Chapter 7

# Concentration of Measure

*Often we want to show that some random quantity is close to its mean with high probability. Results of this kind are known as* **concentration inequalities**. *In this chapter we consider some important concentration results such as Hoeffding's inequality, Bernstein's inequality and McDiarmid's inequality. Then we consider* **uniform bounds** *that guarantee that a set of random quantities are simultaneously close to their means with high probabilty.*

## 7.1 Introduction

Often we need to show that a random quantity is close to its mean. For example, later we will prove Hoeffding's inequality which implies that, if $Z_1, \ldots, Z_n$ are Bernoulli random variables with mean $\mu$ then

$$\mathbb{P}(|\overline{Z} - \mu| > \epsilon) \leq 2e^{-2n\epsilon^2}$$

where $\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$.

More generally, we want a result of the form

$$\mathbb{P}\Big(\big|f(Z_1, \ldots, Z_n) - \mu_n(f)\big| > \epsilon\Big) < \delta_n \tag{7.1}$$

where $\mu_n(f) = \mathbb{E}(f(Z_1, \ldots, Z_n))$ and $\delta_n \to 0$ as $n \to \infty$. Such results are known as *concentration inequalities* and the phenomenon that many random quantities are close to their mean with high probability is called *concentration of measure*. These results are

fundamental for establishing performance guarantees of many algorithms. For statistical learning theory, we will need *uniform bounds* of the form

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|f(Z_1,\ldots,Z_n)-\mu_n(f)\right|>\epsilon\right)<\delta_n \tag{7.2}$$

over a class of functions $\mathcal{F}$.

**7.3 Example.** To motivate the need for such results, consider empirical risk minimization in classification. Suppose we have data $(X_1, Y_1)$, ..., $(X_n, Y_n)$ where $Y_i \in \{0, 1\}$ and $X_i \in \mathbb{R}^d$. Let $h : \mathbb{R}^d \to \{0, 1\}$ be a classifier. The training error is

$$\widehat{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \neq h(X_i))$$

and the true classification error is

$$R(h) = \mathbb{P}(Y \neq h(X)).$$

We would like to know if $\widehat{R}(h)$ is close to $R(h)$ with high probability. This is precisely of the form (7.1) with $Z_i = (X_i, Y_i)$ and $f(Z_1, \ldots, Z_n) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \neq h(X_i))$.

Now let $\mathcal{H}$ be a set of classifiers. Let $\widehat{h}$ minimize the training error $\widehat{R}(h)$ over $\mathcal{H}$ and let $h_*$ minimize the true error $R(h)$ over $\mathcal{H}$. Can we guarantee that the risk $R(\widehat{h})$ of the selected classifier is close to the risk $R(h_*)$ of the best classifier? Let $\mathcal{E}$ denote the event that $\sup_{h\in\mathcal{H}} |\widehat{R}_n(h) - R(h)| \leq \epsilon$. When the event $\mathcal{E}$ holds, we have that

$$R(h_*) \leq R(\widehat{h}) \leq \widehat{R}_n(\widehat{h}) + \epsilon \leq \widehat{R}_n(h_*) + \epsilon \leq R(h_*) + 2\epsilon$$

where we used the following facts: $h_*$ minimizes $R$, $\mathcal{E}$ holds, $\widehat{h}$ minimizes $\widehat{R}_n$, $\mathcal{E}$ holds and $h_*$ minimizes $R$. It follows that, when $\mathcal{E}$ holds, $|R(\widehat{h}) - R(h_*)| \leq 2\epsilon$. Concentration of measure is used to prove that $\mathcal{E}$ holds with high probability. □

Besides classification, concentration inequalities are used for studying many other methods such as clustering, random projections and density estimation.

**Notation**

If $P$ is a probability measure and $f$ is a function then we write

$$Pf = P(f) = \int f(z)dP(z) = \mathbb{E}(f(Z)).$$

Given $Z_1, \ldots, Z_n$, let $P_n$ denote the empirical measure that puts mass $1/n$ at each data point:

$$P_n(A) = \frac{1}{n}\sum_{i=1}^{n} I(Z_i \in A)$$

where $I(Z_i \in A) = 1$ if $Z_i \in A$ and $I(Z_i \in A) = 0$ otherwise. Then we write

$$P_n f = P_n(f) = \int f(z)dP_n(z) = \frac{1}{n}\sum_{i=1}^{n} f(Z_i).$$

## 7.2 Basic Inequalities

### 7.2.1 Hoeffding's Inequality

Suppose that $Z$ has a finite mean and that $\mathbb{P}(Z \geq 0) = 1$. Then, for any $\epsilon > 0$,

$$\mathbb{E}(Z) = \int_0^\infty z\, dP(z) \geq \int_\epsilon^\infty z\, dP(z) \geq \epsilon \int_\epsilon^\infty dP(z) = \epsilon\, \mathbb{P}(Z > \epsilon) \tag{7.4}$$

which yields *Markov's inequality*:

$$\mathbb{P}(Z > \epsilon) \leq \frac{\mathbb{E}(Z)}{\epsilon}. \tag{7.5}$$

An immediate consequence of Markov's inequality is *Chebyshev's inequality*

$$\mathbb{P}(|Z - \mu| > \epsilon) = \mathbb{P}(|Z - \mu|^2 > \epsilon^2) \leq \frac{\mathbb{E}(Z-\mu)^2}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2} \tag{7.6}$$

where $\mu = \mathbb{E}(Z)$ and $\sigma^2 = \mathsf{Var}(Z)$. If $Z_1, \ldots, Z_n$ are iid with mean $\mu$ and variance $\sigma^2$ then, since $\mathsf{Var}(\overline{Z}_n) = \sigma^2/n$, Chebyshev's inequality yields

$$\mathbb{P}(|\overline{Z}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}. \tag{7.7}$$

While this inequality is useful, it does not decay exponentially fast as $n$ increases. To improve the inequality, we use *Chernoff's method*: for any $t > 0$,

$$\mathbb{P}(Z > \epsilon) = \mathbb{P}(e^Z > e^\epsilon) = \mathbb{P}(e^{tZ} > e^{t\epsilon}) \leq e^{-t\epsilon}\mathbb{E}(e^{tZ}). \tag{7.8}$$

We then minimize over $t$ and conclude that:

$$\mathbb{P}(Z > \epsilon) \leq \inf_{t \geq 0} e^{-t\epsilon}\, \mathbb{E}(e^{tZ}). \tag{7.9}$$

To use the above result we need to bound the moment generating function $\mathbb{E}(e^{tZ})$.

**7.10 Lemma.** *Let $Z$ be a mean $\mu$ random variable such that $a \leq Z \leq b$. Then, for any $t$,*

$$\mathbb{E}(e^{tZ}) \leq e^{t\mu + t^2(b-a)^2/8}. \tag{7.11}$$

***Proof.*** For simplicity, assume that $\mu = 0$. Since $a \leq Z \leq b$, we can write $Z$ as a convex combination of $a$ and $b$, namely, $Z = \alpha b + (1-\alpha)a$ where $\alpha = (Z-a)/(b-a)$. By the convexity of the function $y \to e^{ty}$ we have

$$e^{tZ} \leq \alpha e^{tb} + (1-\alpha)e^{ta} = \frac{Z-a}{b-a}e^{tb} + \frac{b-Z}{b-a}e^{ta}.$$

Take expectations of both sides and use the fact that $\mathbb{E}(Z) = 0$ to get

$$\mathbb{E}e^{tZ} \leq -\frac{a}{b-a}e^{tb} + \frac{b}{b-a}e^{ta} = e^{g(u)} \tag{7.12}$$

where $u = t(b-a)$, $g(u) = -\gamma u + \log(1 - \gamma + \gamma e^u)$ and $\gamma = -a/(b-a)$. Note that $g(0) = g'(0) = 0$. Also, $g''(u) \leq 1/4$ for all $u > 0$. By Taylor's theorem, there is a $\xi \in (0, u)$ such that

$$g(u) \ = g(0) + ug'(0) + \frac{u^2}{2}g''(\xi) = \frac{u^2}{2}g''(\xi) \leq \frac{u^2}{8} = \frac{t^2(b-a)^2}{8}.$$

Hence, $\mathbb{E}e^{tZ} \leq e^{g(u)} \leq e^{t^2(b-a)^2/8}$. □

**7.13 Theorem (Hoeffding).** *If $Z_1, Z_2, \ldots, Z_n$ are independent with $\mathbb{P}(a \leq Z_i \leq b) = 1$ and common mean $\mu$ then for any $t > 0$*

$$\mathbb{P}(|\overline{Z}_n - \mu| > \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2} \tag{7.14}$$

*where $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$.*

***Proof.*** For simplicity assume that $\mathbb{E}(Z_i) = 0$. Now we use the Chernoff method. For any $t > 0$, we have, from Markov's inequality, that

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Z_i \geq \epsilon\right) &= \mathbb{P}\left(\frac{t}{n}\sum_{i=1}^n Z_i \geq t\epsilon\right) = \mathbb{P}\left(e^{(t/n)\sum_{i=1}^n Z_i} \geq e^{t\epsilon}\right) \\ &\leq e^{-t\epsilon}\mathbb{E}\left(e^{(t/n)\sum_{i=1}^n Z_i}\right) = e^{-t\epsilon}\prod_i \mathbb{E}(e^{(t/n)Z_i}) \qquad (7.15) \\ &\leq e^{-t\epsilon}e^{(t^2/n^2)\sum_{i=1}^n (b_i-a_i)^2/8} \qquad (7.16)\end{aligned}$$

where the last inequality follows from Lemma 7.10. Now we minimize the right hand side over $t$. In particular, we set $t = 4\epsilon n^2/\sum_{i=1}^n (b_i - a_i)^2$ and get $\mathbb{P}\left(\overline{Z}_n \geq \epsilon\right) \leq e^{-2n\epsilon^2/c}$. By a similar argument, $\mathbb{P}\left(\overline{Z}_n \leq -\epsilon\right) \leq e^{-2n\epsilon^2/c}$ and the result follows. □

**7.17 Corollary.** *If $Z_1, Z_2, \ldots, Z_n$ are independent with $\mathbb{P}(a_i \leq Z_i \leq b_i) = 1$ and common mean $\mu$, then, with probability at least $1 - \delta$,*

$$|\overline{Z}_n - \mu| \leq \sqrt{\frac{c}{2n}\log\left(\frac{2}{\delta}\right)} \qquad (7.18)$$

*where $c = \frac{1}{n}\sum_{i=1}^n (b_i - a_i)^2$.*

**7.19 Corollary.** *If $Z_1, Z_2, \ldots, Z_n$ are independent Bernoulli random variables with $\mathbb{P}(Z_i = 1) = p$ then, for any $\epsilon > 0$, $\mathbb{P}(|\overline{Z}_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2}$. Hence, with probability at least $1 - \delta$ we have that $|\overline{Z}_n - p| \leq \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}$.*

**7.20 Example (Classification).** Returning to the classification problem, let $h$ be a classifier and let $f(z) = I(y \neq h(x)$ where $z = (x, y)$. Then Hoeffding's inequality implies that $|R(h) - \widehat{R}_n(h)| \leq \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}$ with probability at least $1 - \delta$. □

The following result extends Hoeffding's inequality to more general functions $f(z_1, \ldots, z_n)$.

