)\} = 1$ for all t

It is convenient to recall the definition of pseudo-distance and pseudo-metric space. A pseudo-distance d on T is a nonnegative symmetric function of two variables $s,t \in T$ that satisfies the triangle inequality but for which $d(s,t) = 0$ does not necessarily imply $s = t$ A pseudo-metric space (T,d) is a set T equipped with a pseudo-distance d. Clearly, a pseudo-metric space becomes a metric space by taking the quotient with respect to the equivalence relation $s \simeq t$ iff $d(s,t) = 0$. For instance, the space $\mathcal{L}^p$ of functions is a pseudo-metric space for the L^p (pseudo-)norm, and the space of equivalence classes, L^p, is a metric space for the same norm. One only seldom needs to distinguish between the two. If the index set T of a process X is a metric or pseudo-metric space (T,d), we say that X is continuous in probability if $X(t_n) \rightarrow X(t)$ in probability whenever $d(t_n, t) \rightarrow 0$. In this case, if T_0 is a d -dense subset of T, the law of the process on $\left(\mathbb{R}^T, \mathcal{C}\right)$ is determined by the finite-dimensional distributions $\mathcal{L}(X(t_1), \ldots, X(t_n))$ for all $n \in \mathbb{N}$ and $t_i \in T_0$ Here are two more definitions of interest. Definition 2.1 .2 A process $X(t), t \in T, (T, d)$ a metric or pseudo-metric space, is separable if there exists $T_0 \subset T, T_0$ countable, and $\Omega_0 \subset \Omega$ with $\operatorname{Pr}(\Omega_0) = 1$ such that for all $\omega \in \Omega_0$ $t \in T$ and $\varepsilon > 0$

$$X(t, \omega) \in \overline{\{X(s, \omega)} : s \in T_0 \cap B_d(t, \varepsilon)\Big\}$$

where $B_d(t, \varepsilon)$ is the open d -ball about t of radius $\varepsilon . X$ is measurable if the map $(\Omega \times T, \Sigma \otimes \mathcal{T}) \rightarrow \mathbb{R}$ given by $(\omega, t) \longrightarrow X(\omega, t)$ is jointly measurable, where $\mathcal{T}$ is the σ -algebra generated by the d -balls of T.

By definition, if $X(t), \quad t \in T$, is separable, then there are points from T_0 in any neighborhood of $t, t \in T$; hence (T, d) is separable; that is, (T, d) possesses a countable dense subset. Note that if X is separable, then $\sup_{t \in T} X(t) = \sup_{s \in T_0} X(s)$ a.s., and the latter, being a countable supremum, is measurable; that is, suprema over uncountable sets are measurable. The same holds for $|X(t)|$ Often we require the sample paths $t \mapsto X(t, \omega)$ to have certain properties for almost every ω, notably, to be bounded or bounded and uniformly continuous ω a.s.

Definition 4.1.2 A process $X(t), t \in T$, is sample bounded if it has a version $\tilde{X}$ whose sample paths $t \mapsto \tilde{X}(t, \omega)$ are almost all uniformly bounded, that is, $\sup_{t \in T} |\tilde{X}(t)| < \infty$ a.s. If (T, d) is a metric or pseudo-metric space, then X is sample continuous (more properly, sample bounded and uniformly continuous) if it has a version $\tilde{X}(t)$ whose sample paths are almost all bounded and uniformly d -continuous.

Note that if X is sample continuous, then the finite-dimensional distributions of X are the marginals of a probability measure μ defined on the cylindrical σ -algebra $\mathcal{C} \cap C_u(T, d)$ of $C_u(T, d)$, the space of bounded uniformly continuous functions on (T, d) $\mathcal{L}(X(t_1), \ldots, X(t_k)) = \mu \circ (\delta_{t_1}, \ldots, \delta_{t_k})^{-1}, t_i \in T, k < \infty$ (here and in what follows, δ_t is unit mass at t). The vector space $C_u(T, d)$, equipped with the supremum norm $\|f\|_\infty = \sup_{t \in T} |f(t)|$, is a Banach space, that is, a complete metric space for which the vector space operations are continuous. The Banach space $C_u(T, d)$ is separable if (and only if) (T, d) is totally bounded, and in this case, $C_u(T, d)$ is isometric to $C(\bar{T}, d)$, where $(\bar{T}, d)$ is the completion of (T, d), which is compact. Then, assuming (T, d) totally bounded, we have $\|f\|_\infty = \sup_{t \in T_0} |f(t)|$, where T_0 is any countable dense subset of T; in particular, the closed balls of $C_u(T, d)$ are measurable for the cylindrical σ -algebra: $\{f : \|f - f_0\|_\infty \leq r\} = \cap_{t \in T_0} \{f : |f(t) - f_0(t)| \leq r\}$. This implies that the open sets are also measurable because, by separability of $C_u(T, d)$, every open set in this space is the union of a countable number of closed balls. This proves that the Borel and the cylindrical σ -algebras of $C_u(T, d)$ coincide if (T, d) is totally bounded. Hence, in this case, the finite-dimensional distributions of X are the marginal measures of a Borel probability measure μ on $C_u(T,d)$. since $C_u(T,d)$ is separable and complete (for the supremum norm), the probability law μ of X is tight in view of the following basic result that we shall use frequently in this book (see Exercise 2.6 for its proof). Recall that a probability measure μ is tight if for all $\varepsilon > 0$ there is K compact such that $\mu\left(K^c\right) < \varepsilon$

Proposition 4.1.1 **— Oxtoby-Ulam.** If μ is a Borel probability measure on a complete separable metric space, then μ is tight.

In general, given a Banach space B, a B -valued random variable X is a Borel measurable map from a probability space into B. Thus, the preceding considerations prove the following proposition. It is convenient to introduce an important Banach space: given a set $T, \ell_\infty\ (T) \subset \mathbb{R}^T$ will denote the set of bounded functions $x : T \mapsto \mathbb{R}$. Note that this is a Banach space if we equip it with the supremum norm $\|x\|_T = \sup_{t\in T}|x(t)|$ and that the inclusion of $C_u(T)$ into $\ell_\infty(T)$ is isometric. Observe that $\ell_\infty(T)$ is separable for the supremum norm if and only if T is finite.

Proposition 4.1.2 If (T,d) is a totally bounded metric or pseudo-metric space and $X(t)$ $t \in T$, is a sample continuous process, then X has a version which is a $C_u(T,d)$ -valued random variable, and its probability law is a tight Borel measure with support contained in $C_u(T,d)$ and hence a tight Borel probability measure on $\ell_\infty(T)$

■ **Example 4.1 — Banach space-valued random variables as sample continuous processes..** Let B be a separable Banach space, let B^* be its dual space and let B_1^* denote the

closed-unit ball of B_1^* about the origin. Then there exists a countable set $D \subset B_1^*$ such that $\|x\| = \sup_{f\in D} f(x)$ for all $x \in B$: if $\{x_i\} \subset B$ is a countable dense subset of B and $f_i \in B_1^*$ are such that $f_i\left(x_i\right) = \|x_i\|$ (note that f_i exists by the Hahn-Banach theorem), then $D = \{f_i\}$ is such a set. The inclusion $B \mapsto C_u(D, \|\cdot\|)$, where $\|\cdot\|$ is the norm on B_1^*, is an isometric imbedding, and every B -valued random variable X defines a process $f \mapsto f(X), f \in D$, with all its sample paths bounded and uniformly continuous. Hence, any results proved for sample bounded and uniformly continuous processes indexed by totally bounded metric spaces do apply to Banach space-valued random variables for B separable.

If $X(t), t \in T$, is a sample bounded process, then its probability law is defined on the cylindrical σ -algebra of $\ell_\infty(T), \Sigma = \mathcal{C} \cap \ell_\infty(T)$. since $\ell_\infty(T)$ is a metric space for the supremum norm, it also has another natural σ -algebra, the Borel σ -algebra. We conclude with the interesting fact that if the law of the bounded process X extends to a tight Borel measure on $\ell_\infty(T)$, then X is sample continuous with respect to a metric d for which (T,d) is totally bounded.

Proposition 4.1.3 Let $X(t), \quad t \in T$, be a sample bounded stochastic process. Then the finite-dimensional probability laws of X are those of a tight Borel probability measure on $\ell_\infty(T)$ if and only if there exists on T a pseudo-distance d for which (T,d) is totally bounded and such that X has a version with almost all its sample paths uniformly continuous for d.

Proof. Let us assume that the probability law of X is a tight Borel measure μ on $\ell_\infty(T)$; let $K_n, n \in \mathbb{N}$, be an increasing sequence of compact sets in $\ell_\infty(T)$ such that $\mu\left(\cup_{n=1}^{\infty} K_n\right) = 1$ and set $K = \cup_{n=1}^{\infty} K_n$. Define a pseudo-metric d as

$$d(s,t) = \sum_{n=1}^{\infty} 2^{-n}\left(1 \wedge d_n(s,t)\right)$$

where

$$d_n(s,t) = \sup\left\{|f(t) - f(s)| : f \in K_n\right\}$$

To prove that (T,d) is totally bounded, given $\varepsilon > 0$, let m be such that $\sum_{n=m+1}^{\infty} 2^{-n} < \varepsilon/4$ since the set $\cup_{n=1}^{m} K_n$ is compact, it is totally bounded, and therefore, it contains a finite subset $\{f_1, \ldots, f_r\}$ which is $\varepsilon/4$ dense in $\cup_{n=1}^{m} K_n$ for the supremum norm; that is, for each $f \in \cup_{n=1}^{m} K_n$, there is $i \leq r$ such that $\|f - f_i\|_\infty \leq \varepsilon/4$. since $\cup_{n=1}^{m} K_n$ is a bounded subset of $\ell_\infty(T)$ (as it is compact), it follows that the subset $A = \{(f_1(t), \ldots, f_r(t)) : t \in T\}$ of $\mathbb{R}^r$ is bounded, hence precompact, hence totally bounded, and therefore there exists a finite set $T_\varepsilon = \{t_i : 1 \leq i \leq N\}$ such that for each $t \in T$ there is $i = i(t) \leq N$ such that $\max_{1 \leq s \leq r} | f_s(t) - f_s(t_i) | \leq \varepsilon/4$. It follows that T_ε is ε dense in T for the pseudo-metric d : for $n \leq m, t \in T$ and $t_i = t_{i(t)}$, we have

$$d_n(t,t_i) = \sup_{f \in K_n} |f(t) - f(t_i)| \leq \max_{s \leq r} |f_s(t) - f_s(t_i)| + \varepsilon/2 \leq \frac{3\varepsilon}{4}$$

and therefore

$$d(t,t_i) \leq \frac{\varepsilon}{4} + \sum_{n=1}^{m} 2^{-n} d_n(t,t_i) \leq \varepsilon$$

proving that (T,d) is totally bounded. Next, since $\mu(K) = 1$, the identity map of $(\ell_\infty(T), \mathcal{B}, \mu)$ is a version of X with almost all its trajectories in K. Thus, to prove that X has a version with almost all its sample paths bounded and uniformly d -continuous, it suffices to show that the functions from K have these properties. If $f \in K_n$, then $|f(s) - f(t)| \leq d_n(s,t) \leq 2^n d(s,t)$ for all $s,t \in T$ with $d(s,t) < 2^{-n}$, proving that f is uniformly continuous, and f is bounded because K_n is bounded. Conversely, let $X(t), t \in T$, be a process with a version whose sample paths are almost all in $C_u(T,d)$ for a distance or pseudo-distance d on T for which (T,d) is totally bounded, and let us continue denoting X such a version (recall the notation $C_u(T,d)$ as the space of bounded uniformly continuous functions on (T,d)).