**7.21 Theorem (McDiarmid).** *Let $Z_1, \ldots, Z_n$ be independent random variables. Suppose that*

$$\sup_{z_1,\ldots,z_n,z_i'} \left| f(z_1,\ldots,z_{i-1},z_i,z_{i+1},\ldots,z_n) - f(z_1,\ldots,z_{i-1},z_i',z_{i+1},\ldots,z_n) \right| \leq c_i \tag{7.22}$$

*for $i = 1, \ldots, n$. Then*

$$\mathbb{P}\left( \left| f(Z_1,\ldots,Z_n) - \mathbb{E}(f(Z_1,\ldots,Z_n)) \right| \geq \epsilon \right) \leq 2\, \exp\left( -\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2} \right). \tag{7.23}$$

***Proof.*** Let $Y = f(Z_1, \ldots, Z_n)$ and $\mu = \mathbb{E}(f(Z_1, \ldots, Z_n))$. Then

$$\mathbb{P}\Big( |Y - \mu| \geq \epsilon \Big) = \mathbb{P}\Big( Y - \mu \geq \epsilon \Big) + \mathbb{P}\Big( Y - \mu \leq -\epsilon \Big).$$

We will bound the first quantity. The second follows similarly. Let $V_i = \mathbb{E}(Y|Z_1, \ldots, Z_i) - \mathbb{E}(Y|Z_1, \ldots, Z_{i-1})$. Then

$$f(Z_1,\ldots,Z_n) - \mathbb{E}(f(Z_1,\ldots,Z_n)) = \sum_{i=1}^n V_i$$

and $\mathbb{E}(V_i|Z_1, \ldots, Z_{i-1}) = 0$. Using a similar argument as in Lemma 7.10, we have

$$\mathbb{E}(e^{tV_i}|Z_1,\ldots,Z_{i-1}) \leq e^{t^2 c_i^2/8}. \tag{7.24}$$

Now, for any $t > 0$,

$$\begin{aligned}
\mathbb{P}\left(Y - \mu \geq \epsilon\right) &= \mathbb{P}\left( \sum_{i=1}^n V_i \geq \epsilon \right) = \mathbb{P}\left( e^{t\sum_{i=1}^n V_i} \geq e^{t\epsilon} \right) \leq e^{-t\epsilon}\mathbb{E}\left( e^{t\sum_{i=1}^n V_i} \right) \\
&= e^{-t\epsilon}\mathbb{E}\left( e^{t\sum_{i=1}^{n-1} V_i} \mathbb{E}\left( e^{tV_n} \,\middle|\, Z_1,\ldots,Z_{n-1} \right) \right) \\
&\leq e^{-t\epsilon} e^{t^2 c_n^2/8} \mathbb{E}\left( e^{t\sum_{i=1}^{n-1} V_i} \right) \cdots \leq e^{-t\epsilon} e^{t^2 \sum_{i=1}^n c_i^2}.
\end{aligned}$$

The result follows by taking $t = 4\epsilon / \sum_{i=1}^n c_i^2$. □

**Remark:** If $f(z_1, \ldots, z_n) = \frac{1}{n}\sum_{i=1}^n z_i$ then we get back Hoeffding's inequality.

**7.25 Example.** Let $X_1, \ldots, X_n \sim P$ and let $P_n(A) = \frac{1}{n}\sum_{i=1}^n I(X_i \in A)$. Define $\Delta_n \equiv f(X_1, \ldots, X_n) = \sup_A |P_n(A) - P(A)|$. Changing one observation changes $f$ by at most

$1/n$. Hence,

$$\mathbb{P}\left( |\Delta_n - \mathbb{E}(\Delta_n)| > \epsilon \right) \leq 2e^{-2n\epsilon^2}.$$

□

### 7.2.2 Sharper Inequalities

Hoeffding's inequality does not use any information about the random variables except the fact that they are bounded. If the variance of $X_i$ is small, then we can get a sharper inequality from Bernstein's inequality. We begin with a preliminary result.

**7.26 Lemma.** *Suppose that* $|X| \leq c$ *and* $\mathbb{E}(X) = 0$. *For any* $t > 0$,

$$\mathbb{E}(e^{tX}) \leq \exp\left\{ t^2\sigma^2 \left( \frac{e^{tc} - 1 - tc}{(tc)^2} \right) \right\} \tag{7.27}$$

*where* $\sigma^2 = \mathrm{Var}(X)$.

***Proof.*** Let $F = \sum_{r=2}^{\infty} \frac{t^{r-2}\mathbb{E}(X^r)}{r!\sigma^2}$. Then,

$$\mathbb{E}(e^{tX}) \ = \mathbb{E}\left( 1 + tx + \sum_{r=2}^{\infty} \frac{t^r X^r}{r!} \right) = 1 + t^2\sigma^2 F \leq e^{t^2\sigma^2 F}. \tag{7.28}$$

For $r \geq 2$, $\mathbb{E}(X^r) = \mathbb{E}(X^{r-2}X^2) \leq c^{r-2}\sigma^2$ and so

$$F \leq \sum_{r=2}^{\infty} \frac{t^{r-2}c^{r-2}\sigma^2}{r!\sigma^2} = \frac{1}{(tc)^2} \sum_{i=2}^{\infty} \frac{(tc)^r}{r!} = \frac{e^{tc} - 1 - tc}{(tc)^2}. \tag{7.29}$$

Hence, $\mathbb{E}(e^{tX}) \leq \exp\left\{ t^2\sigma^2 \frac{e^{tc}-1-tc}{(tc)^2} \right\}$. □

**7.30 Theorem (Bernstein).** *If* $\mathbb{P}(|X_i| \leq c) = 1$ *and* $\mathbb{E}(X_i) = \mu$ *then, for any* $\epsilon > 0$,

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) \leq 2\exp\left\{ -\frac{n\epsilon^2}{2\sigma^2 + 2c\epsilon/3} \right\} \tag{7.31}$$

*where* $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \mathsf{Var}(X_i)$.

***Proof.*** For simplicity, assume that $\mu = 0$. From Lemma 7.26,

$$\mathbb{E}(e^{tX_i}) \leq \exp\left\{ t^2\sigma_i^2 \frac{e^{tc} - 1 - tc}{(tc)^2} \right\} \tag{7.32}$$

where $\sigma_i^2 = \mathbb{E}(X_i^2)$. Now,

$$\mathbb{P}\big(\overline{X}_n > \epsilon\big) = \mathbb{P}\left(\sum_{i=1}^n X_i > n\epsilon\right) = \mathbb{P}\Big(e^{t\sum_{i=1}^n X_i} > e^{tn\epsilon}\Big) \tag{7.33}$$

$$\leq e^{-tn\epsilon}\mathbb{E}(e^{t\sum_{i=1}^n X_i}) = e^{-tn\epsilon}\prod_{i=1}^n \mathbb{E}(e^{tX_i}) \tag{7.34}$$

$$\leq e^{-tn\epsilon}\exp\left\{nt^2\sigma^2\frac{e^{tc}-1-tc}{(tc)^2}\right\}. \tag{7.35}$$

Take $t = (1/c)\log(1+\epsilon c/\sigma^2)$ to get

$$\mathbb{P}(\overline{X}_n > \epsilon) \leq \exp\left\{-\frac{n\sigma^2}{c^2}h\left(\frac{c\epsilon}{\sigma^2}\right)\right\} \tag{7.36}$$

where $h(u) = (1+u)\log(1+u) - u$. The results follows by noting that $h(u) \geq u^2/(2 + 2u/3)$ for $u \geq 0$. □

A useful corollary is the following.

**7.37 Lemma.** *Let $X_1, \ldots, X_n$ be iid and suppose that $|X_i| \leq c$ and $\mathbb{E}(X_i) = \mu$. With probability at least $1-\delta$,*

$$|\overline{X}_n - \mu| \leq \sqrt{\frac{2\sigma^2\log(1/\delta)}{n}} + \frac{2c\log(1/\delta)}{3n}. \tag{7.38}$$

*In particular, if $\sigma \leq \sqrt{2c^2\log(1/\delta)/(9n)}$, then with probability at least $1-\delta$,*

$$|\overline{X}_n - \mu| \leq \frac{C}{n} \tag{7.39}$$

*where $C = 4c\log\left(\frac{1}{\delta}\right)/3$.*

We also get a very specific inequality in the special case that $X$ is Gaussian.