Then X is a random variable taking values in $C_u(T,d)$, and its marginal laws correspond to a Borel probability measure on $C_u(T,d)$ (see the argument following Definition 2.1 .3). But since (T,d) is precompact, $C_u(T,d)$ is separable, and the law of X is in fact a tight Borel measure by the Oxtoby-Ulam theorem (Proposition 2.1.4). But a tight Borel probability measure on $C_u(T,d)$ is a tight Borel measure on $\ell_\infty(T)$ because the inclusion of $C_u(T,d)$ into ℓ_∞ is continuous.

■

Gaussian Processes: Introduction and First Properties

We now look at Gaussian processes. Recall that a finite-dimensional random vector or a multivariate random variable $Z=(Z_1,\ldots,Z_n), n\in\mathbb{N}$, is an n -dimensional Gaussian vector, or a multivariate normal random vector, or its coordinates are jointly normal, if the random variables $\langle a,Z\rangle=\sum_{i=1}^n a_i Z_i, a=(a_1,\ldots,a_n)\in\mathbb{R}^n$, are normal variables, that is, variables with laws $N\left(m(a),\sigma^2(a)\right), \sigma(a)\geq 0, m\in\mathbb{R}$. If $m=m(a)=0$ for all $a\in\mathbb{R}^n$, we say that the Gaussian vector is centred.

Definition 4.1.3 A stochastic process $X(t), t\in T$, is a Gaussian process if for all $n\in\mathbb{N}$, $a_i\in\mathbb{R}$ and $t_i\in T$, the random variable $\sum_{i=1}^n a_i X\left(t_i\right)$ is normal or, equivalently, if all the finite-dimensional marginals of X are multivariate normal. X is a centred Gaussian process if all these random variables are normal with mean zero.

Definition 4.1.4 A covariance Φ on T is a map $\Phi: T\times T\to\mathbb{R}$ such that for all $n\in\mathbb{N}$ and $t_1,\ldots,t_n\in T$, the matrix $\left(\Phi\left(t_i,t_j\right)\right)_{i,j=1}^n$ is symmetric and nonnegative definite (i.e., $\Phi\left(t_i,t_j\right)=\Phi\left(t_j,t_i\right)$ and $\sum_{i,j} a_i a_j \Phi\left(t_i,t_j\right)\geq 0$ for all a_i)

The following is a consequence of the Kolmogorov consistency theorem.

Proposition 4.1.4 Given a covariance Φ on T and a function f on T, there is a Gaussian process $X(t)$ such that $E(X(t))=f(t)$ and $E[(X(t)-f(t))(X(s)-f(s))]=\Phi(s,t)$ for all $s,t\in T.\Phi$ is called the covariance of the process and f its expectation, and we say that X is a centred Gaussian process if and only if $f\equiv 0$.

Proof. If $F\subset T$ is finite, take $\mu_F=N\left((f(t):t\in F),\Phi|_{F\times F}\right)$. It is easy to see that the set $\{\mu_F: F\subset T, F \text{ finite }\}$ is a consistent system of marginals. Hence, by the Kolmogorov consistency theorem, there is a probability on $\left(\mathbb{R}^T,\mathcal{C}\right)$, hence a process, with $\{\mu_F\}$ as its set of finite-dimensional marginals. ■

▪ **Example 4.2** A basic example of a Gaussian process is the isonormal or white noise process on a separable Hilbert space H, where $\{X(h): h\in H\}$ has a covariance diagonal for the inner product $\langle\cdot,\cdot\rangle$ of $H: EX(h)=0$ and $EX(h)X(g)=\langle h,g\rangle_H$ for all $g,h\in H$. The existence of this process does not even require the Kolmogorov consistency theorem but only the existence of an infinite sequence of random variables (i.e., the existence of an infinite product probability space): if $\{g_i\}$ is a sequence of independent $N(0,1)$ random variables and $\{\psi_i\}$ is an orthonormal basis of H, the process defined by linear and continuous extension of $\tilde{X}(\psi_i) = g_i$ (i.e., by $\tilde{X}(\sum a_i\psi_i) = \sum a_i g_i$ whenever $\sum a_i^2 < \infty$) is clearly a version of X. Note for further use that if $V \subset L^2(\Omega,\Sigma,\Pr)$ is the closed linear span of the sequence $\{g_i\}$, then the map $\tilde{X} : H \mapsto V$ is an isometry.

From now on, all our Gaussian processes will be centred, even if sometimes we omit mentioning it. If X is a centred Gaussian process on T, the L^2 -pseudo-distance between $X(t)$ and $X(s)$ defines a pseudo-distance d_X on T

$$d_X^2(s,t) := E(X(t) - X(s))^2 = \Phi(t,t) + \Phi(s,s) - 2\Phi(s,t)$$

that we call the intrinsic distance of the process. With this pseudo-metric, T is isometric to the subspace $\{X(t) : t \in T\}$ of $L^2(\Omega,\Sigma,\Pr)$. Clearly, a centred Gaussian process X is continuous in probability for the pseudo-distance d_X; in particular, its probability law in $(\mathbb{R}^T,\mathcal{C})$ is determined by the finite-dimensional marginals based on subsets of any d_X-dense subset T_0 of T.

It is important to note that the probability law of a centred Gaussian process X is completely determined by its intrinsic distance d_X (or by the covariance Φ). Thus, all the probabilistic information about a centred Gaussian process is contained in the metric (or pseudo-metric) space (T,d_X). This is a very distinctive feature of Gaussian processes.

Here is a first, albeit trivial, example of the exact translation of a property of the metric space (T,d_X) into a probabilistic property of X, actually, necessarily of a version of X.

Proposition 4.1.5 For a Gaussian process X indexed by T, the following are equivalent:

1. The pseudo-metric space (T,d_X) is separable, and
2. X, as a process on (T,d_X), has a separable, measurable (strict) version.

Proof. If point 2 holds, let $\bar{X}$ be a separable and measurable version of X (in particular, $d_{\bar{X}} = d_X$), and let T_0 be a countable set as in the definition of separability. Then, as mentioned earlier, the very definition of separability implies that $T_0 \cap B_{d_X}(t,\varepsilon) \neq \emptyset$ for all $t \in T$ and $\varepsilon > 0$. Thus, T_0 is dense in (T,d_X), and therefore, (T,d_X) is separable.

Assume now that (T,d_X) is separable, and let T_0 be a countable d_X -dense subset of T. Also assume, as we may by taking equivalence classes, that $d_X(s,t) \neq 0$ for all $s,t \in T_0, s \neq t$. If $T_0 = \{s_i : i \in \mathbb{N}\}$, define, for each n, the following partition of T :

$$C_n(s_m) = B(s_m,2^{-n}) \setminus \bigcup_{k<m} B(s_k,2^{-n}), \quad m \in \mathbb{N}$$

For each $t \in T$, let $s_n(t)$ be the only $s \in T_0$ such that $t \in C_n(s)$, and define $X_n(t) = X(s_n(t))$ Now $X_n(t,\omega)$ is jointly measurable because $X_n^{-1}(A) = \bigcup_{i\in\mathbb{N}} [C_n(s_i) \times \{\omega : X(s_i,\omega) \in A\}]$

since, for any $t \in T, \Pr\{|X_n(t) - X(t)| > 1/n\} \leq n^2 E(X(s_n(t)) - X(t))^2 \leq n^2/2^{2n}$, it follows by Borel-Cantelli that $X_n(t) \to X(t)$ a.s.

Define $\bar{X}(t,\omega) = \limsup_n X_n(t,\omega)$, which, for each t, is ∞ at most on a set of measure 0. Then the process $\bar{X}(t,\omega)$ is measurable because it is a limsup of measurable functions. Also, for each $t, \bar{X}(t) = X(t)$ on a set of measure 1; that is, $\bar{X}$ is a strict version of X. Next we show that $\bar{X}$ is separable. Given $r \in \mathbb{N}$, there exists n_r large enough that $d_X(s_r, s_l) > 1/2^{n_r}$ for all $l < r$; hence, for $n \geq n_r, X_n(s_r) = X(s_r)$. This shows that $\bar{X}(s) = X(s)$ for all $s \in T_0$. Then, for all $\omega \in \Omega$

$$\bar{X}(t,\omega) = \limsup X_n(t,\omega) = \limsup X(s_n(t),\omega) = \limsup \bar{X}(s_n(t),\omega)$$

proving that $\bar{X}$ is separable. ■

Just as with normal random variables, Gaussian processes also satisfy the Gaussian stability property, namely, that if two Gaussian processes with index set T are independent, then their sum is a Gaussian process with covariance the sum of covariances (and mean the sum of means); in particular, if X and Y are independent and equally distributed Gaussian processes (meaning that they have the same finite-dimensional marginal distributions or, what is the same, the same law on the cylindrical σ -algebra $\mathcal{C}$ of $\mathbb{R}^T$), then the process $\alpha X + \beta Y$ has the same law as $(\alpha^2 + \beta^2)^{1/2} X$. This property has many consequences, and here is a nice instance of its use.

Theorem 4.1.6 [$0 - 1$ law] Let $F \subset \mathbb{R}^T$ be a C -measurable linear subspace, and let X be a (centred) Gaussian process indexed by T. Then

$$\Pr\{X \in F\} = 0 \text{ or } 1$$

Proof. Let X_1 and X_2 be independent copies of X. Define sets

$$A_n = \{X_1 + nX_2 \in F\} \quad \text{and} \quad B_n = \{X_2 \notin F\} \cap A_n, \quad n \in \mathbb{N}$$

since $X_1 + nX_2$ is a version of $\sqrt{1+n^2}X$ and F is a vector space, we have

$$\begin{aligned} \Pr\{B_n\} &= \Pr\{A_n\} - \Pr[A_n \cap \{X_2 \in F\}] \\ &= \Pr\{X \in F\} - \Pr\{X_1 + nX_2 \in F, X_2 \in F\} \\ &= \Pr\{X \in F\} - \Pr\{X_1 \in F, X_2 \in F\} \\ &= \Pr\{X \in F\} - [\Pr\{X \in F\}]^2 \end{aligned}$$

Clearly, $B_n \cap B_m = \varnothing$ if $n \neq m$; hence, since by the preceding inequalities $\Pr\{B_n\}$ does not depend on n, it follows that $\Pr\{B_n\} = 0$ for all n. But then, again by the same inequalities, $\Pr\{X \in F\}$ can only be 0 or 1. ■

Corollary 4.1.7 Let X be a centred Gaussian process on T and $\|\cdot\|$ be a $\mathcal{C}$ -measurable pseudo-norm on $\mathbb{R}^T$. Then

$$P\{\|X\| < \infty\} = 0 \text{ or } 1$$

Proof. The set $\{x \in \mathbb{R}^T : \|x\| < \infty\} = \cup_n \{x \in \mathbb{R}^T : \|x\| < n\}$ is a measurable vector space, and the 0 - 1 law yields the result. ■

■ **Example 4.3** If X is Gaussian, separable and centred, then there exists $T_0 \subset T$, T_0 countable, such that $\sup_{t\in T}|X(t)| = \sup_{t\in T_0}|X(t)|$ a.s, but $\|x\|_{T_0} := \sup_{t\in T_0}|x(t)|$ is a measurable pseudo-norm, and hence it is finite with probability 0 or 1.