**7.40 Theorem.** *Suppose that $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$.*

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right). \tag{7.41}$$

***Proof.*** Let $X \sim N(0,1)$ with density $\phi(x) = (2\pi)^{-1/2}e^{-x^2/2}$ and distribution function $\Phi(x) = \int_{-\infty}^x \phi(s)ds$. For any $\epsilon > 0$,

$$\mathbb{P}(X > \epsilon) = \int_\epsilon^\infty \phi(s)ds \leq \frac{1}{\epsilon}\int_\epsilon^\infty s\phi(s)ds = -\frac{1}{\epsilon}\int_\epsilon^\infty \phi'(s)ds = \frac{\phi(\epsilon)}{\epsilon} \leq \frac{1}{\epsilon}e^{-\epsilon^2/2}. \tag{7.42}$$

By symmetry we have that

$$\mathbb{P}(|X| > \epsilon) \leq \frac{2}{\epsilon} e^{-\epsilon^2/2}.$$

Now suppose that $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Then $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i \sim N(\mu, \sigma^2/n)$. Let $Z \sim N(0,1)$. Then,

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) = \mathbb{P}\Big(\sqrt{n}|\overline{X}_n - \mu|/\sigma > \sqrt{n}\epsilon/\sigma\Big) = \mathbb{P}\Big(|Z| > \sqrt{n}\epsilon/\sigma\Big) \tag{7.43}$$

$$\leq \frac{2\sigma}{\epsilon\sqrt{n}} e^{-n\epsilon^2/(2\sigma^2)} \leq e^{-n\epsilon^2/(2\sigma^2)} \tag{7.44}$$

for all large $n$. □

### 7.2.3 Bounds on Expected Values

Suppose we have an exponential bound on $\mathbb{P}(X_n > \epsilon)$. In that case we can bound $\mathbb{E}(X_n)$ as follows.

**7.45 Theorem.** *Suppose that $X_n \geq 0$ and that for every $\epsilon > 0$,*

$$\mathbb{P}(X_n > \epsilon) \leq c_1 e^{-c_2 n \epsilon^2} \tag{7.46}$$

*for some $c_2 > 0$ and $c_1 > 1/e$. Then, $\mathbb{E}(X_n) \leq \sqrt{\frac{C}{n}}$ where $C = (1 + \log(c_1))/c_2$.*

***Proof.*** Recall that for any nonnegative random variable $Y$, $\mathbb{E}(Y) = \int_0^\infty \mathbb{P}(Y \geq t)dt$. Hence, for any $a > 0$,

$$\mathbb{E}(X_n^2) = \int_0^\infty \mathbb{P}(X_n^2 \geq t)dt = \int_0^a \mathbb{P}(X_n^2 \geq t)dt + \int_a^\infty \mathbb{P}(X_n^2 \geq t)dt \leq a + \int_a^\infty \mathbb{P}(X_n^2 \geq t)dt.$$

Equation (7.46) implies that $\mathbb{P}(X_n > \sqrt{t}) \leq c_1 e^{-c_2 n t}$. Hence,

$$\mathbb{E}(X_n^2) \leq a + \int_a^\infty \mathbb{P}(X_n^2 \geq t)dt = a + \int_a^\infty \mathbb{P}(X_n \geq \sqrt{t})dt \leq a + c_1 \int_a^\infty e^{-c_2 n t} dt = a + \frac{c_1 e^{-c_2 n a}}{c_2\, n}.$$

Set $a = \log(c_1)/(nc_2)$ and conclude that

$$\mathbb{E}(X_n^2) \leq \frac{\log(c_1)}{nc_2} + \frac{1}{nc_2} = \frac{1 + \log(c_1)}{nc_2}.$$

Finally, we have $\mathbb{E}(X_n) \leq \sqrt{\mathbb{E}(X_n^2)} \leq \sqrt{\frac{1+\log(c_1)}{nc_2}}$. □

Now we consider bounding the maximum of a set of random variables.

**7.47 Theorem.** *Let $X_1, \ldots, X_n$ be random variables. Suppose there exists $\sigma > 0$ such that $\mathbb{E}(e^{tX_i}) \leq e^{t\sigma^2/2}$ for all $t > 0$. Then*

$$\mathbb{E}\left(\max_{1\leq i\leq n} X_i\right) \leq \sigma\sqrt{2\log n}. \tag{7.48}$$

***Proof.*** By Jensen's inequality,

$$\begin{aligned}\exp\left\{t\mathbb{E}\left(\max_{1\leq i\leq n} X_i\right)\right\} &\leq \mathbb{E}\left(\exp\left\{t\max_{1\leq i\leq n} X_i\right\}\right)\\ &= \mathbb{E}\left(\max_{1\leq i\leq n}\exp\{tX_i\}\right) \leq \sum_{i=1}^n \mathbb{E}\left(\exp\{tX_i\}\right) \leq ne^{t^2\sigma^2/2}.\end{aligned}$$

Thus, $\mathbb{E}\left(\max_{1\leq i\leq n} X_i\right) \leq \frac{\log n}{t} + \frac{t\sigma^2}{2}$. The result follows by setting $t = \sqrt{2\log n}/\sigma$. □

## 7.3 Uniform Bounds

### 7.3.1 Binary Functions

A binary function on a space $\mathcal{Z}$ is a function $f : \mathcal{Z} \to \{0, 1\}$. Let $\mathcal{F}$ be a class of binary functions on $\mathcal{Z}$. For any $z_1, \ldots, z_n$ define

$$\mathcal{F}_{z_1,\ldots,z_n} = \Big\{(f(z_1), \ldots, f(z_n)) : \ f \in \mathcal{F}\Big\}. \tag{7.49}$$

$\mathcal{F}_{z_1,\ldots,z_n}$ is a finite collection of binary vectors and $|\mathcal{F}_{z_1,\ldots,z_n}| \leq 2^n$. The set $\mathcal{F}_{z_1,\ldots,z_n}$ is called *the projection of $\mathcal{F}$ onto $z_1, \ldots, z_n$*.

**7.50 Example.** Let $\mathcal{F} = \{f_t : \ t \in \mathbb{R}\}$ where $f_t(z) = 1$ if $z > t$ and $f_t(z) = 0$ of $z \leq t$. Consider three real numbers $z_1 < z_2 < z_3$. Then

$$\mathcal{F}_{z_1,z_2,z_3} = \Big\{(0,0,0),\ (0,0,1),\ (0,1,1),\ (1,1,1)\Big\}.$$

□

Define the *growth function* or *shattering number* by

$$s(\mathcal{F}, n) = \sup_{z_1,\ldots,z_n} \left|\mathcal{F}_{z_1,\ldots,z_n}\right|. \tag{7.51}$$

A binary function $f$ can be thought of as an indicator function for a set, namely, $A = \{z : \ f(x) = 1\}$. Conversely, any set can be thought of as a binary function, namely, its indicator function $I_A(z)$. We can therefore re-express the growth function in terms of sets. If $\mathcal{A}$ is a class of subsets of $\mathbb{R}^d$ then $s(\mathcal{A}, n)$ is defined to be $s(\mathcal{F}, n)$ where $\mathcal{F} = |\{I_A : A \in \mathcal{A}\}|$ is the set of indicator functions and then $s(\mathcal{A}, n)$ is again called the *shattering number*. It follows that

$$s(\mathcal{A}, n) = \max_F s(\mathcal{A}, F)$$

where the maximum is over all finite sets of size $n$ and $s(\mathcal{A}, F) = |\{A \cap F : \ A \in \mathcal{A}|$ denotes the number of subsets of $F$ picked out by $\mathcal{A}$. We say that a finite set $F$ of size $n$ is *shattered* by $\mathcal{A}$ if $s(\mathcal{A}, F) = 2^n$.

**7.52 Theorem.** *Let $\mathcal{A}$ and $\mathcal{B}$ be classes of subsets of $\mathbb{R}^d$.*

1. $s(\mathcal{A}, n+m) \leq s(\mathcal{A}, n)s(\mathcal{A}, m)$.
2. *If* $\mathcal{C} = \mathcal{A} \bigcup \mathcal{B}$ *then* $s(\mathcal{C}, n) \leq s(\mathcal{A}, n) + s(\mathcal{B}, n)$
3. *If* $\mathcal{C} = \{A \bigcup B : \ A \in \mathcal{A}, B \in \mathcal{B}\}$ *then* $s(\mathcal{C}, n) \leq s(\mathcal{A}, n)s(\mathcal{B}, n)$.
4. *If* $\mathcal{C} = \{A \bigcap B : \ A \in \mathcal{A}, B \in \mathcal{B}\}$ *then* $s(\mathcal{C}, n) \leq s(\mathcal{A}, n)s(\mathcal{B}, n)$.

***Proof.*** See exercise 10. □

**VC Dimension.** Recall that a finite set $F$ of size $n$ is *shattered* by $\mathcal{A}$ if $s(\mathcal{A}, F) = 2^n$. The VC dimension (named after Vapnik and Chervonenkis) of $\mathcal{A}$ is the size of the largest set that can be shattered by $\mathcal{A}$.

The *VC dimension* of a class of set $\mathcal{A}$ is

$$\mathsf{VC}(\mathcal{A}) = \sup\Big\{n : \ s(\mathcal{A}, n) = 2^n\Big\}. \tag{7.53}$$

The *VC dimension* of a class of binary functions $\mathcal{F}$ is

$$\mathsf{VC}(\mathcal{F}) = \sup\Big\{n : \ s(\mathcal{F}, n) = 2^n\Big\}. \tag{7.54}$$

If the VC dimension is finite, then the growth function cannot grow too quickly. In fact, there is a phase transition: $s(\mathcal{F}, n) = 2^n$ for $n < d$ and then the growth switches to

polynomial.