■ **Example 4.4** The B-valued Gaussian variables where B is a separable Banach space constitute a very general and important class of Gaussian processes, and we define them now. Given a separable Banach space B, a B-valued random variable X is centred Gaussian if $f(X)$ is a mean zero normal variable for every $f \in B^*$, the topological dual of B. By linearity, this is equivalent to the statement that $f_1(X),\ldots,f_n(X)$ are jointly centred normal for every $n \in \mathbb{N}$ and $f_i \in B^*$. In particular, if X is a B-valued centred Gaussian random variable, then the map $X : B^* \mapsto \mathcal{L}^2(\Omega,\Sigma,\Pr)$, defined by $X(f) = f(X)$, is a centred Gaussian process. If $B = E$ has dimension d, X is centred Gaussian iff the coordinates of X in a basis of E are jointly normal with mean zero (hence, the same is true for the coordinates of X in any basis).

Now we turn to a very useful property of Gaussian processes X, namely, that the supremum norm of a Gaussian process concentrates about its mean, as well as about its median, with very high probability, in fact as if it were a real normal variable with variance the largest variance of the individual variables $X(t)$. This result is a consequence of an even deeper result, the isoperimetric inequality for Gaussian measures, although it has simpler direct proofs, particularly if one is allowed some latitude and does not aim at the best result. Here is one such proof that uses the stability property in an elegant and simple way.

We should recall that a function $f : V \mapsto \mathbb{R}$, where V is a metric space, is lipschitz with Lipschitz constant $c = \|f\|_{\text{Lip}}$ if $c := \sup_{x\neq y}|f(x) - f(y)|/d(x,y) < \infty$. Rademacher's theorem asserts that if $f : \mathbb{R}^n \mapsto \mathbb{R}$ is Lipschitz, then it is a.e. differentiable and the essential supremum of the norm of its derivative is bounded by its Lipschitz constant $\|f\|_{\text{Lip.}}$. We remark that although we will use this result in the theorem that follows, it is not needed for its application to a concentration of maxima of jointly normal variables because one can compute by hand the derivative of the Lipschitz function $x \mapsto \max_{i\leq d}|x_i|, x \in \mathbb{R}^d$.

Theorem 4.1.8 Let $(B, \|\cdot\|_B)$ be a finite-dimensional Banach space, and let X be an B-valued centred Gaussian random variable. Let $f : B \mapsto \mathbb{R}$ be a Lipschitz function. Let $\Psi : \mathbb{R} \mapsto \mathbb{R}$ be a nonnegative, convex, measurable function. Then the following inequality holds:

$$E[\Psi(f(X) - Ef(X))] \leq E\left[\Psi\left(\frac{\pi}{2}\langle f'(X), Y\rangle\right)\right] \tag{4.24}$$

where Y is an independent copy of X(X and Y have the same probability law and are independent), and $\langle\cdot,\cdot\rangle$ denotes the duality action of B^* on B.

Proof. Since the range of X is a full subspace, we may assume without loss of generality that B equals the range of X (i.e., the support of the law of X is B). This has the effect that the law of X and Lebesgue measure on B are mutually absolutely continuous (as the density of X is strictly positive on its supporting subspace). For $\theta \in [0, 2\pi)$, define $X(\theta) = X \sin\theta + Y \cos\theta$. Then $X'(\theta) = X\cos\theta - Y\sin\theta$, and notice that $X(\theta)$ and $X'(\theta)$ are (normal and) independent: it suffices to check covariances, and if $f, g \in B^*$, we have

$$E\left[f(X(\theta))g\left(X'(\theta)\right)\right] = E(f(X)g(X))\sin\theta\cos\theta - E(f(Y)g(Y))\sin\theta\cos\theta = 0$$

In other words, the joint probability laws of X and Y and of $X(\theta)$ and $X'(\theta)$ coincide.

since for any increasing sequence θ_i

$$\begin{aligned}\sum |f(X(\theta_i)) - f(X(\theta_{i-1}))| &\leq \|f\|_{\text{Lip}} \sum \|X(\theta_i) - X(\theta_{i-1})\| \\ &\leq \|f\|_{\text{Lip}} (\|X\| + \|Y\|) \sum |\theta_i - \theta_{i-1}|\end{aligned}$$

it follows that the function $\theta \mapsto f(X(\theta))$ is absolutely continuous, and therefore, we have

$$f(X) - f(Y) = f(X(\pi/2)) - f(X(0)) = \int_0^{\pi/2} \frac{d}{d\theta} f(X(\theta)) d\theta$$

Using convexity of Ψ, Fubini's theorem and the preceding, we obtain

$$\begin{aligned}E\Psi(f(X) - Ef(X)) &= E\Psi(f(X) - Ef(Y)) \leq E\Psi(f(X) - f(Y)) \\ &= E\Psi\left(\int_0^{\pi/2} \frac{d}{d\theta} f(X(\theta)) d\theta\right) \leq \frac{2}{\pi} E \int_0^{\pi/2} \Psi\left(\frac{\pi}{2}\frac{d}{d\theta} f(X(\theta))\right) d\theta \\ &= \frac{2}{\pi} \int_0^{\pi/2} E\Psi\left(\frac{\pi}{2}\frac{d}{d\theta} f(X(\theta))\right) d\theta\end{aligned}$$

Now f is m a.e. differentiable with a bounded derivative by Rademacher's theorem, where m is Lebesgue measure on B, and since $\mathcal{L}(X(\theta))$ is absolutely continuous with respect to Lebesgue measure for every $\theta \in [0, \pi/2)(X(\theta)$ has the same support as $X), f'$ exists a.s. relative to the law of $X(\theta)$. since $X''(\theta)$ exists for each θ, it follows from the chain rule that given $\theta, df(X(\theta))/d\theta = \langle f'(X(\theta)), X'(\theta)\rangle$ a.s. Then, since $\mathcal{L}(X, Y) =$

$\mathcal{L}(X(\theta), X'(\theta))$, we have

$$E\Psi\left(\frac{\pi}{2}\frac{d}{d\theta}f(X(\theta))\right) = E\Psi\left(\frac{\pi}{2}\langle f'(X), Y\rangle\right)$$

which, combined with the preceding string of inequalities, proves the theorem. ■

R It turns out, as we will see in the next section, that Lipschitz functions are the natural tool for extracting concentration results from isoperimetric inequalities, on the one hand, and on the other, as we will see now, the supremum norm of a vector in $\mathbb{R}^n$ is a Lipschitz function, so concentration inequalities for Lipschitz functions include as particular cases concentration inequalities for the supremum norm and for other norms as well.

■ **Example 4.5 — Concentration for the maximum of a finite number of jointly normal variables.** To estimate the distribution of $\max_{i\le n}|g_i|$ for a finite sequence $g_1,\ldots,g_n$ of jointly normal variables using the preceding theorem, we take $B = \ell_\infty^n$, which is $\mathbb{R}^n$ with the norm $f(x) = \max_{i\le n}|x_i|$, where $x = (x_1,\ldots,x_n)$, which we take as our function f, and we take $X = (g_1,\ldots,g_n)$.f is obviously Lipschitz, so the previous theorem will apply to it. We also have that for each $1\le i\le n, f(x) = x_i$ on the set $\{x : x_i > |x_j|, 1\le j\le n, j\ne i\}$ and $f(x) = -x_i$ on $\{x : -x_i > |x_j|, 1\le j\le n, j\ne i\}$. It follows that m a.s. the gradient of f has all but one coordinate equal to zero, and this coordinate is 1 or -1. If $g_i \ne \pm g_j$ for $i\ne j$, which we can assume without loss of generality (by deleting repeated coordinates without changing the maximum), then this also holds a.s. for the law of X. Let $\sigma_i^2 = Eg_i^2$ and $\sigma^2 = \max_{i\le n}\sigma_i^2$. For almost every $X = x$ fixed, $\langle f'(x), Y\rangle$ is $\pm g_i$ for some i, that is, in law, the same as $\sigma_i g, g$ standard normal. Therefore, if we assume that the function Ψ is as in the preceding theorem and that, moreover, it is even and nondecreasing on $[0,\infty)$, then, letting E_Y denote integration with respect to the variable Y only, the preceding observation implies that, X a.s.,

$$E_Y\Psi\left(\frac{\pi}{2}\langle f'(x), Y\rangle\right) \le E\Psi\left(\frac{\pi}{2}\sigma g\right)$$

We conclude that for any $n\in\mathbb{N}$, if $g_1,\ldots,g_n$ are jointly normal random variables and if $\sigma^2 = \max_{i\le n} Eg_i^2$, then for any nonnegative, even, convex function Ψ nondecreasing on $[0,\infty)$

$$E\Psi\left(\max_{i\le n}|g_i| - E\max_{i\le n}|g_i|\right) \le E\Psi\left(\frac{\pi}{2}\sigma g\right) \tag{4.25}$$

where g denotes a standard normal random variable. Now $Ee^{t|g|} \le E\left(e^{tg} + e^{-tg}\right) = 2e^{t^2/2}$. Thus, if $\Psi_\lambda(x) = e^{\lambda|x|}$, we have

$$E\Psi_\lambda\left(\frac{\pi}{2}\sigma g\right) \le 2e^{\lambda^2\pi^2\sigma^2/8}$$

and, by (4.25) and Chebyshev's inequality,

$$\Pr\left\{\left|\max_{i\le n}\left|g_i\right|-E\max_{i\le n}\left|g_i\right|\right|>u\right\}\le 2e^{-\lambda u+\lambda^2\pi^2\sigma^2/8},\quad u\ge 0$$

With $\lambda u/2=\lambda^2\pi^2\sigma^2/8$, that is, $\lambda=4u/\left(\pi^2\sigma^2\right)$, this inequality gives the following approximate concentration inequality about its mean for the maximum of any finite number of normal random variables:

$$\Pr\left\{\left|\max_{i\le n}\left|g_i\right|-E\max_{i\le n}\left|g_i\right|\right|>u\right\}\le 2e^{-\frac{1}{\pi^2}\frac{u^2}{2\sigma^2}},\quad u\ge 0 \tag{4.26}$$

The last inequality and the one in the next theorem are suboptimal: the factor $1/\pi^2$ in the exponent is superfluous, as we will see in two of the sections that follow. We can translate (4.25) and (4.26) into a concentration inequality for the supremum norm of a separable Gaussian process (and draw as well some consequences).

Theorem 4.1.9 Let $\{X(t),t\in T\}$ be a separable centred Gaussian process such that

$$\Pr\left\{\sup_{t\in T}|X(t)|<\infty\right\}>0$$

Let Ψ be an even, convex, measurable function, nondecreasing on $[0,\infty)$. Let g be $N(0,1)$. Then a. $\sigma=\sigma(X):=\sup_{t\in T}\left(EX^2(t)\right)^{1/2}<\infty$ and $E\sup_{t\in T}|X(t)|<\infty$ and b. The following inequalities hold:

$$E\Psi\left(\sup_{t\in T}|X(t)|-E\sup_{t\in T}|X(t)|\right)\le E\Psi\left(\frac{\pi}{2}\sigma g\right)$$

and

$$\Pr\left\{\left|\sup_{t\in T}\left|X(t)\right|-E\sup_{t\in T}\left|X(t)\right|\right|>u\right\}\le 2e^{-\left(Ku^2/2\sigma^2\right)}$$

where $K=\frac{1}{\pi^2}$ (As mentioned earlier, the optimal constant K in this theorem will be shown to be 1.)

Proof. By assumption and the 0 - 1 law (Theorem 2.1.13; see the example following Corollary 4.1.7 , sup $_{t\in T}|X(t)|<\infty$ a.s. Let $0<z_{1/2}<1$ be such that $\Pr\{|g|>z_{1/2}\}=1/2$, and let $M<\infty$ be such that $\Pr\left\{\sup_{t\in T}|X(t)|>M\right\}<1/2$. Then, for each t

$$1/2>\Pr\{|X(t)|>M\}=\Pr\left\{|g|>M/\left(EX(t)^2\right)^{1/2}\right\}$$

which implies that $\sigma=\sup_{t\in T}\left(EX^2(t)\right)^{1/2}\le M/z_{1/2}<\infty$. Let $T_0=\{t_i\}_{i=1}^n$ be a countable set such that $\sup_{t\in T}|X(t)|=\sup_{t\in T_0}|X(t)|$. For every $n\in\mathbb{N}$, we have, by inequality (4.26),

$$\Pr\left\{\left|\max_{i\le n}\left|X\left(t_i\right)\right|-E\max_{i\le n}\left|X\left(t_i\right)\right|\right|>\sigma u\right\}\le 2e^{-u^2/2\pi^2}$$

since $\sup_{t\in T}|X(t)| < \infty$ a.s., this variable has a finite median m, and also for all n

$$\Pr\left\{\max_{i\leq n}|X(t_i)| \leq m\right\} \geq \frac{1}{2}$$

If u_0 is such that $2e^{-u_0^2/2\pi^2} < 1/2$, these two inequalities imply that for all $n \in \mathbb{N}$, the intersection of the two sets $\{x : |E\max_{i\leq n}|X(t_i)| - x| \leq \sigma u_0\}$ and $\{x : x \leq m\}$ is not empty and hence that $E\max_{i\leq n}|X(t_i)| \leq m + \sigma u_0 < \infty$. a) is proved. We have $\sup_{t\in T}|X(t)| = \lim_{n\to\infty}\max_{i\leq n}|X(t_i)|$ a.s. and, by monotone convergence, also in $L^1(\Pr)$. Hence, the first inequality in (b) follows by inequality (4.25), continuity of Ψ and Fatou's lemma. The second inequality follows from the first by Chebyshev's inequality in the same way as (4.26) follows from (4.25) . ■

4.1.2 Isoperimetric Inequalities with Applications to Concentration

The Gaussian isoperimetric inequality, in its simplest form, identifies the half-spaces as the sets of $\mathbb{R}^n$ with the smallest Gaussian perimeter among those with a fixed Gaussian measure, where the Gaussian measure in question is the standard one, that is, the probability law of n independent standard normal random variables, and where the Gaussian perimeter of a set is taken as the limit of the measure of the difference of an ε -enlargement of the set and the set itself divided by ε. The proof of this theorem was obtained originally by translating the isoperimetric inequality on the sphere to the Gaussian setting by means of Poincaré's lemma, which states that the limiting distribution of the orthogonal projection onto a Euclidean space of fixed dimension n of the uniform distribution on the sphere of $\mathbb{R}^{m+1}$ with radius $\sqrt{m}$ is the standard Gaussian measure of $\mathbb{R}^n$. The isoperimetric inequality on the sphere is a deep result that goes back to P. Lévy and E. Schmidt, ca. 1950 (although the equivalent isoperimetric problem on the plane goes back to the Greeks-recall, for instance, 'Dido's problem'). The Gaussian isoperimetric inequality does imply best possible concentration inequalities for Lipschitz functions on $\mathbb{R}^n$ and for functions on $\mathbb{R}^N$ that are Lipschitz 'in the direction of ℓ_2 ', although concentration inequalities have easier proofs, as seen in the preceding section and as will be seen again in further sections. The Gaussian isoperimetric inequality in general Banach spaces requires the notion of reproducing kernel Hilbert space and will be developed in a further section as well. This section contains proofs as short as we could find of the isoperimetric inequalities on the sphere and for the standard Gaussian measure on $\mathbb{R}^n, n \leq \infty$, with applications to obtain the best possible concentration inequality with respect to the standard Gaussian measure for Lipschitz functions f about their medians and for the supremum norm of a separable Gaussian process X when $\sup_{t\in T}|X(t)| < \infty$ a.s.

The Isoperimetric Inequality on the Sphere

Let $S^n = \left\{x \in \mathbb{R}^{n+1} : \|x\|^2 = \sum_{i=1}^{n+1} x_i^2 = 1\right\}$, where $x = (x_1, \ldots, x_{n+1})$; let p be an arbitrary point in S^n that we take to be the north pole, $p = (0, \ldots, 0, 1)$; and let μ be the uniform probability distribution on S^n (equal to the normalized volume element - surface area for S^2 equal also to the normalized Haar measure of the rotation group). Let d be the geodesic distance on S^n, defined, for any two points, as the length of the shortest segment of the great circle joining them.

A closed cap centred at a point $x \in S^n$ is a geodesic closed ball around x, that is, a set of the form $C(x,\rho) := \{y : d(x,y) \leq \rho\}$. Here ρ is the radius of the cap, and clearly, the μ -measure

of a cap is a continuous function of its radius, varying between 0 and 1. Often we will not specify the centre or the radius of $C = C(x,\rho)$, particularly if the centre is the north pole. The isoperimetric inequality on the sphere states that the caps are the sets of shortest perimeter among all the measurable sets of a given surface area. What we will need is an equivalent formulation, in terms of neighbourhoods of sets, defined as follows: the closed ε neighbourhood of a set A is defined as $A_\varepsilon = \{x : d(x,A) \leq \varepsilon\}$, with the distance between a point and a set being defined, as usual, by $d(x,A) = \inf\{d(x,y) : y \in A\}$. The question is: among all measurable subsets of the sphere with surface area equal to the surface area of A find sets B for which the surface areas of their neighbourhoods $B_\varepsilon, 0 < \varepsilon < 1$, are smallest. The following theorem shows that an answer are the caps (they are in fact the answer, but uniqueness will not be considered: we are only interested in the value of inf $\mu(A_\varepsilon), \varepsilon > 0$).

Theorem 4.1.10 Let $A \neq \emptyset$ be a measurable subset of S^n, and let C be a cap such that t $\mu(C) = \mu(A)$. Then, for all $\varepsilon > 0$

$$\mu(C_\varepsilon) \leq \mu(A_\varepsilon) \tag{4.27}$$

The proof is relatively long, and some prior digression may help. The idea is to construct transformations $A \mapsto A^*$ on measurable subsets of the sphere that preserve area, that is, $\mu(A) = \mu(A^*)$, and decrease perimeter, a condition implied by $\mu\left((A^*)_\varepsilon\right) \leq \mu(A_\varepsilon) = \mu\left((A_\varepsilon)^*\right)$ $\varepsilon > 0$, because the perimeter of A is the limit as $\varepsilon \to 0$ of $\mu(A_\varepsilon \backslash A)/\varepsilon$. Then iterating transformations that satisfy these two properties should eventually produce the solution, in our case a cap. Or, more directly, one may obtain a cap using a more synthetic compactness argument instead of iteration. In the sense that A^* concentrates the same area as A on a smaller perimeter, A^* is closer to the solution of the problem than A is. A^* is called a symmetrisation of A.

Proof. Proof If $\mu(A) = 0$, then C consists of a single point, and (2.4) holds. Next, we observe that by regularity of the measure μ, it suffices to prove the theorem for A compact. By regularity, there exist A^m compact, $A^m \subset A, A^m$ increasing and such that $\mu(A^m) \nearrow \mu(A)$. Let C^m be caps with the same centre as C and with $\mu(C^m) = \mu(A^m)$. since the measure of a cap is a continuous one-to-one function of its geodesic radius, we also have

$\mu\left(C_{\varepsilon}^{m}\right) \nearrow \mu\left(C_{\varepsilon}\right)$, and if the theorem holds for compact sets, then

$$\mu\left(A_{\varepsilon}\right) \geq \lim \mu\left(A_{\varepsilon}^{m}\right) \geq \lim \mu\left(C_{\varepsilon}^{m}\right)=\mu\left(C_{\varepsilon}\right)$$

and the theorem holds in general. Thus, we will assume that A is compact and that $\mu(A) \neq 0$. We divide the proof into several parts.

Part 1: Construction and main properties of the symmetrisation operation. Given an n -dimensional subspace $H \subset \mathbb{R}^{n+1}$ that does not contain the point p, let $\sigma=\sigma_{H}$ be the reflection about H; that is, if $x=u+v$ with $u \in H$ and v orthogonal to H, then $\sigma(x)=u-v$. Clearly, σ is an isometry (so it preserves μ -measure), and it is involutive; that is, $\sigma^{2}=\sigma$. It also satisfies a property that, together with the preceding two, is crucial for the symmetrisation operation to work, namely, that if x and y are on the same half-space with respect to H, then

$$d(x,y) \leq d(x,\sigma(y)) \tag{4.28}$$

To see this, observe that the geodesic distance is an increasing function of the Euclidean distance, so it suffices to prove (2.5) for the Euclidean distance. Changing orthogonal coordinates if necessary, we may and do assume that $H=\left\{x: x_{n+1}=0\right\}$, so if x and y are in the same hemisphere, then $\operatorname{sign}\left(x_{n+1}\right)=\operatorname{sign}\left(y_{n+1}\right)$, which implies that the $(n+1)$ th coordinate of $x-y$ is dominated in absolute value by the $(n+1)$ th coordinate of $x-\sigma(y)$, whereas the first n coordinates of these two vectors coincide. Hence, $\sum_{i=1}^{n+1}\left(x_{i}-y_{i}\right)^{2} \leq \sum_{i=1}^{n+1}\left(x_{i}-\sigma(y)_{i}\right)^{2}$ H divides S^{n} into two open hemispheres, and we denote by S_{+} the open hemisphere that contains p, S_{-} the other hemisphere, and $S_{0}=S^{n} \cap H$. The symmetrisation of A with respect to $\sigma=\sigma_{H}, s_{H}(A)=A^{*}$ is defined as

$$s_{H}(A)=A^{*}:=\left[A \cap\left(S_{+} \cup S_{0}\right)\right] \cup\left\{a \in A \cap S_{-}: \sigma(a) \in A\right\} \cup\left\{\sigma(a): a \in A \cap S_{-}, \sigma(a) \notin A\right\} \tag{4.29}$$

Note that A^{*} is obtained from A by reflecting towards the northern hemisphere every $a \in A \cap S_{-}$ for which $\sigma(A)$ is not already in A. It is easy to see (Exercise 2.2.1) that if A is compact, then so is A^{*} and that if C is a cap with centre at p or at any other point in the northern hemisphere, then $C^{*}=C$. Next, observe that the three sets in the definition are disjoint and that, σ being an isometry, the measure of the third set equals $\mu\left\{a \in A \cap S_{-}: \sigma(a) \notin A\right\}$, which implies that

$$\mu\left(A^{*}\right)=\mu(A), \quad A \in \mathcal{B}\left(S^{n+1}\right) \tag{4.30}$$

This is one of the two properties of the symmetrisation operation that we need. We now show that the ε -neighbourhoods of A^{*} are less massive than those of A (thus making A^{*} 'closer' to being a cap than A is), actually, we prove more, namely, that for all $A \in \mathcal{B}\left(S^{n}\right)$ and $\varepsilon>0$, then

$$\left(A^{*}\right)_{\varepsilon} \subseteq\left(A_{\varepsilon}\right)^{*}, \quad \text{hence} \quad \mu\left(\left(A^{*}\right)_{\varepsilon}\right) \leq \mu\left(\left(A_{\varepsilon}\right)^{*}\right)=\mu\left(A_{\varepsilon}\right) \tag{4.31}$$

To see this, let $x \in (A^*)_\varepsilon$ and let $y \in A^*$ be such that $d(x,y) \leq \varepsilon$ (such a $y \in A^*$ exists by compactness). Then, using (2.5) and that σ is an involutive isometry, we obtain, when x and y lay on different half-spaces,

$$d(\sigma(x),y) = d(x,\sigma(y)) \leq d(\sigma(x),\sigma(y)) = d(x,y) \leq \varepsilon$$

Thus, since $y \in A^*$ implies that either $y \in A$ or $\sigma(y) \in A$, in either case we have that both $x \in A_\varepsilon$ and $\sigma(x) \in A_\varepsilon$; hence, $x \in (A_\varepsilon)^*$. If x and y are in S_-, then y and $\sigma(y)$ are both in A, and therefore, by the last identity earlier, $x \in A_\varepsilon$ and $\sigma(x) \in A_\varepsilon$; hence, $x \in (A_\varepsilon)^*$ in this case as well. If x and y are in S_+, then either y or $\sigma(y)$ is in A; hence, either x or $\sigma(x)$ is in A_ε, which together with $x \in S_+$ implies that $x \in (A_\varepsilon)^*$. The cases where x and/or y are in S_0 are similar, even easier, and they are omitted. The inclusion in (4.31) is proved, and the inequality there follows from the inclusion and from (4.30).