**7.55 Theorem (Sauer's Theorem).** *Suppose that $\mathcal{F}$ has finite VC dimension $d$. Then,*

$$s(\mathcal{F}, n) \leq \sum_{i=0}^{d} \binom{n}{i} \tag{7.56}$$

*and for all $n \geq d$,*

$$s(\mathcal{F}, n) \leq \left(\frac{en}{d}\right)^d. \tag{7.57}$$

***Proof.*** When $n = d = 1$, (7.56) clearly holds. We proceed by induction. Suppose that (7.56) holds for $n-1$ and $d-1$ and also that it holds for $n-1$ and $d$. We will show that it holds for $n$ and $d$. Let $h(n,d) = \sum_{i=0}^{d} \binom{n}{i}$. We need to show that $\mathsf{VC}(\mathcal{F}) \leq d$ implies that $s(\mathcal{F}, n) \leq h(n,d)$. Let $F_1 = \{z_1, \ldots, z_n\}$ and $F_2 = \{z_2, \ldots, z_n\}$. Let $\mathcal{F}_1 = \{(f(z_1), \ldots, f(z_n) : f \in \mathcal{F}\}$ and $\mathcal{F}_2 = \{(f(z_2), \ldots, f(z_n) : f \in \mathcal{F}\}$. For $f, g \in \mathcal{F}$, write $f \sim g$ if $g(z_1) = 1 - f(z_1)$ and $g(z_j) = f(z_j)$ for $j = 2, \ldots, n$. Let

$$\mathcal{G} = \Big\{ f \in \mathcal{F} : \text{ there exists } g \in \mathcal{F} \text{ such that } g \sim f \Big\}.$$

Define $\mathcal{F}_3 = \{(f(z_2), \ldots, f(z_n)) : f \in \mathcal{G}\}$. Then $|\mathcal{F}_1| = |\mathcal{F}_2| + |\mathcal{F}_3|$. Note that $\mathsf{VC}(\mathcal{F}_2) \leq d$ and $\mathsf{VC}(\mathcal{F}_3) \leq d-1$. The latter follows since, if $\mathcal{F}_3$ shatters a set, then we can add $z_1$ to create a set that is shattered by $\mathcal{F}_1$. By assumption $|\mathcal{F}_2| \leq h(n-1, d)$ and $|\mathcal{F}_3| \leq h(n-1, d-1)$. Hence,

$$|\mathcal{F}_1| \leq h(n-1,d) + h(n-1,d-1) = h(n,d).$$

Thus, $s(\mathcal{F}, n) \leq h(n,d)$ which proves (7.56).

To prove (7.57), we use the fact that $n \geq d$ and so:

$$\begin{aligned}\sum_{i=0}^{d} \binom{n}{i} &\leq \left(\frac{n}{d}\right)^d \sum_{i=0}^{d} \binom{n}{i} \left(\frac{d}{n}\right)^i \leq \left(\frac{n}{d}\right)^d \sum_{i=0}^{n} \binom{n}{i} \left(\frac{d}{n}\right)^i \\ &\leq \left(\frac{n}{d}\right)^d \left(1 + \frac{d}{n}\right)^n \leq \left(\frac{n}{d}\right)^d e^d.\end{aligned}$$

□

The VC dimensions of some common examples are summarized in Table 7.1.

Now we can extend the concentration inequalities to hold uniformly over sets of binary functions. We start with finite collections.

**7.58 Theorem.** *Suppose that $\mathcal{F} = \{f_1, \ldots, f_N\}$ is a finite set of binary functions. Then, with probability at least $1 - \delta$,*

$$\sup_{f \in \mathcal{F}} |P_n(f) - P(f)| \leq \sqrt{\frac{2}{n} \log\left(\frac{2N}{\delta}\right)}. \tag{7.59}$$

| Class $\mathcal{A}$ | VC dimension $V_{\mathcal{A}}$ |
|---|---|
| $\mathcal{A} = \{A_1, \ldots, A_N\}$ | $\log_2 N$ |
| Intervals $[a, b]$ on the real line | 2 |
| Discs in $\mathbb{R}^2$ | 3 |
| Closed balls in $\mathbb{R}^d$ | $d + 2$ |
| Rectangles in $\mathbb{R}^d$ | $2d$ |
| Half-spaces in $\mathbb{R}^d$ | $d + 1$ |
| Convex polygons in $\mathbb{R}^2$ | $\infty$ |

**Table 7.1.** *The VC dimension of some classes $\mathcal{A}$.*

***Proof.*** It follows from Hoeffding's inequality that, for each $f \in \mathcal{F}$, $\mathbb{P}\left(|P_n(f) - P(f)| > \epsilon\right) \leq 2e^{-n\epsilon^2/2}$. Hence,

$$\begin{aligned}\mathbb{P}\left(\max_{f\in\mathcal{F}} |P_n(f) - P(f)| > \epsilon\right) &= \mathbb{P}\left(|P_n(f) - P(f)| > \epsilon \text{ for some } f \in \mathcal{F}\right) \\ &\leq \sum_{j=1}^{N} \mathbb{P}\left(|P_n(f_j) - P(f_j)| > \epsilon\right) \leq 2Ne^{-n\epsilon^2/2}.\end{aligned}$$

The conclusion follows. □

Now we consider results for the case where $\mathcal{F}$ is infinite. We begin with an important result due to Vapnik and Chervonenkis.

**7.60 Theorem (Vapnik and Chervonenkis).** *Let $\mathcal{F}$ be a class of binary functions. For any $t > \sqrt{2/n}$,*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |(P_n - P)f| > t\right) \leq 4\, s(\mathcal{F}, 2n)e^{-nt^2/8} \tag{7.61}$$

*and hence, with probability at least $1 - \delta$,*

$$\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| \leq \sqrt{\frac{8}{n}\log\left(\frac{4\, s(\mathcal{F}, 2n)}{\delta}\right)}. \tag{7.62}$$

Before proving the theorem, we need the *symmetrization lemma*. Let $Z'_1, \ldots, Z'_n$ denote a second independent sample from $P$. Let $P'_n$ denote the empirical distribution of this

second sample. The variables $Z_1', \ldots, Z_n'$ are called a *ghost sample*.

**7.63 Lemma (Symmetrization).** *For all* $t > \sqrt{2/n}$,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |(P_n - P)f| > t\right) \leq 2\mathbb{P}\left(\sup_{f\in\mathcal{F}} |(P_n - P_n')f| > t/2\right). \tag{7.64}$$

***Proof.*** Let $f_n \in \mathcal{F}$ maximize $|(P_n - P)f|$. Note that $f_n$ is a random function as it depends on $Z_1, \ldots, Z_n$. We claim that if $|(P_n - P)f_n| > t$ and $|(P - P_n')f_n| \leq t/2$ then $|(P_n' - P_n)f_n| > t/2$. This follows since

$$\begin{aligned} t &< |(P_n - P)f_n| = |(P_n - P_n' + P_n' - P)f_n| \leq |(P_n - P_n')f_n| + |(P_n' - P)f_n| \\ &\leq |(P_n - P_n')f_n| + \frac{t}{2} \end{aligned}$$

and hence $|(P_n' - P_n)f_n| > t/2$. So

$$\begin{aligned} I(|(P_n - P)f_n| > t)\, I(|(P - P_n')f_n| \leq t/2) &= I(|(P_n - P)f_n| > t, |(P - P_n')f_n| \leq t/2) \\ &\leq I(|(P_n' - P_n)f_n| > t/2). \end{aligned}$$

Now take the expected value over $Z_1', \ldots, Z_n'$ and conclude that

$$I(|(P_n - P)f_n| > t)\, \mathbb{P}'(|(P - P_n')f_n| \leq t/2) \leq \mathbb{P}'(|(P_n' - P_n)f_n| > t/2). \tag{7.65}$$

By Chebyshev's inequality,

$$\mathbb{P}'(|(P - P_n')f_n| \leq t/2) \geq 1 - \frac{4\mathsf{Var}'(f_n)}{nt^2} \geq 1 - \frac{1}{nt^2} \geq \frac{1}{2}.$$

(Here we used the fact that $W \in [0,1]$ implies that $\mathsf{Var}(W) \leq 1/4$.) Inserting this into (7.65) we have that

$$I(|(P_n - P)f_n| > t) \leq 2\,\mathbb{P}'(|(P_n' - P_n)f_n| > t/2).$$

Thus,

$$I\Big(\sup_{f\in\mathcal{F}} |(P_n - P)f| > t\Big) \leq 2\,\mathbb{P}'\Big(\sup_{f\in\mathcal{F}} |(P_n' - P_n)f| > t/2\Big).$$

Now take the expectation over $Z_1, \ldots, Z_n$ to conclude that

$$\mathbb{P}\Big(\sup_{f\in\mathcal{F}} |(P_n - P)f| > t\Big) \leq 2\,\mathbb{P}\Big(\sup_{f\in\mathcal{F}} |(P_n' - P_n)f| > t/2\Big).$$

□

The importance of symmetrization is that we have replaced $(P_n - P)f$, which can take any real value, with $(P_n - P_n')f$, which can take only finitely many values. Now we prove the Vapnik-Chervonenkis theorem.

***Proof.*** Let $V = \mathcal{F}_{Z_1',\ldots,Z_n',Z_1,\ldots,Z_n}$. For any $v \in V$ write $(P_n' - P_n)v$ to mean $(1/n)(\sum_{i=1}^{n} v_i - \sum_{i=n+1}^{2n} v_i)$. Using the symmetrization lemma and Hoeffding's inequality,

$$\begin{aligned}
\mathbb{P}(\sup_{f\in\mathcal{F}} |(P_n - P)f| > t) &\leq 2\,\mathbb{P}(\sup_{f\in\mathcal{F}} |(P_n' - P_n)f| > t/2) \\
&= 2\,\mathbb{P}(\max_{v\in V} |(P_n' - P_n)v| > t/2) \\
&\leq 2 \sum_{v\in V} \mathbb{P}(|(P_n' - P_n)v| > t/2) \\
&\leq 2 \sum_{v\in V} 2e^{-nt^2/8} \leq 4\, s(\mathcal{F}, 2n) e^{-nt^2/8}.
\end{aligned}$$

□

Recall that, for a class with finite VC dimension $d$, $s(\mathcal{F}, n) \leq (en/d)^d$. hence we have:

**7.66 Corollary.** *If $\mathcal{F}$ has finite VC dimension $d$, then, with probability at least $1 - \delta$,*

$$\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| \leq \sqrt{\frac{8}{n}\left(\log\left(\frac{4}{\delta}\right) + d\log\left(\frac{ne}{d}\right)\right)}. \tag{7.67}$$

### 7.3.2 Radamacher Complexity

A more general way to develop uniform bounds is to use a quantity called Rademacher complexity. In this section we assume that $\mathcal{F}$ is a class of functions $f$ such that $0 \leq f(z) \leq 1$.