Part 2: Preparation for the compactness argument. Let $(\mathcal{K},h)$ denote the set of nonempty compact subsets of S^n equipped with the Hausdorff distance, defined as $h(A,B) = \inf\{\varepsilon : A \subseteq B_\varepsilon, B \subseteq A_\varepsilon\}, A,B \in \mathcal{K}.(\mathcal{K},h)$ is a compact metric space (Exercise 2.2.2). Given a compact nonempty set $A \subseteq S^n$, let $\mathcal{A}$ be the minimal closed subset of $\mathcal{K}$ that contains A and is preserved by s_H for all n -dimensional subspaces H of $\mathbb{R}^{n+1}$ that do not contain the north pole p (meaning that if $A \in \mathcal{K}$, then $s_H(A) \in \mathcal{K}$ for all H with $p \notin H$). $\mathcal{A}$ exists and is nonempty because $\mathcal{K}$ is a closed $\{s_H\}$ -invariant collection of sets that contains A. Also note that since $(\mathcal{K},h)$ is compact and $\mathcal{A}$ closed, $\mathcal{A}$ is compact. We have Claim: If $B \in \mathcal{A}$, then (a) $\mu(B) = \mu(A)$, and (b) for all $\varepsilon > 0, \mu(B_\varepsilon) \leq \mu(A_\varepsilon)$ Proof of the claim. It suffices to show that the collection of closed sets $\mathcal{F}$ satisfying a) and b) is preserved by s_H for all H not containing p and is a closed subset of $\mathcal{K}$ because then $\mathcal{A} \subseteq \mathcal{F}$ follows by minimality of $\mathcal{A}$. That $s_H(\mathcal{F}) \subseteq \mathcal{F}$ follows from (2.7) and (2.8). Let now $B^n \in \mathcal{F}$ and $h(B^n,B) \to 0$. Let $\varepsilon > 0$ be fixed. Given $\delta > 0$, there exists n_δ such that $B \subseteq B^n_\delta$ for all $n \geq n_\delta$; hence, $B_\varepsilon \subseteq B^n_{\delta+\varepsilon}$ and $\mu(B_\varepsilon) \leq \mu\left(B^n_{\delta+\varepsilon}\right) \leq \mu(A_{\delta+\varepsilon})$. Letting $\delta \searrow 0$ shows that B satisfies condition (b). Letting $\varepsilon \searrow 0$ in condition (b) for B shows that $\mu(B) \leq \mu(A)$. Using that for all n large enough we also have $B^n \subseteq B_\delta$, we get that $\mu(A) = \mu(B^n) \leq \mu(B_\delta)$ and, letting $\delta \searrow 0$, that $\mu(A) \leq \mu(B)$, proving condition (a). The claim is proved. Part 3: Completion of the proof of Theorem 2.2.1. Clearly, because of the claim about $\mathcal{A}$, it suffices to show that if C is the cap centred at p such that $\mu(A) = \mu(C)$, then $C \in \mathcal{A}$.

Define $f(B) = \mu(B \cap C), B \in \mathcal{A}$. We show first that f is upper semicontinuous on $\mathcal{A}$. If $h(B^n,B) \to 0$, then, given $\delta > 0$, for all n large enough, $B^n \subseteq B_\delta$, which, as is easy to see, implies that $B^n \cap C \subseteq (B \cap C_\delta)_\delta$. Hence, $\limsup_n \mu(B^n \cap C) \leq \mu((B \cap C_\delta)_\delta)$, but because B and C are closed, if $\delta_n \searrow 0$, then $\cap_n (B \cap C_{\delta_n})_{\delta_n} = B \cap C$, thus obtaining $\limsup_n \mu(B^n \cap C) \leq \mu(B \cap C)$ since f is upper semicontinuous on $\mathcal{A}$ and $\mathcal{A}$ is compact, f attains its maximum at some $B \in \mathcal{A}$. The theorem will be proved if we show that $C \subseteq B$. Assume that $C \notin B$. Then, since $\mu(C) = \mu(A) = \mu(B)$ and both C and B are closed, we have that both $B\backslash C$ and $C\backslash B$ have positive μ -measure. Thus, the Lebesgue density theorem, which holds on S^n (see Exercise 2.2 .3 for definitions and a sketch of the proof), implies that there exist points of density $x \in B\backslash C$ and $y \in C\backslash B$. Let H be the subspace of dimension n orthogonal to the vector $x-y$, and let us keep the shorthand notation σ for the reflection with respect to H D^* for $s_H(D), S_+, S_-$ for the two hemispheres determined by H, and S_0 for $S^n \cap H$. Then $\sigma(y)=x$. since $y \in C$ and $x \notin C$, we have both, that p is not in H (the reflection of a point in C with respect to a hyperplane through p is necessarily in C) and that y is closer to p than x is; that is, $d(y,p) \leq d(x,p) = d\left(\sigma_H(y),p\right)$. Then it follows from this last obsesrvation and (4.28) that $y \in S_+$ and $x \in S_-$. Let $x \in (B \cap C)^*$. Then, if $x \in B \cap C \cap (S_+ \cup S_0)$ or if $x \in B \cap C \cap S_-$ and $\sigma(x) \in B \cap C$, we obviously have $x \in B^* \cap C$. Now, if $z \in C \cap S_-$, then $\sigma(z) \in C($ as $\sigma(z)$ is closer to p than z is); hence, if $x=\sigma(z)$ with $z \in B \cap C \cap S_-$ and $\sigma(z) \notin B \cap C$, then $\sigma(z)$ is not in B and therefore $x \in B^* \cap C$. We conclude that $(B \cap C)^* \subseteq B^* \cap C$ and, in particular, that

$$\mu(B \cap C) = \mu\left((B \cap C)^*\right) \leq \mu\left(B^* \cap C\right)$$

By definition of density point, for $\delta > 0$ small enough, $C(x,\delta) \subset S_-, \sigma(C(x,\delta)) = C(y,\delta) \subset S_+, \mu((B\backslash C) \cap C(x,\delta)) \geq 2\mu(C(x,\delta))/3$, and $\mu((C\backslash B) \cap C(y,\delta)) \geq 2\mu(C(y,\delta))/3$. Then the set

$$D = ((B\backslash C) \cap C(x,\delta)) \cap \sigma((C\backslash B) \cap C(y,\delta)) \tag{4.32}$$

satisfies

$$\mu(D) \geq \mu(C(x,\delta))/3 > 0, \quad D \subset (B\backslash C) \cap S_- \quad \text{and} \quad \sigma(D) \subset C\backslash B \tag{4.33}$$

The inclusions in (2.10) imply that $\sigma(D) \subset B^* \cap C$ and $\sigma(D) \cap (B \cap C)^* = \varnothing($ as $z \in (B \cap C)^*$ implies either $z \in B \cap C$ or $\sigma(z) \in B \cap C$). This together with (2.9) and $\mu(D) > 0$ proves

$$\mu\left(B^* \cap C\right) \geq \mu\left((B \cap C)^* \cup \sigma(D)\right) = \mu\left((B \cap C)^*\right) + \mu(D) > \mu\left((B \cap C)^*\right)$$

which, because $B^* \in \mathcal{A}$, contradicts the fact that f attains it maximum at B.

■

The Gaussian Isoperimetric Inequality for the Standard Gaussian Measure on $\mathbb{R}^N$

In this subsection we translate the isoperimetric inequality on the sphere to an isoperimetric inequality for the probability law γ_n of n independent $N(0,1)$ random variables by means of Poincaré's lemma, which states that this measure can be obtained as the limit of the projection of the uniform distribution on $\sqrt{m} S^{n+m}$ onto $\mathbb{R}^n$ when $m \rightarrow \infty$. We also let $n \rightarrow \infty$.

In what follows, $g_i, i \in \mathbb{N}$, is a sequence of independent $N(0,1)$ random variables, and as mentioned earlier, $\gamma_n = \mathcal{L}\left(g_1, \ldots, g_n\right)$. We call γ_n the standard Gaussian measure on $\mathbb{R}^n$. We also set $\gamma = \mathcal{L}\left(\{g_i\}_{i=1}^{\infty}\right)$, the law of the process $i \mapsto g_i, i \in \mathbb{N}$, a probability measure on the cylindrical σ -algebra $\mathcal{C}$ of $\mathbb{R}^{\mathrm{N}}$, which we also refer to as the standard Gaussian measure on $\mathbb{R}^N$.

Here is the Gaussian isoperimetric problem: for a measurable subset A of $\mathbb{R}^n$, and $\varepsilon > 0$, define its Euclidean neighbourhoods A_ε as $A_\varepsilon := \{x \in \mathbb{R}^n : d(x,A) \leq \varepsilon\} = A + \varepsilon O_n$, where d denotes Euclidean distance and O_n is the closed d -unit ball centred at $0 \in \mathbb{R}^n$. The problem is this: given a Borel set A, find among the Borel sets $B \subset \mathbb{R}^n$ with the same γ_n -measure as A those for which the γ_n -mesure of the neighbourhood B_ε is smallest, for all $0 < \varepsilon < 1/2$. The solution will be shown to be the affine half-space $(\{x : \langle x,u\rangle \leq \lambda\}, u$ any unit vector, $\lambda \in \mathbb{R})$ of the same measure as A. Note that $\gamma_n\{x : \langle x,u\rangle \leq \lambda\} = \gamma_1\{x \leq \lambda\}$.

Prior to stating and proving the main results, we describe the relationship between the uniform distribution on the sphere of increasing radius and dimension and the standard Gaussian measure on $\mathbb{R}^n$.

Lemma 4.1 — Poincaré's lemma. Let μ_{n+m} be the uniform distribution on $\sqrt{m}S^{n+m}$, the sphere of $\mathbb{R}^{n+m+1}$ of radius $\sqrt{m}$ and centred at the origin. Let π_m be the orthogonal projection $\mathbb{R}^{n+m+1} \mapsto \mathbb{R}^n = \left\{x \in \mathbb{R}^{n+m+1} : x_i = 0, n < i \leq n+m+1\right\}$, and let $\tilde{\pi}_m$ be the restriction of π_m to $\sqrt{m}S^{n+m}$. Let $v_m = \mu_{n+m} \circ \tilde{\pi}_m^{-1}$ be the projection onto $\mathbb{R}^n$ of μ_{n+m}. Then v_m has a density f_m such that if ϕ_n is the density of $\gamma_n, \lim_{m\to\infty} f_m(x) = \phi_n(x)$ for all $x \in \mathbb{R}^n$ Therefore,

$$\gamma_n(A) = \lim_{m\to\infty} \mu_{n+m}\left(\tilde{\pi}_m^{-1}(A)\right)$$

for all Borel sets A of $\mathbb{R}^n$.

Proof. Set $G_n := (g_1, \ldots, g_n)$ and $G_{n+m+1} := (g_1, \ldots, g_{n+m+1})$. The rotational invariance of the standard Gaussian law on Euclidean space implies that μ_{n+m} is the law of the vector $\sqrt{m}G_{n+m+1}/\left|G_{n+m+1}\right|^{1/2}$. Hence, v_m is the law of $\sqrt{m}G_n/\left|G_{n+m+1}\right|^{1/2}$. This allows for computations with normal densities that we only sketch. For any measurable set A of $\mathbb{R}^n$

$$v_m(A) = \frac{1}{(2\pi)^{(n+m+1)/2}} \int_{\mathbb{R}^{m+1}} \int_{\tilde{A}(y)} e^{-\left(|z|^2+|y|^2\right)/2} dz dy$$

where $z \in \mathbb{R}^n$ and $y \in \mathbb{R}^{m+1}$, and $\tilde{A} = \left\{z \in \mathbb{R}^n : \sqrt{m/\left(|z|^2+|y|^2\right)} z \in A\right\}$. Make the change of variables $z \mapsto x, x = \sqrt{m/\left(|z|^2+|y|^2\right)} z$ or $z = |y|x/\sqrt{m-|x|^2}, |x| \leq \sqrt{m}$. Its Jacobian is $\partial(z)/\partial(x) = m|y|^n/\left(m-|x|^2\right)^{1+n/2}$, thus obtaining

$$\begin{aligned} v_m(A) &= \frac{1}{(2\pi)^{(n+m+1)/2}} \int_A I\left(|x|^2 < m\right) \frac{m}{\left(m-|x|^2\right)^{n/2+1}} \int_{\mathbb{R}^{m+1}} |y|^n \exp\left(-\frac{1}{2}\frac{m|y|^2}{m-|x|^2}\right) dy dx \\ &= \frac{E\left(\left|G_{m+1}\right|^n\right)}{m^{n/2}} \frac{1}{(2\pi)^{n/2}} \int_A \left(1-\frac{|x|^2}{m}\right)^{(m-1)/2} I\left(|x|^2 < m\right) dx \end{aligned}$$

Hence, the density of v_m is $f_m(x) = C_{n,m}(2\pi)^{-n/2}\left(1-|x|^2/m\right)^{(m-1)/2} I\left(|x|^2 < m\right), x \in \mathbb{R}^n$.

Clearly, $(2\pi)^{-n/2}\left(1-|x|^2/m\right)^{(m-1)/2} I\left(|x|^2 < m\right) \to (2\pi)^{-n/2}e^{-|x|^2/2}$ for all x as $m \to \infty$.

Moreover, since for $0 \le a < m$ and $m \ge 2$ we have $1 - a/m \le e^{-a/2(m-1)}$, it follows that $\left(1-|x|^2/m\right)^{(m-1)/2} I\left(|x|^2 < m\right)$ is dominated by the integrable function $e^{-|x|^2/4}$. Thus, by the dominated convergence theorem, $f_m(x)/C_{n,m} \to (2\pi)^{-n/2}e^{-|x|^2/2}$ in L^1, which implies that $C_{n,m}^{-1} \to 1$, proving the lemma. (Alternatively, just show that $C_{n,m} = E\left(|G_{m+1}|^n\right)/m^{n/2} \to 1$ as $m \to \infty$ by taking limits on well-known expressions for the moments of chi-square random variables.) Now the limit (2.11) for any Borel set follows by dominated convergence. ■

Theorem 4.1.11 For $n < \infty$, let γ_n be the standard Gaussian measure of $\mathbb{R}^n$, let A be a measurable subset of $\mathbb{R}^n$, and let H be a half-space $H = \{x \in \mathbb{R}^n : \langle x, u\rangle \le a\}$, u a unit vector, such that $\gamma_n(H) = \gamma_n(A)$ and hence with $a := \Phi^{-1}(\gamma_n(A))$, where Φ denotes the standard normal distribution function. Then, for all $\varepsilon > 0$,

$$\gamma_n(H + \varepsilon O_n) \le \gamma_n(A + \varepsilon O_n) \tag{4.34}$$

which, by the definition of a, is equivalent to

$$\gamma_n(A + \varepsilon O_n) \ge \Phi\left(\Phi^{-1}(\gamma_n(A)) + \varepsilon\right) \tag{4.35}$$

Proof. First, we check the behaviour of distances under $\tilde{\pi}_m$. If d_{n+m} denotes the geodesic distance of $\sqrt{m}S^{n+m}$, it is clear that the projection $\tilde{\pi}_m$ is a contraction from the sphere onto $\mathbb{R}^n$; that is, $|\tilde{\pi}_m(x) - \tilde{\pi}_m(y)| \le d_{n+m}(x,y)$ for any $x, y \in \sqrt{m}S^{n+m}$. Moreover, if in the half-space $H_b := \{x \in \mathbb{R}^n : \langle x, u\rangle \le b\}$, we have $-\sqrt{m} < b < \sqrt{m}$; then its pre-image $\tilde{\pi}^{-1}(H_b)$ is a nonempty cap, and for $0 < \varepsilon < \sqrt{m} - b$, we have $\left(\tilde{\pi}^{-1}(H_b)\right)_\varepsilon = \tilde{\pi}^{-1}(H_b + \tau(b,\varepsilon)O_n) = \tilde{\pi}^{-1}\left(H_{b+\tau(b,s)}\right)$, where

$$b + \tau = \sqrt{m}\cos\left(\cos^{-1}\frac{b}{\sqrt{m}} \pm \frac{\varepsilon}{\sqrt{m}}\right)$$

which, taking limits in the addition formula for the cosine, immediately gives $\lim_{m\to\infty}\tau(b,\varepsilon) = \varepsilon$

Let now $b < a = \Phi^{-1}(\gamma_n(A))$ so that $H_b = \{x : \langle x, u\rangle \le b\} \subset H$. Then, by Poincaré's lemma,

$$\lim_m \mu_{n+m}\left(\tilde{\pi}_m^{-1}(A)\right) = \gamma_n(A) > \gamma_n(H_b) = \lim_m \mu_{n+m}\left(\tilde{\pi}_m^{-1}(H_b)\right)$$

so for all m large enough, we have both $b \in (-\sqrt{m}, \sqrt{m})$, such that $\pi_m^{-1}(H_b)$ is a nonempty cap in the sphere, and $\mu_{n+m}\left(\tilde{\pi}_m^{-1}(A)\right) \ge \mu_{n+m}\left(\tilde{\pi}_m^{-1}(H_b)\right)$. Then the isoperimetric inequality for μ_{n+m} (Theorem 2.2.1) yields that for each $\varepsilon > 0, b + \varepsilon < \sqrt{m}$, for all m large enough,

$$\mu_{n+m}\left(\left(\tilde{\pi}_m^{-1}(A)\right)_{\varepsilon}\right) \geq \mu_{n+m}\left(\left(\pi_m^{-1}\left(H_b\right)\right)_{\varepsilon}\right) = \mu_{n+m}\left(\pi_m^{-1}\left(H_{b+\tau(b,\varepsilon)}\right)\right)$$

so by Poincaré's lemma again,

$$\gamma_n\left(A + \varepsilon O_n\right) \geq \limsup_m \mu_{n+m}\left(\left(\tilde{\pi}_m^{-1}(A)\right)_{\varepsilon}\right) \geq \limsup_m \mu_{n+m}\left(\left(\tilde{\pi}_m^{-1}\left(H_{b+\tau(b,s)}\right)\right)\right) = \gamma_n\left(H_{b+\varepsilon}\right)$$

since this holds for all $b < a$, it also holds with b replaced by a. ■

Theorem 4.1.11 extends to infinite dimensions, as will be shown in Theorem **??**. An extension to the standard Gaussian measure on $\mathbb{R}^{\mathbb{N}}$, that is, for the law γ of a sequence of independent standard normal random variables, can be obtained directly. Before stating the theorem, it is convenient to make some topological and measure-theoretic considerations. The distance $\rho(x,y) = \sum_{k=1}^{\infty} \min\left(\left|x_k - y_k\right|, 1\right) / 2^k$ metrises the product topology of $\mathbb{R}^{\mathbb{N}}$, and $\left(\mathbb{R}^{\mathbb{N}}, \rho\right)$ is a separable and complete metric space, as is easy to see. That is, $\mathbb{R}^{\mathbb{N}}$ is a Polish space (a topological space that admits a metric for which it is separable and complete). Then the cylindrical σ -algebra $\mathcal{C}$ coincides with the Borel σ -algebra of $\mathbb{R}^{\mathbb{N}}$, and any finite cylindrical (hence Borel) measure is tight (Radon). The product space $\mathbb{R}^{\mathbb{N}} \times \ell_2$ is also Polish, and for each $t \in \mathbb{R}$, the map $f_t : \mathbb{R}^{\mathbb{N}} \times \ell_2 \mapsto \mathbb{R}^{\mathbb{N}}, f_t(x,y) = x + ty$ is continuous. Then the image of f_t is universally measurable, that is, measurable for any Radon measure, in particular, in our case, measurable for any finite measure on the cylindrical σ -algebra $\mathcal{C}$ of $\mathbb{R}^{\mathbb{N}}$. See, for example, theorem 13.2 .6 in section 13.2 in Dudley (2002) .

Theorem 2.2 .4 Let A be a Borel set of $\mathbb{R}^{\mathbb{N}}$ (i.e., $A \in \mathcal{C}$), and let γ be the probability law of $\left(g_i : i \in \mathbb{N}\right), g_i$ independent standard normal. Let O denote the unit ball about zero of $\ell_2 \subset \mathbb{R}^{\mathbb{N}}, O = \left\{x \in \mathbb{R}^{\mathbb{N}} : \sum_i x_i^2 \leq 1\right\}$. Then, for all $\varepsilon > 0$

$$\gamma(A + \varepsilon O) \geq \Phi\left(\Phi^{-1}(\gamma(A)) + \varepsilon\right)$$

The proof is indicated in Exercises 2.2 .5 through 2.2 .7

Application to Gaussian Concentration

We would like to translate the isoperimetric inequality in Theorem **??** into a concentration inequality for functions of $\{g_i\}_{i=1}^n$ about their medians, that is, into a bound for $\gamma\{| f(x) - M | > \varepsilon\}$ for all $\varepsilon > 0$. The following definition describes the functions for which such a translation is almost obvious.

Definition 4.1.5 A function $f : \mathbb{R}^{\mathbb{N}} \mapsto \mathbb{R}$ is Lipschitz in the direction of ℓ_2, or ℓ_2 -Lipschitz for short, if it is measurable and if

$$\|f\|_{\text{Lip2}} := \sup\left\{\frac{|f(x)-f(y)|}{|x-y|} : x,y \in \mathbb{R}^{\mathrm{N}}, x \neq y, x-y \in \ell_2\right\} < \infty$$

where $|x-y|$ is the ℓ_2 norm of $x-y$.

For a measurable function f on $\mathbb{R}^{\mathbb{N}}$, we denote by M_f the median of f with respect to the Gaussian measure γ, defined as $M_f = \inf\{t : \gamma\{x : f(x) \leq t\} > 1/2\}$. Then $\gamma\left(f \leq M_f\right) \geq 1/2$ and $\gamma\left(f \geq M_f\right) \geq 1/2$, and M is the largest number satisfying these two inequalities.