Random variables $\sigma_1, \ldots, \sigma_n$ are called *Rademacher random variables* if they are independent, identically distributed and $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 1/2$. Define the *Rademacher complexity* of $\mathcal{F}$ by

$$\mathsf{Rad}_n(\mathcal{F}) = \mathbb{E}\left(\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i f(Z_i)\right|\right). \tag{7.68}$$

Define the *empirical Rademacher complexity* of $\mathcal{F}$ by

$$\mathsf{Rad}_n(\mathcal{F}, Z^n) = \mathbb{E}_\sigma\left(\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i f(Z_i)\right|\right) \tag{7.69}$$

where $Z^n = (Z_1, \ldots, Z_n)$ and the expectation is over $\sigma$ only.

Intuitively, $\mathsf{Rad}_n(\mathcal{F})$ is large if we can find functions $f \in \mathcal{F}$ that "look like" random noise, that is, they are highly correlated with $\sigma_1, \ldots, \sigma_n$. Here are some properties of the Rademacher complexity.

**7.70 Lemma.**

1. *If* $\mathcal{F} \subset \mathcal{G}$ *then* $\mathsf{Rad}_n(\mathcal{F}, Z^n) \leq \mathsf{Rad}_n(\mathcal{G}, Z^n)$.
2. *Let* $\mathsf{conv}(\mathcal{F})$ *denote the convex hull of* $\mathcal{F}$. *Then* $\mathsf{Rad}_n(\mathcal{F}, Z^n) = \mathsf{Rad}_n(\mathsf{conv}(\mathcal{F}), Z^n)$.
3. *For any* $c \in \mathbb{R}$, $\mathsf{Rad}_n(c\mathcal{F}, Z^n) = |c|\mathsf{Rad}_n(\mathcal{F}, Z^n)$.
4. *Let* $g : \mathbb{R} \to \mathbb{R}$ *be such that* $g(0) = 0$ *and,* $|g(y) - g(x)| \leq L|x - y|$ *for all* $x, y$. *Then* $\mathsf{Rad}_n(g \circ \mathcal{F}, Z^n) \leq 2L\mathsf{Rad}_n(\mathcal{F}, Z^n)$.

***Proof.*** See Exercise 8. □

**7.71 Theorem.** *With probability at least* $1 - \delta$,

$$\sup_{f \in \mathcal{F}} |P_n(f) - P(f)| \leq 2\,\mathsf{Rad}_n(\mathcal{F}) + \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)} \tag{7.72}$$

*and*

$$\sup_{f \in \mathcal{F}} |P_n(f) - P(f)| \leq 2\,\mathsf{Rad}_n(\mathcal{F}, Z^n) + \sqrt{\frac{4}{n}\log\left(\frac{2}{\delta}\right)}. \tag{7.73}$$

***Proof.*** The proof has two steps. First we show that $\sup_{f \in \mathcal{F}} |P_n(f) - P(f)|$ is close to its mean. Then we bound the mean.

**Step 1:** Let $g(Z_1, \ldots, Z_n) = \sup_{f \in \mathcal{F}} |P_n(f) - P(f)|$. If we change $Z_i$ to some other value $Z_i'$ then $|g(Z_1, \ldots, Z_n) - g(Z_1, \ldots, Z_i', \ldots, Z_n)| \leq \frac{1}{n}$. By McDiarmid's inequality,

$$\mathbb{P}\left(|g(Z_1, \ldots, Z_n) - \mathbb{E}[g(Z_1, \ldots, Z_n)]| > \epsilon\right) \leq 2\,e^{-2n\epsilon^2}.$$

Hence, with probability at least $1 - \delta$,

$$g(Z_1, \ldots, Z_n) \leq \mathbb{E}[g(Z_1, \ldots, Z_n)] + \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}. \tag{7.74}$$

**Step 2:** Now we bound $\mathbb{E}[g(Z_1,\ldots,Z_n)]$. Once again we introduce a ghost sample $Z_1',\ldots,Z_n'$ and Rademacher variables $\sigma_1,\ldots,\sigma_n$, Note that $P(f)=\mathbb{E}'P_n'(f)$. Also note that

$$\frac{1}{n}\sum_{i=1}^n(f(Z_i')-f(Z_i)) \stackrel{d}{=} \frac{1}{n}\sum_{i=1}^n \sigma_i(f(Z_i')-f(Z_i))$$

where $\stackrel{d}{=}$ means "equal in distribution." Hence,

$$\begin{aligned}
\mathbb{E}[g(Z_1,\ldots,Z_n)] &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}|P(f)-P_n(f)|\right] = \mathbb{E}\left[\sup_{f\in\mathcal{F}}|\mathbb{E}'(P_n'(f)-P_n(f))|\right] \\
&\le \mathbb{E}\mathbb{E}'\left[\sup_{f\in\mathcal{F}}|P_n'(f)-P_n(f)|\right] = \mathbb{E}\mathbb{E}'\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n(f(Z_i')-f(Z_i))\right|\right] \\
&= \mathbb{E}\mathbb{E}'\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n\sigma_i(f(Z_i')-f(Z_i))\right|\right] \\
&\le \mathbb{E}'\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n\sigma_i\,f(Z_i')\right|\right] + \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n\sigma_i\,f(Z_i)\right|\right] \\
&= 2\mathsf{Rad}_n(\mathcal{F}).
\end{aligned}$$

Combining this bound with (7.74) proves the first result.

To prove the second result, let $a(Z_1,\ldots,Z_n)=\mathsf{Rad}_n(\mathcal{F},Z^n)$ and note that $a(Z_1,\ldots,Z_n)$ changes by at most $1/n$ if we change one observation. McDiarmid's inequality implies that $|\mathsf{Rad}_n(\mathcal{F},Z^n)-\mathsf{Rad}_n(\mathcal{F})|\le\sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}$ with probability at least $1-\delta$. Combining this with the first result yields the second result. □

In the special case where $\mathcal{F}$ is a class of binary functions, we can relate $\mathsf{Rad}_n(\mathcal{F})$ to shattering numbers.

**7.75 Theorem.** *Let $\mathcal{F}$ be a set of binary functions. Then, for all $n$,*

$$\mathsf{Rad}_n(\mathcal{F}) \le \sqrt{\frac{2\log s(\mathcal{F},n)}{n}}. \tag{7.76}$$

***Proof.*** Let $\mathcal{D}=\{Z_1,\ldots,Z_n\}$. Define $S(f,\sigma)=|n^{-1}\sum_{i=1}^n\sigma_i f(Z_i)|$ and $S(v,\sigma)=|n^{-1}\sum_{i=1}^n\sigma_i v_i|$. Now, $-1\le\sigma_i f(Z_i)\le 1$. Note that

$$\mathsf{Rad}_n(\mathcal{F}) = \mathbb{E}\left(\sup_{f\in\mathcal{F}}S(f,\sigma)\right) = \mathbb{E}\left(\mathbb{E}\left(\sup_{f\in\mathcal{F}}S(f,\sigma)\,\middle|\,\mathcal{D}\right)\right) = \mathbb{E}\left(\mathbb{E}\left(\max_{v\in\mathcal{F}_{Z_1,\ldots,Z_n}}S(v,\sigma)\,\middle|\,\mathcal{D}\right)\right).$$

Now, $\sigma_i v_i/n$ has mean 0 and $-1/n\le\sigma_i v_i\le 1/n$ so, by Lemma 7.10, $\mathbb{E}(e^{t\sigma_i v_i})\le e^{t^2/(2n^2)}$ for any $t>0$. From Theorem 7.47,

$$\mathbb{E}\left(\max_{v\in\mathcal{F}_{Z_1,\ldots,Z_n}}S(v,\sigma)\,\middle|\,\mathcal{D}\right)\le\sqrt{\frac{2\log|V_n|}{n}}=\sqrt{\frac{2\log s(\mathcal{F},n)}{n}}$$

and the result follows. □

In fact, there is a sharper relationship between $\mathsf{Rad}_n(\mathcal{F})$ and VC dimension.

**7.77 Theorem.** *Suppose that $\mathcal{F}$ has finite VC dimension $d$. There exists a universal constant $C > 0$ such that $\mathsf{Rad}_n(\mathcal{F}) \leq C\sqrt{d/n}$.*

For a proof, see, for example, Devroye and Lugosi (2001).

Combining these results with Theorem 7.75 and Theorem 7.77 we get the following result.

**7.78 Corollary.** *With probability at least $1 - \delta$,*

$$\sup_{f \in \mathcal{F}} |P_n(f) - P(f)| \leq \sqrt{\frac{8 \log s(\mathcal{F}, n)}{n}} + \sqrt{\frac{1}{2n} \log\left(\frac{2}{\delta}\right)}. \tag{7.79}$$

*If $\mathcal{F}$ has finite VC dimension $d$ then, with probability at least $1 - \delta$,*

$$\sup_{f \in \mathcal{F}} |P_n(f) - P(f)| \leq 2C\sqrt{\frac{d}{n}} + \sqrt{\frac{1}{2n} \log\left(\frac{2}{\delta}\right)}. \tag{7.80}$$

### 7.3.3 Bounds For Classes of Real Valued Functions

Suppose now that $\mathcal{F}$ is a class of real-valued functions. There are various methods to obtain uniform bounds. We consider two such methods: covering numbers and bracketing numbers.