Theorem 4.1.12 If f is an ℓ_2 -Lipschitz function on $\mathbb{R}^N$, and if M_f is its median with respect to γ, then

$$\begin{aligned}
&\gamma\left\{x : f(x) \geq M_f + \varepsilon\right\} \leq \left(1 - \Phi\left(\varepsilon / \|f\|_{\text{Lip2}}\right)\right) \\
&\gamma\left\{x : f(x) \leq M_f - \varepsilon\right\} \leq \left(1 - \Phi\left(\varepsilon / \|f\|_{\text{Lip2}}\right)\right)
\end{aligned}$$

in particular

$$\gamma\left\{x : \left|f(x) - M_f\right| \geq \varepsilon\right\} \leq 2\left(1 - \Phi\left(\varepsilon / \|f\|_{\text{Lip2}}\right)\right) \leq e^{-\varepsilon^2 / 2\|f\|\text{Lip2}}$$

for all $\varepsilon > 0$

Proof. Let $A^+ = \left\{x \in \mathbb{R}^{\mathbb{N}} : f(x) \geq M_f\right\}$ and $A^- = \left\{x \in \mathbb{R}^{\mathbb{N}} : f(x) \leq M_f\right\}$. Then $\gamma\left(A^+\right) \geq 1/2, \gamma\left(A^-\right) \geq 1/2$. Moreover, if $x \in A^+ + \varepsilon O$, then there exists $h \in O$ such that $x - \varepsilon h \in A^+$; hence, $f(x - \varepsilon h) \geq M_f$ and $f(x) + \varepsilon\|f\|_{\text{Lip 2}} \geq f(x - \varepsilon h) \geq M_f$; that is, $A^+ + \varepsilon O \subset \left\{x : f(x) \geq M_f - \varepsilon\|f\|_{\text{Lip 2}}\right\}$. Then the Gaussian isoperimetric inequality (2.14) for $A = A^+$ gives (recall $\Phi^{-1}(1/2) = 0$)

$$\gamma\left\{f < M_f - \varepsilon\|f\|_{\text{Lip 2}}\right\} \leq 1 - \gamma\left(A^+ + \varepsilon O\right) \leq 1 - \Phi(\varepsilon)$$

which is the second inequality in (2.15). Likewise, $A^- + \varepsilon O \subset \left\{x : f(x) \leq M_f + \varepsilon\|f\|_{\text{Lip 2}}\right\}$, and the isoperimetric inequality applied to A^+ gives the first inequality in (2.15). Finally, (2.16) follows by combination of the previous two inequalities and a known bound for the tail probabilities of a normal variable (Exercise 2.2 .8). ■

Let now $X(t), t \in T$, be a separable centred Gaussian process such that $\Pr\left\{\sup_{t \in T}|X(t)| < \infty\right\} > 0$. Then $\sup_{t \in T}|X(t)| = \sup_{t \in T_0}|X(t)| < \infty$ a.s., where $T_0 = \{t_k\}_{k=1}^{\infty}$ is a countable subset of T (see Example 2.1 .15). Ortho-normalizing (in $L^2(\Pr)$), the jointly normal sequence $\left\{X\left(t_k\right)\right\}$ yields $X\left(t_k\right) = \sum_{i=1}^k a_{ki} g_i$, where g_i are independent standard normal variables, and $\sum_{i=1}^k a_{ki}^2 = EX^2\left(t_k\right)$. Then the probability law of the process $X\left(t_k\right), k \in \mathbb{N}$, coincides with the law of the random variable defined on the probability space $\left(\mathbb{R}^{\mathbb{N}}, \mathcal{C}, \gamma\right), X : \mathbb{R}^{\mathbb{N}} \mapsto \mathbb{R}$ $\tilde{X}\left(t_k, x\right) = \sum_{i=1}^k a_{ki} x_i$. This is so because the coordinates of $\mathbb{R}^{\mathbb{N}}$, considered as random variables on the probability space $\left(\mathbb{R}^{\mathbb{N}}, \mathcal{C}, \gamma\right)$, are i.i.d. $N(0,1)$. Now define a function $f : \mathbb{R}^{\mathbb{N}} \mapsto \mathbb{R}$ by

$$f(x) = \sup_k \left| \sum_{i=1}^{k} a_{ki} x_i \right|$$

The probability law of f under γ is the same as the law of sup $_{t \in T_0} |X(t)|$, which, in turn, is the same as the law of sup $_{t \in T} |X(t)|$. Moreover, if $h \in O$, the unit ball of ℓ_2, by Cauchy-Schwarz,

$$|f(x+h) - f(x)|^2 = \sup_k \left| \sum_{i=1}^{k} a_{ki} h_i \right| \leq \sup_k \left[\sum_{i=1}^{k} a_{ki}^2 \sum_{i=1}^{k} h_i^2 \right] \leq \sup_k \sum_{i=1}^{k} a_{ki}^2 = \sup_k EX^2 (t_k)$$

Therefore,

$$\|f\|_{\mathrm{Lip2}} \leq \sigma^2(X), \quad \text{where } \sigma^2 = \sigma^2(X) := \sup_{t \in T} EX^2(t)$$

Recall from an argument at the beginning of the proof of Theorem 2.1 .20 that for the processes X we are considering here, $\sigma^2 < \infty$ and the median $M < \infty$. Then Theorem 2.2 .6 applies to the function f and gives the following concentration inequality:

Theorem 4.1.13 — The Borell-Sudakov-Tsirelson concentration inequality for Gaussian processes. Let $X(t), t \in T$, be a centred separable Gaussian process such that t $\Pr\left\{\sup_{t \in T} |X(t)| < \infty\right\} > 0$, and let M be the median of $\sup_{t \in T} |X(t)|$ and σ^2 the supremum of the variances $EX^2(t)$. Then, for all $u > 0$

$$\Pr\left\{\sup_{t \in T} |X(t)| > M + u\right\} \leq 1 - \Phi(u/\sigma), \quad \Pr\left\{\sup_{t \in T} |X(t)| < M - u\right\} \leq 1 - \Phi(u/\sigma) \tag{4.36}$$

and hence,

$$\Pr\left\{\left|\sup_{t \in T}\right| X(t)| - M| > u\right\} \leq 2(1 - \Phi(u/\sigma)) \leq e^{-u^2/2\sigma^2} \tag{4.37}$$

Inequality (4.37) is also true with the median M of $\sup_{t \in T} |X(t)|$ replaced by the expectation $E\left(\sup_{t \in T} |X(t)|\right)$, as we will see in Section 2.5 as a consequence of the Gaussian logarithmic Sobolev inequality (other proofs are possible; see Section 2.1 for a simple proof of a weaker version). But such a result, in its sharpest form, does not seem to be obtainable from (4.37). However, notice that if we integrate in (4.37) and let g be a $N(0,1)$ random variable, we obtain

$$|\mathrm{E}\sup_{t \in T}| X(t)| - M| \leq E\left|\sup_{t \in T}\right| X(t)| - M| \leq \sigma E|g| = \sqrt{2/\pi}\sigma \tag{4.38}$$

an inequality which is interesting in its own right and which gives, by combining with the same (4.37)

$$\Pr\left\{\left|\sup_{t\in T}\left|X(t)\right|-E\sup_{t\in T}\left|X(t)\right|\right|>u+\sqrt{2/\pi}\sigma\right\}\leq e^{-u^2/2\sigma^2} \tag{4.39}$$

which is of the right order for large values of u. Theorem 4.1.13, or even (4.39), expresses the remarkable fact that the supremum of a Gaussian process $X(t)$, centred at its mean or at its median, has tail probabilities not worse than those of a normal variable with the largest of the variances $EX^2(t), t \in T$. In particular, if we knew the size of $E\sup_{t\in T}|X(t)|$, we would have a very exact knowledge of the distribution of sup $_{t\in T}|X(t)|$. This will be the object of the next two sections.

We complete this section with simple applications of Theorem 4.1.13 to integrability and moments of the supremum of a Gaussian processes.

Corollary 4.1.14 Let $X(t), t \in T$, be a Gaussian process as in Theorem 4.1.13. Let M and σ also be as in this theorem, and write $\|X\| := \sup_{t\in T}|X(t)|$ to ease notation. Then there exists $K < \infty$ such that with the same hypothesis and notation as in the preceding corollary, for all $p \geq 1$

$$\left(E\|X\|^p\right)^{1/p} \leq 2E\|X\| + \left(E|g|^p\right)^{1/p}\sigma \leq K\sqrt{p}E\|X\|$$

for some absolute constant K

Proof. Just integrate inequality (2.18) with respect to $pt^{p-1}dt$ and then use that $M \leq 2E\|X\|$ (by Chebyshev) and that $\sigma \leq \sqrt{\pi/2}\sup_{t\in T}E|X(t)|$. See Exercise 2.1 .2 ■

Corollary 4.1.15 Let $X(t), t \in T$, be a Gaussian process as in Theorem 2.2.7, and let $\|X\|, M$ and σ be as in Corollary 2.2.8. Then

$$\lim_{u\to\infty}\frac{1}{u^2}\log\Pr\{\|X\|>u\} = -\frac{1}{2\sigma^2}$$

and

$$Ee^{\lambda\|X\|^2} < \infty \text{ if and only if } \lambda < \frac{1}{2\sigma^2}$$

Proof. The first limit follows from the facts that the first inequality in (2.17) can be rewritten as

$$\frac{1}{(u-M)^2}\log\Pr\{\|X\|>u\} \leq -\frac{1}{\sigma^2}$$

and that $\Pr\{\|X\|>u\} \geq \Pr\{|X(t)|>u\}$ for all $t \in T($ as, for a $N(0,1)$ variable g, we do have $u^{-2}\log\Pr\{|g|>u/a\} \to -1/2a^2$, e.g., by l'Hôpital's rule). For the second statement, just apply the first limit to $Ee^{\lambda\|X\|} = 1 + \int_0^\infty \int_0^{\lambda\|X\|^2} e^v dv d\mathcal{L}(\|X\|)(u) = 1 + \int_0^\infty e^v \Pr\{\|X\| >$

$\sqrt{v/\lambda}\}dv$

■

The Metric Entropy Bound for Suprema of Sub-Gaussian Processes

In this section we define sub-Gaussian processes and obtain the celebrated Dudley's entropy bound for their supremum norm. We are careful about the constants, as they are of some consequence in statistical estimations, at the expense of making the 'chaining argument' (proof of Theorem **??**) slightly more complicated than it could be. Combined with concentration inequalities, these bounds yield good estimates of the distribution of the supremum of a Gaussian process. They also constitute sufficient conditions for sample boundedness and sample continuity of Gaussian and sub-Gaussian processes and provide moduli of continuity for their sample paths which are effectively sharp in light of Sudakov's inequality derived in the next section.

A square integrable random variable ξ is said to be $sub-$Gaussian with parameter $\sigma > 0$ if for all $\lambda \in \mathbb{R}$

$$Ee^{\lambda\xi} \leq e^{\lambda^2\sigma^2/2}$$

Developing the two exponentials, dividing by $\lambda > 0$ and by $\lambda < 0$ and letting $\lambda \to 0$ in each case yield $E\xi = 0$; that is, sub-Gaussian random variables are automatically centred. Then, if in the two developments once the expectation term is cancelled, we divide by λ^2 and let $\lambda \to 0$, we obtain $E\xi^2 \leq \sigma^2$.