If $Q$ is a measure and $p \geq 1$, define

$$\|f\|_{L_p(Q)} = \left(\int |f(x)|^p dQ(x)\right)^{1/p}.$$

When $Q$ is Lebesgue measure we simply write $\|f\|_p$. We also define

$$\|f\|_\infty = \sup_x |f(x)|.$$

A set $\mathcal{C} = \{f_1, \ldots, f_N\}$ is an *$\epsilon$-cover of $\mathcal{F}$* (or an $\epsilon$-net) if, for every $f \in \mathcal{F}$ there exists a $f_j \in \mathcal{C}$ such that $\|f - f_j\|_{L_p(Q)} < \epsilon$.

**7.81 Definition.** *The size of the smallest $\epsilon$-cover is called the **covering number** and is denoted by $N_p(\epsilon, \mathcal{F}, Q)$. The **uniform covering number** is defined by*

$$N_p(\epsilon, \mathcal{F}) = \sup_Q N_p(\epsilon, \mathcal{F}, Q)$$

*where the supremum is over all probability measures $Q$.*

Now we show how covering numbers can be used to obtain bounds.

**7.82 Theorem.** *Suppose that $\|f\|_\infty \leq B$ for all $f \in \mathcal{F}$. Then,*

$$\mathbb{P}\Big(\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| > \epsilon\Big) \leq 2N(\epsilon/3, \mathcal{F}, L_\infty)e^{-n\epsilon^2/(18B^2)}.$$

***Proof.*** Let $N = N(\epsilon/3, \mathcal{F}, L_\infty)$ and let $C = \{f_1, \ldots, f_N\}$ be an $\epsilon/3$ cover. For any $f \in \mathcal{F}$ there is an $f_j \in C$ such that $\|f - f_j\|_\infty \leq \epsilon/3$. So

$$\begin{aligned} |P_n(f) - P(f)| &\leq |P_n(f) - P_n(f_j)| + |P_n(f_j) - P(f_j)| + |P(f_j) - P(f)| \\ &\leq |P_n(f_j) - P(f_j)| + \frac{2\epsilon}{3}. \end{aligned}$$

Hence,

$$\begin{aligned} \mathbb{P}\left(\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| > \epsilon\right) &\leq \mathbb{P}\left(\max_{f_j\in C} |P_n(f_j) - P(f_j)| + \frac{2\epsilon}{3} > \epsilon\right) \\ &= \mathbb{P}\left(\max_{f_j\in C} |P_n(f_j) - P(f_j)| > \frac{\epsilon}{3}\right) \leq \sum_{j=1}^N \mathbb{P}\left(|P_n(f_j) - P(f_j)| > \frac{\epsilon}{3}\right) \\ &\leq 2N(\epsilon/3, \mathcal{F}, L_\infty)e^{-n\epsilon^2/(18B^2)} \end{aligned}$$

from the union bound and Hoeffding's inequality. □

The VC dimension can be used to bound covering numbers.

**7.83 Theorem.** *Let $\mathcal{F}$ be a class of functions $f : \mathbb{R}^d \to [0, B]$ with $VC$ dimension $d$ such that $2 \leq d < \infty$. Let $p \geq 1$ and $0 < \epsilon < B/4$. Then*

$$N_p(\epsilon, \mathcal{F}) \leq 3\left(\frac{2eB^p}{\epsilon^p} \log\left(\frac{3eB^p}{e^p}\right)\right)^d.$$

(For a proof, see Devroye, Gyorfi and Lugosi (1996).) However, there are cases where the covering numbers are finite and yet the VC dimension is infinite.

**Bracketing Numbers.** Another measure of complexity is the bracketing number. Given a pair of functions $\ell$ and $u$ with $\ell \leq u$, we define the *bracket*

$$[\ell, u] = \Big\{ h : \ \ell(x) \leq h(x) \leq u(x) \text{ for all } x \Big\}.$$

A collection of pairs of functions $(\ell_1, u_1), \ldots, (\ell_N, u_N)$ is a *bracketing* of $\mathcal{F}$ if,

$$\mathcal{F} \subset \bigcup_{j=1}^{B} [\ell_j, u_j].$$

The collection is an $\epsilon$-$L_q(P)$-bracketing if it is a bracketing and if

$$\left( \int |u_j(x) - \ell_j(x)|^q dP(x) \right)^{\frac{1}{q}} \leq \epsilon$$

for $j = 1, \ldots, N$. The *bracketing number* $N_{[]}(\epsilon, \mathcal{F}, L_q(P))$ is the size of the smallest $\epsilon$ bracketing. Bracketing number are a little larger than covering numbers but provide stronger control of the class $\mathcal{F}$.

**7.84 Theorem.**

1. $N_p(\epsilon, \mathcal{F}, P) \leq N_{[]}(2\epsilon, \mathcal{F}, L_p(P))$.
2. *Let* $X_1, \ldots, X_n \sim P$. *If Suppose that* $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ *for all* $\epsilon > 0$. *Then, for every* $\delta > 0$,
$$\mathbb{P}\left( \sup_{f \in \mathcal{F}} |P_n(f) - P(f)| > \delta \right) \to 0 \tag{7.85}$$
*as* $n \to \infty$.

***Proof.*** See exercise 11. □

**7.86 Theorem.** *Let* $A = \sup_f \int |f| dP$ *and* $B = \sup_f \|f\|_\infty$. *Then*

$$P^n\left( \sup_{f \in \mathcal{F}} |P_n(f) - P(f)| > \epsilon \right) \leq 2N_{[]}(\epsilon/8, \mathcal{F}, L_1(P)) \exp\left( -\frac{3n\epsilon^2}{4B[6A + \epsilon]} \right) + 2N_{[]}(\epsilon/8, \mathcal{F}, L_1(P)) \exp\left( -\frac{3n\epsilon}{40B} \right).$$

*Hence, if* $\epsilon \leq 2A/3$,

$$P^n\left( \sup_{f \in \mathcal{F}} |P_n(f) - P(f)| > \epsilon \right) \leq 4N_{[]}(\epsilon/8, \mathcal{F}, L_1(P)) \exp\left( -\frac{96n\epsilon^2}{76AB} \right). \tag{7.87}$$

***Proof.*** (This proof follows Yukich (1985).) For notational simplicity in the proof, let us write, $N(\epsilon) \equiv N_{[]}(\epsilon, \mathcal{F}, L_1(P))$. Define $z_n(f) = \int f(dP_n - dP)$. Let $[\ell_1, u_1], \ldots, [\ell_N, u_N]$ be a minimal $\epsilon/8$ bracketing. We may assume that for each $j$, $\|u_j\| \leq B$ and $\|\ell_j\| \leq B$. (Otherwise, we simply truncate the brackets.) For each $j$, choose some $f_j \in [\ell_j, u_j]$.

Consider any $f \in \mathcal{F}$ and let $[\ell_j, u_j]$ denote a bracket containing $f$. Then

$$|z_n(f)| \leq |z_n(f_j)| + |z_n(f - f_j)|.$$

Furthermore,

$$\begin{aligned}
|z_n(f - f_j)| &= \left| \int (f - f_j)(dP_n - dP) \right| \leq \int |f - f_j| \, (dP_n + dP) \leq \int |u_j - \ell_j| \, (dP_n + dP) \\
&= \int |u_j - \ell_j| \, (dP_n - dP) + 2 \int |u_j - \ell_j| dP \\
&= \int |u_j - \ell_j| \, (dP_n - dP) + 2 \left( \frac{\epsilon}{8} \right) = z_n(|u_j - \ell_j|) + \frac{\epsilon}{4}.
\end{aligned}$$

Hence,

$$|z_n(f)| \leq |z_n(f_j)| + \left[ [z_n(|u_j - \ell_j|) + \frac{\epsilon}{4} \right].$$

Thus,

$$\begin{aligned}
P^n(\sup_{f \in \mathcal{F}} |z_n(f)| > \epsilon) &\leq P^n(\max_j |z_n(f_j)| > \epsilon/2) + P^n(\max_j |z_n(|u_j - \ell_j|)| + \epsilon/4 > \epsilon/2) \\
&\leq P^n(\max_j |z_n(f_j)| > \epsilon/2) + P^n(\max_j |z_n(|u_j - \ell_j|)| > \epsilon/4).
\end{aligned}$$

Now

$$\mathsf{Var}(f_j) \leq \int f_j^2 \, dP = \int |f_j| \, |f_j| dP \leq \|f_j\|_\infty \int |f_j| dP \leq AB.$$

Hence, by Bernstein's inequality,

$$P^n\Big( \max_j |z_n(f_j)| > \epsilon/2 \Big) \leq 2 \sum_{j=1}^N \exp\left( -\frac{1}{2} \frac{n(\epsilon/2)^2}{AB + B\epsilon/6} \right) \leq 2N(\epsilon/8) \exp\left( -\frac{3}{4B} \frac{n\epsilon^2}{6A + \epsilon} \right).$$

Similarly,

$$\begin{aligned}
\mathsf{Var}(|u_j - \ell_j|) &\leq \int (u_j - \ell_j)^2 dP \leq \int |u_j - \ell_j| \, |u_j - \ell_j| dP \\
&\leq \|u_j - \ell_j\|_\infty \int |u_j - \ell_j| dP \leq 2B \frac{\epsilon}{8} = \frac{B\epsilon}{4}.
\end{aligned}$$

Also, $\|u_j - \ell_j\|_\infty \leq 2B$. Hence, by Bernstein's inequality,

$$\begin{aligned}
P^n\Big( \max_j z_n(|u_j - \ell_j|) > \epsilon/4 \Big) &\leq 2 \sum_{j=1}^N \exp\left( -\frac{1}{2} \frac{n(\epsilon/4)^2}{2B\frac{\epsilon}{4} + 2B(\epsilon/4)/3} \right) \\
&\leq 2N(\epsilon/8) \exp\left( -\frac{3n\epsilon}{40B} \right).
\end{aligned}$$

□

The following result is from Example 19.7 from van der Vaart (1998).

**7.88 Lemma.** *Let $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ where $\Theta$ is a bounded subset of $\mathbb{R}^d$. Suppose there exists a function $m$ such that, for every $\theta_1, \theta_2$,*

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq m(x)\,\|\theta_1 - \theta_2\|.$$

*Then,*

$$N_{[]}(\epsilon, \mathcal{F}, L_q(P)) \leq \left(\frac{4\sqrt{d}\,\mathsf{diam}(\Theta)\int |m(x)|^q dP(x)}{\epsilon}\right)^d.$$

***Proof.*** Let

$$\delta = \frac{\epsilon}{4\sqrt{d}\int |m(x)|^q dP(x)}.$$

We can cover $\Theta$ with (at most) $N = (\mathsf{diam}(\Theta)/\delta)^d$ cubes $C_1, \ldots, C_N$ of size $\delta$. Let $c_1, \ldots, c_N$ denote the centers of the cubes. Note that $C_j \subset B(x_j, \sqrt{d}\delta)$ where $B(x, r)$ denotes a ball of radius $r$ centered at $x$. Hence, $\bigcup_j B(c_j, \sqrt{d}\delta)$ covers $\Theta$. Let $\theta_j$ be the projection of $c_j$ onto $\Theta$. Then $\bigcup_j B(\theta_j, 2\delta\sqrt{d})$ covers $\Theta$. In summary, for every $\theta \in \Theta$ there is a $\theta_j \in \{\theta_1, \ldots, \theta_N\}$ such that

$$\|\theta - \theta_j\| \leq 2\delta\sqrt{d} \leq \frac{\epsilon}{2\int |m(x)|^q dP(x)}.$$

Define $\ell_j = f_{\theta_j} - \epsilon m(x)/2$ and $u_j = f_{\theta_j} + \epsilon m(x)/2$. We claim that the brackets $[\ell_1, u_1], \ldots, [\ell_N, u_N]$ cover $\mathcal{F}$. To see this, choose any $f_\theta \in \mathcal{F}$. Let $\theta_j$ be the closest element $\{\theta_1, \ldots, \theta_N\}$ to $\theta$. Then

$$\begin{aligned} f_\theta(x) &= f_{\theta_j}(x) + f_\theta(x) - f_{\theta_j}(x) \leq f_{\theta_j}(x) + |f_\theta(x) - f_{\theta_j}(x)| \\ &\leq f_{\theta_j}(x) + m(x)\|\theta - \theta_j\| \leq f_{\theta_j}(x) + \frac{m(x)\epsilon}{2\int |m(x)|^q dP(x)} = u_j(x). \end{aligned}$$

By a similar argument, $f_\theta(x) \geq \ell_j(x)$. Also, $\int (u_j - \ell_j)^q dP \leq \epsilon^q$. Finally, note that the number of brackets is

$$N = (\mathsf{diam}(\Theta)/\delta)^d = \left(\frac{4\sqrt{d}\,\mathsf{diam}(\Theta)\int |m(x)|^q dP(x)}{\epsilon}\right)^d.$$

□

**7.89 Example (Density Estimation).** Let $X_1, \ldots, X_n \sim P$ where $P$ has support on a compact set $\mathcal{X} \subset \mathbb{R}^d$. Consider the kernel density estimator $\widehat{p}_h(x) = \frac{1}{h^d}\sum_i K(\|x - X_i\|/h)$ where $K$ is a smooth symmetric function and $h > 0$ is a bandwidth. We study $\widehat{p}_h$ in detail in the chapter on nonparametric density estimation. Here we bound the sup norm distance between $\widehat{p}_h(x)$ and is mean $p_h(x) = \mathbb{E}(\widehat{p}_h(x))$.

**7.90 Theorem.** *Suppose that $K(x) \leq K(0)$ for all $x$ and that*

$$|K(y) - K(x)| \leq L\|x - y\|$$

*for all $x, y$. Then*

$$P^n\Big(\sup_x |\widehat{p}(x) - p_h(x)| > \epsilon\Big) \leq 2\left(\frac{32L\sqrt{d}\,\mathsf{diam}(\mathcal{X})}{h^{d+1}\epsilon}\right)^d \left[\exp\left(-\frac{3n\epsilon^2 h^d}{4K(0)(6+\epsilon)}\right) + \exp\left(-\frac{3n\epsilon h^d}{40K(0)}\right)\right].$$

*Hence, if $\epsilon \leq 2/3$ then*

$$P^n\Big(\sup_x |\widehat{p}(x) - p_h(x)| > \epsilon\Big) \leq 4\left(\frac{32L\sqrt{d}\,\mathsf{diam}(\mathcal{X})}{h^{d+1}\epsilon}\right)^d \exp\left(-\frac{3n\epsilon^2 h^d}{28K(0)}\right).$$

***Proof.*** Let $\mathcal{F} = \{f_x : x \in \mathcal{X}\}$ where $f_x(u) = h^{-d}K(\|x-u\|/h)$. We apply Theorem 7.86 with $A = 1$ and $B = K(0)/h^d$. We need to bound $N_{[]}(\epsilon, \mathcal{F}, L_1(P))$. Now

$$\begin{aligned}
|f_x(u) - f_y(u)| &= \frac{1}{h^d}\left|K\left(\frac{\|x-u\|}{h}\right) - K\left(\frac{\|y-u\|}{h}\right)\right| \\
&\leq \frac{L}{h^{d+1}}\Big|\,\|x-u\| - \|y-u\|\,\Big| \\
&\leq \frac{L}{h^{d+1}}\|x-y\|.
\end{aligned}$$

Apply Lemma 7.88 with $m(x) = L/h^{d+1}$. Thus implies that

$$N_{[]}(\epsilon, \mathcal{F}, L_1(P)) \leq \left(\frac{4L\sqrt{d}\,\mathsf{diam}\mathcal{X}}{h^{d+1}\epsilon}\right)^d.$$

Hence, Theorem 7.86 yields,

$$P^n\Big(\sup_x |\widehat{p}(x) - p_h(x)| > \epsilon\Big) \leq 2\left(\frac{32L\sqrt{d}\,\mathsf{diam}(\mathcal{X})}{h^{d+1}\epsilon}\right)^d \left[\exp\left(-\frac{3n\epsilon^2 h^d}{4K(0)(6+\epsilon)}\right) + \exp\left(-\frac{3n\epsilon h^d}{40K(0)}\right)\right].$$

□

**7.91 Corollary.** *Suppose that $h = h_n = (C_n/n)^\xi$ where $\xi \geq 1/d$ and $C_n = (\log n)^a$ for some $a \geq 0$. Then*

$$P^n\Big(\sup_x |\widehat{p}(x) - p_h(x)| > \epsilon\Big) \leq 4\left(\frac{32L\sqrt{d}\,\mathsf{diam}(\mathcal{X})}{\epsilon}\right)^d \left(\frac{n}{C_n}\right)^{\xi(d+1)} \exp\left(-\frac{3\epsilon^2 C_n^{\xi d} n^{1-d\xi}}{28K(0)}\right).$$

*Hence, for sufficiently large $n$,*

$$P^n(\sup_x |\widehat{p}(x) - p_h(x)| > \epsilon) \leq c_1 \exp\left(-c_2\epsilon^2 C_n^{\xi d} n^{1-d\xi}\right).$$

Note that the proofs of the last two results did not depend on $P$. Hence, if $\mathcal{P}$ is the set of distribution with support on $\mathcal{X}$, we have that

$$\sup_{P\in\mathcal{P}} P^n\Big(\sup_x |\widehat{p}(x) - p_h(x)| > \epsilon\Big) \leq c_1 \exp\left(-c_2\epsilon^2 C_n^{\xi d} n^{1-d\xi}\right).$$

□

**7.92 Example.** Here are some further examples. In exercise 12 you are asked to prove these results.

1. Let $\mathcal{F}$ be the set of cdf's on $\mathbb{R}$. Then $N_{[]}(\epsilon, \mathcal{F}, L_2(P)) \leq 2/\epsilon^2$.
2. (Sobolev Spaces.) Let $\mathcal{F}$ be the functions $f$ on $[0,1]$ such that $\|f\|_\infty \leq 1$, the $(k-1)$ derivative is absolutely continuous and $\int (f^{(k)}(x))^2 dx \leq 1$. Then, there is a constant $C > 0$ such that
$$N_{[]}(\epsilon, \mathcal{F}, L_\infty(P)) \leq \exp\left[C\left(\frac{1}{\epsilon}\right)^{\frac{1}{k}}\right].$$
3. Let $\mathcal{F}$ be the set of monotone functions $f$ on $\mathbb{R}$ such that $\|f\|_\infty \leq 1$. Then, there is a constant $C > 0$ such that
$$N_{[]}(\epsilon, \mathcal{F}, L_\infty(P)) \leq \exp\left[C\left(\frac{1}{\epsilon}\right)\right].$$

□

## 7.4 Additional results

### 7.4.1 Talagrand's Inequality

One of the most important developments in concentration of measure is Talagrand's inequality (Talagrand 1994, 1996) which can be thought of as a uniform version of Bernstein's

inequality. Let $\mathcal{F}$ be a class of functions and define $Z_n = \sup_{f\in\mathcal{F}} |P_n(f)|$.

**7.93 Theorem.** *Let $v \geq \mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f^2(X_i)$ and $U \geq \sup_{f\in\mathcal{F}} \|f\|_\infty$. There exists a universal constant $K$ such that*

$$\mathbb{P}\left(\left|\sup_{f\in\mathcal{F}} |P_n(f)| - \mathbb{E}(\sup_{f\in\mathcal{F}} |P_n(f)|)\right| > t\right) \leq K\exp\left\{-\frac{nt}{KU}\log\left(1+\frac{tU}{v}\right)\right\}. \quad (7.94)$$

To make use of Talagrand's inequality, we need to estimate $\mathbb{E}(\sup_{f\in\mathcal{F}} |P_n(f)|)$.