Aside from normal variables, perhaps the main examples of sub-Gaussian variables are the linear combinations of independent Rademacher (or symmetric Bernoulli) random

variables $\xi = \sum_{i=1}^n a_i\varepsilon_i$, where ε_i are independent identically distributed and $\Pr\{\varepsilon_i = 1\} = \Pr\{\varepsilon_i = -1\} = 1/2$. To see that these variables are sub-Gaussian, just note that by Taylor expansion, if ε is a Rademacher variable,

$$Ee^{\lambda\varepsilon} = \left(e^{\lambda} + e^{-\lambda}\right)/2 \leq e^{\lambda^2/2}, \quad \lambda \in \mathbb{R}$$

so that, by independence,

$$Ee^{\lambda\sum a_i\varepsilon_i} \leq e^{\lambda^2\sum a_i^2/2}$$

Both for Gaussian and for linear combinations of independent Rademacher variables, $\sigma^2 = E\xi^2$ The distributions of sub-Gaussian variables have sub-Gaussian tails: Chebyshev's inequality in exponential form, namely,

$$\Pr\{\xi \geq t\} = \Pr\left\{e^{\lambda\xi} \geq e^{\lambda t}\right\} \leq e^{\lambda^2\sigma^2/2-\lambda t}, \quad t > 0, \lambda > 0$$

with $\lambda = t/\sigma^2$ and applied as well to $-\xi$, gives that if ξ is sub-Gaussian for σ^2, then

$$\begin{aligned}&\Pr\{\xi \geq t\} \leq e^{-t^2/2\sigma^2} \text{ and } \Pr\{\xi \leq -t\} \leq e^{-t^2/2\sigma^2}, \quad \text{hence,}\\ &\Pr\{|\xi| \geq t\} \leq 2e^{-t^2/2\sigma^2}, \quad t > 0\end{aligned} \tag{4.40}$$

The last inequality in (4.40) in the case of linear combinations of independent Rademacher variables is called Hoeffding 's inequality. Of course, we can be more precise about the tail probabilities of normal variables: simple calculus gives that for all $t > 0$,

$$\frac{t}{t^2+1} e^{-t^2/2} \leq \int_t^\infty e^{-u^2/2} du \leq \min\left(t^{-1}, \sqrt{\pi/2}\right) e^{-t^2/2} \tag{4.41}$$

(see Exercise 2.2 .8). Back to the inequalities (2.22), we notice that if they hold for ξ, then ξ/c enjoys square exponential integrability for some $0 < c < \infty$: if $c^2 > 2\sigma^2$, then

$$Ee^{\xi^2/c^2} = \int_0^\infty 2te^{t^2} \Pr\{|\xi| > ct\} dt \leq \frac{2}{c^2/2\sigma^2 - 1} < \infty \tag{4.42}$$

The collection of random variables ξ on $(\Omega, \Sigma, \Pr)$ that satisfy this integrability property constitutes a vector space, denoted by $L^{\psi 2}(\Omega, \Sigma, \Pr)$, and the functional

$$\|\xi\|_{\psi_2} = \inf\{c > 0 : E\psi_2(|\xi|/c) \leq 1\}$$

where $\psi_2(x) := e^{x^2} - 1$(a convex function which is zero at zero) is a pseudo-norm on it for which L^{ψ_2}, with identification of a.s. equal functions, is a Banach space (Exercise 2.3 .5). With this definition, inequality (4.42) shows that

$$\Pr\{|\xi| \geq t\} \leq 2e^{-t^2/2\sigma^2} \quad \text{for all } t > 0 \quad \text{implies} \quad \|\xi\|_{\psi_2} \leq \sqrt{6}\sigma \tag{4.43}$$

To complete the set of relationships developed so far, suppose that $\xi \in L^{\psi_2}$ and $E\xi = 0$, and let us show that ξ is sub-Gaussian. We have

$$Ee^{\lambda\xi} - 1 \leq E\sum_{k=2}^\infty \left|\lambda^k \xi^k\right| / k! \leq \frac{\lambda^2}{2} E\left(\xi^2 e^{|\lambda\xi|}\right)$$

Now we estimate the exponent $|\lambda\xi|$ on the region $|\xi| > 2\lambda\|\xi\|_{\psi_2}^2$ and on its complement to obtain, after multiplying and dividing by $\|\xi\|_{\psi_2}^2$ and using that $a < e^{a/2}$ for all $a > 0$,

$$\begin{aligned}\frac{\lambda^2}{2} E\left(\xi^2 e^{|\lambda\xi|}\right) &\leq \frac{\lambda^2\|\xi\|_{\psi_2}^2}{2} e^{2\lambda^2\|\xi\|_{\psi_2}^2} E\left(\frac{\xi^2}{\|\xi\|_{\psi_2}^2} e^{\xi^2/2\|\xi\|_{\psi_2}^2}\right)\\ &\leq \lambda^2\|\xi\|_{\psi_2}^2 e^{2\lambda^2\|\xi\|_{\psi_2}^2} Ee^{\xi^2/\|\xi\|_{\psi_2}^2}/2 \leq \lambda^2\|\xi\|_{\psi_2}^2 e^{2\lambda^2\|\xi\|_{\psi_2}^2}\end{aligned}$$

Using $1 + a \leq e^a$, the last two bounds give

$$Ee^{\lambda\xi} \leq e^{3\lambda^2\|\xi\|_{\psi_2}^2} \tag{4.44}$$

showing that ξ is sub-Gaussian with $\sigma \leq \sqrt{6}\|\xi\|_{\psi_2}$. If ξ is symmetric, just developing the exponential gives the better inequality $Ee^{\lambda\xi} \leq e^{\lambda^2\|\xi\|_{\psi_2}^2/2}$ We collect these facts:

Lemma 4.2 If ξ is sub-Gaussian for a constant $\sigma > 0$, then it satisfies the sub-Gaussian tail inequalities (4.40), and therefore, $\xi \in L^{\psi_2}$, with $\|\xi\|_{\psi_2} \leq \sqrt{6}\sigma$. Conversely, if ξ is in L^{ψ_2} and is centred, then it is sub-Gaussian for the constant $\sigma \leq \sqrt{6}\|\xi\|_{\psi_2}$, and in particular, it also satisfies the inequalities (4.40) for $\sigma = \sqrt{6}\|\xi\|_{\psi_2}$

In other words, ignoring constants, for ξ centred, the conditions (a) $\xi \in L^{\psi_2}$ and (b) ξ satisfies the sub-Gaussian tail inequalities (4.40) for some σ_1 and $(c)\xi$ is sub-Gaussian for some σ_2 are all equivalent.

Lemma 2.3 .1 extends to random variables whose tail probabilities are bounded by a constant times the sub-Gaussian probabilities in (4.40) as follows.

Lemma 4.3 Assume that

$$\Pr\{|\xi| \geq t\} \leq 2Ce^{-t^2/2\sigma^2}, \quad t > 0 \tag{4.45}$$

for some $C \geq 1$ and $\sigma > 0$, a condition implied by the Laplace transform condition

$$Ee^{\lambda\xi} \leq Ce^{\lambda^2\sigma^2/2}, \quad \lambda \in \mathbb{R} \tag{4.46}$$

Then ξ also satisfies

$$\|\xi\|_{\psi_2} \leq \sqrt{2(2C+1)}\sigma \tag{4.47}$$

Moreover, if in addition $E\xi = 0$, then also

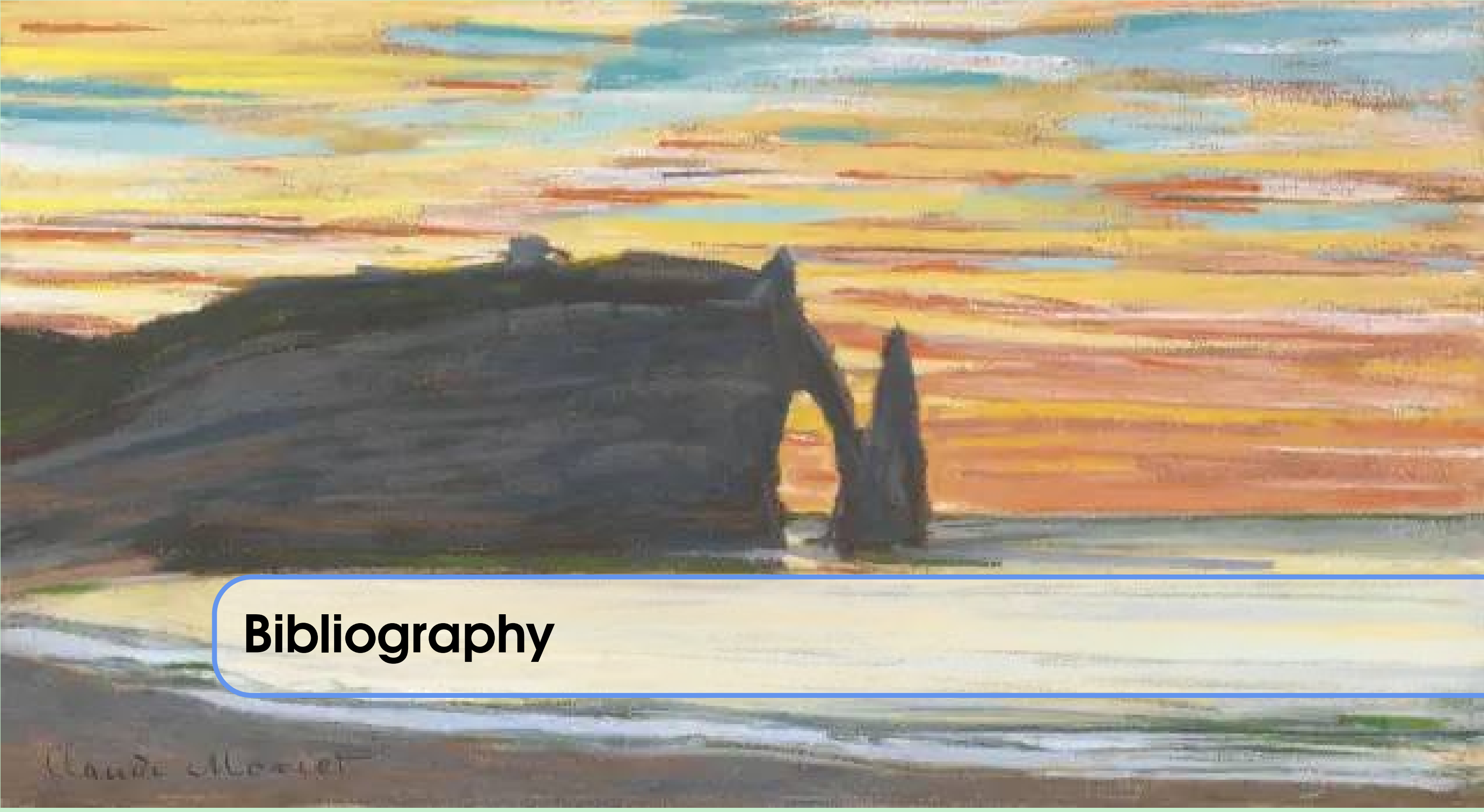
$$Ee^{\lambda\xi} \leq e^{3\lambda^2(2(2C+1))\sigma^2}, \quad \lambda \in \mathbb{R} \tag{4.48}$$

that is, ξ is sub-Gaussian with constant $\tilde{\sigma}^2 = 12(2C+1)\sigma^2$.

Proof. The proof of inequality (4.40) shows that (4.46) implies (4.45). The preceding proof showing that (4.40) implies (4.43), with only formal changes, proves that (4.45) implies (4.47). Finally, inequality (4.48) follows from (4.47) and (4.44). ■

This lemma is useful in that showing that a variable ξ is sub-Gaussian reduces to proving the tail probability bounds (4.45) for some $C > 1$, which may be easier than proving them for $C = 1$

Part Four



Bibliography