**7.95 Theorem (Giné and Guillou, 2001).** *Suppose that there exist $A$ and $d$ such that*

$$\sup_P N(\epsilon, L_2(P), \epsilon a) \leq (A/\epsilon)^d$$

*where $a = \|F\|_{L_2(P)}$ and $F(x) = \sup_{f\in\mathcal{F}} |f(x)|$. Then*

$$\mathbb{E}(\sup_{f\in\mathcal{F}} |P_n(f)|) \leq C\left(dU\log\left(\frac{AU}{\sigma}\right) + \sqrt{dn}\sigma\sqrt{\log\left(\frac{AU}{\sigma}\right)}\right)$$

*for some $C > 0$.*

Combining these results gives Giné and Guillou's version of Talagrand's inequality:

**7.96 Theorem.** *Let $v \geq \mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f^2(X_i)$ and $U \geq \sup_{f\in\mathcal{F}} \|f\|_\infty$. There exists a universal constant $K$ such that*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| > t\right) \leq K\exp\left\{-\frac{nt}{KU}\log\left(1+\frac{tU}{K(\sqrt{n}\sigma + U\sqrt{\log(AU/\sigma)})^2}\right)\right\} \quad (7.97)$$

*whenever*

$$t \geq \frac{C}{n}\left(U\log\left(\frac{AU}{\sigma}\right) + \sqrt{n}\sigma\sqrt{\log\left(\frac{AU}{\sigma}\right)}\right).$$

**7.98 Example.** Density Estimation. Gine and Guillou (2002) apply Talagrand's inequality to get bounds on density estimators. Let $X_1, \ldots, X_n \sim P$ where $X_i \in \mathbb{R}^d$ and suppose that $P$ has density $p$. The kernel density estimator of $p$ with bandwidth $h$ is

$$\widehat{p}_h(x) = \frac{1}{n}\sum_{i=1}^n K\left(\frac{\|x - X_i\|}{h}\right).$$

Applying the results above to $\widehat{p}_h(x)$ we see that (under very weak conditions on $K$) for all small $\epsilon$ and large $n$,

$$\mathbb{P}(\sup_{x\in\mathbb{R}^d} |\widehat{p}_h(x) - p_h(x)| > \epsilon) \leq c_2 e^{-c_2 n h^d \epsilon^2} \quad (7.99)$$

where $p_h(x) = \mathbb{E}(\widehat{p}_h(x))$ and $c_1, c_2$ are positive constants. This agrees with the earlier result Theorem 7.90. □

### 7.4.2 A Bound on Expected Values

Now we consider bounding the expected value of the maximum of an infinite set of random variables. Let $\{X_f : f \in \mathcal{F}\}$ be a collection of mean 0 random variables indexed by $f \in \mathcal{F}$ and let $d$ be a metric on $\mathcal{F}$. Let $N(\mathcal{F}, r)$ be the covering number of $\mathcal{F}$, that is, the smallest number of balls of radius $r$ required to cover $\mathcal{F}$. Say that $\{X_f : f \in \mathcal{F}\}$ is *sub-Gaussian* if, for every $t > 0$ and every $f, g \in \mathcal{F}$,

$$\mathbb{E}(e^{t(X_f - X_g)}) \leq e^{t^2 d^2(f,g)/2}.$$

We say that $\{X_f : f \in \mathcal{F}\}$ is *sample continuous* if, for every sequence $f_1, f_2, \ldots, \in \mathcal{F}$ such that $d(f_i, f) \to 0$ for some $f \in \mathcal{F}$, we have that $X_{f_i} \to X_f$ a.s. The following theorem is from Cesa-Bianchi and Lugosi (2006) and is a variation of a theorem due to Dudley (1978).

**7.100 Theorem.** *Suppose that $\{X_f : f \in \mathcal{F}\}$ is sub-Gaussian and sample continuous. Then*

$$\mathbb{E}\left(\sup_{f \in \mathcal{F}} X_f\right) \leq 12 \int_0^{D/2} \sqrt{\log N(\mathcal{F}, \epsilon)} d\epsilon \tag{7.101}$$

*where $D = \sup_{f,g \in \mathcal{F}} d(f, g)$.*

***Proof.*** The proof uses Dudley's *chaining technique*. We follow the version in Theorem 8.3 of Cesa-Bianchi and Lugosi (2006). Let $\mathcal{F}_k$ be a minimal cover of $\mathcal{F}$ of radius $D2^{-k}$. Thus $|\mathcal{F}_k| = N(\mathcal{F}, D2^{-k})$. Let $f_0$ denote the unique element in $\mathcal{F}_0$. Each $X_f$ is a random variable and hence is a mapping from some sample space $S$ to the reals. Fix $s \in S$ and let $f^*$ be such that $\sup_{F \in \mathcal{F}} X_f(s) = X_{f^*}(s)$. (If an exact maximizer does not exist, we can choose an approximate maximizer but we shall assume an exact maximizer.) Let $f_k \in \mathcal{F}_k$ minimize the distance to $f_*$. Hence,

$$d(f_{k-1}, f_k) \leq d(f_*, f_k) + d(f_*, f_{k-1}) \leq 3D2^{-k}.$$

Now $\lim_{k \to \infty} f_k = f_*$ and by sample continuity

$$\sup_f X_f(s) = X_{f_*}(s) = X_{f_0}(s) + \sum_{k=1}^{\infty} (X_{f_k}(s) - X_{f_{k-1}}(s)).$$

Recall that $\mathbb{E}(X_{f_0}) = 0$. Therefore,

$$\mathbb{E}\left(\sup_f X_f\right) \leq \sum_{k=1}^{\infty} \mathbb{E}\left(\max_{f,g} (X_f - X_g)\right)$$

where the max is over all $f \in \mathcal{F}_k$ and $g \in \mathcal{F}_{k-1}$ such that $d(f,g) \leq 3D2^{-k}$. There are at most $N(\mathcal{F}, D2^{-k})^2$ such pairs. By Theorem 7.47,

$$\mathbb{E}\left(\max_{f,g}(X_f - X_g)\right) \leq 3D2^{-k}\sqrt{2\log N(\mathcal{F}, D2^{-k})^2}.$$

By summing over $k$ we have

$$\begin{aligned}\mathbb{E}\left(\sup_f X_f\right) &\leq \sum_{k=1}^{\infty} 3D2^{-k}\sqrt{2\log N(\mathcal{F}, D2^{-k})^2} = 12\sum_{k=1}^{\infty} D2^{-(k+1)}\sqrt{\log N(\mathcal{F}, D2^{-k})} \\ &\leq 12\int_0^{D/2}\sqrt{N(\mathcal{F},\epsilon)}d\epsilon.\end{aligned}$$

□

**7.102 Example.** Let $Y_1, \ldots, Y_n$ be a sample from a continuous cdf $F$ on $[0,1]$ with bounded density. Let $X_s = \sqrt{n}(F_n(s) - F(s))$ where $F_n$ is the empirical distribution function. The collection $\{X_s : s \in [0,1]\}$ can be shown to be sub-Gaussian an sample continuous with respect to the Euclidean metric on $[0,1]$. The covering number is $N([0,1], r) = 1/r$. Hence,

$$\mathbb{E}\left(\sup_{0\leq s\leq 1}\sqrt{n}(F_n(s) - F(s))\right) \leq 12\int_0^{1/2}\sqrt{\log(1/\epsilon)}d\epsilon \leq C$$

for some $C > 0$. Hence,

$$\mathbb{E}\left(\sup_{0\leq s\leq 1}(F_n(s) - F(s))\right) \leq \frac{C}{\sqrt{n}}.$$

□

## 7.5 Summary

The most important results in this chapter are Hoeffding's inequality:

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) \leq 2e^{-2n\epsilon^2/c},$$

Bernstein's inequality

$$\mathbb{P}(|\overline{X}_n| > \epsilon) \leq 2\exp\left\{-\frac{n\epsilon^2}{2\sigma^2 + 2c\epsilon/3}\right\}$$

the Vapnik-Chervonenkis bound,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}|(P_n - P)f| > t\right) \leq 4\, s(\mathcal{F}, 2n)e^{-nt^2/8}$$

and the Rademacher bound: with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| \leq 2\,\mathsf{Rad}_n(\mathcal{F}) + \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}.$$

These, and similar results, provide the theoretical basis for many statistical machine learning methods. The literature cantains many refinements and extensions of these results.

## 7.6 Bibliographic Remarks

Concentration of measure is a vast and still growing area. Some good references are Devroye, Gyorfi and Lugosi (1996), van der Vaart and Wellner (1996), Chapter 19 of van der Vaart (1998), Dubhashi and Panconesi (2009), and Ledoux (2005).

## Exercises

7.1 Suppose that $X \geq 0$ and $\mathbb{E}(X) < \infty$. Show that $\mathbb{E}(X) = \int_0^\infty P(X \geq t)dt$.

7.2 Show that $h(u) \geq u^2/(2+2u/3)$ for $u \geq 0$ where $h(u) = (1+u)\log(1+u) - u$.

7.3 In the proof of McDiarmid's inequality, verify that $\mathbb{E}(V_i|X_1,\ldots,X_{i-1}) = 0$.

7.4 Prove Lemma 7.37.

7.5 Prove equation (7.24).

7.6 Prove the results in Table 7.1.

7.7 Derive Hoeffding's inequality from McDiarmid's inequality.

7.8 Prove lemma 7.70.

7.9 Consider Example 7.102. Show that $\{X_s : s \in [0,1]\}$ is sub-Gaussian. Show that $\int_0^{1/2}\sqrt{\log(1/\epsilon)}d\epsilon \leq C$ for some $C > 0$.

7.10 Prove Theorem 7.52. .

7.11 Prove Theorem 7.84.

7.12 Prove the results in Example 7.92.