Concentration of Measure

김지수 (Jisu KIM)

딥러닝의 통계적 이해 (Deep Learning: Statistical Perspective), 2024년 2학기

The lecture note is a minor modification of the lecture notes from Prof. Larry Wasserman's "Statistical Machine Learning".

1 Introduction

Often we need to show that a random quantity $f(Z_1, \ldots, Z_n)$ is close to its mean $\mu(f) = \mathbb{E}(f(Z_1, \ldots, Z_n))$. That is, we want a result of the form

$$\mathbb{P}\Big(\big|f(Z_1, \ldots, Z_n) - \mu(f)\big| > \epsilon\Big) < \delta. \tag{1}$$

Such results are known as *concentration of measure*. These results are fundamental for establishing performance guarantees of many algorithms. In fact, for statistical learning theory, we will need *uniform bounds* of the form

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\Big|f(Z_1, \ldots, Z_n) - \mu(f)\Big| > \epsilon\right) < \delta \tag{2}$$

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

2 Examples

Example (Empirical Risk Minimization for Classification)**.** Consider empirical risk minimization in classification. The data are (X_1, Y_1), ..., (X_n, Y_n) where $Y_i \in \{0, 1\}$ and $X_i \in \mathbb{R}^d$. Given a classifier $h : \mathbb{R}^d \to \{0, 1\}$, the training error is

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

The true classification error is

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

We would like to know if $\hat{R}(h)$ is close to $R(h)$ with high probability. This is precisely of the form (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 $\hat{h}$ minimize the training error $\hat{R}(h)$ over $\mathcal{H}$ and let h_* minimize the true error $R(h)$ over $\mathcal{H}$. We will see in the next chapter that if a uniform inequality like (2) holds then $R(\hat{h})$ is close to $R(h_*)$.

Example (k-means Clustering)**.** The risk of k-means clustering with centers $c = (c_1, \ldots, c_k)$ is

$$R(c) = \mathbb{E}\left[\min_j \|X - c_j\|^2\right]$$

and the empirical risk is

$$\hat{R}(c) = \frac{1}{n}\sum_{i=1}^{n}\left[\min_j \|X_i - c_j\|^2\right].$$

In practice we minimize $\hat{R}(c)$ but we would really like to minimize $R(c)$. To show that minimizing $\hat{R}(c)$ is approximately the same as minimizing $R(c)$ we need to show that

$$\mathbb{P}\left(\sup_c \Big\|\hat{R}(c) - R(c)\Big\| > \epsilon\right)$$

is small.

Example (Cross Validation)**.** Concentration of measure can be used to prove that cross-validation chooses good classifiers and good regression estimators. First consider regression. Let $\mathcal{M} = \{\hat{m}_t : t \in T\}$ be a set of regression estimators depending on the training data and depending on a tuning parameter t. Assume that T is finite. Let $\hat{m}_* \in \mathcal{M}$ minimize $\int |\hat{m}(x) - r(x)|^2 dP(x)$ where $r(x) = \mathbb{E}(Y|X = x)$ is the true regression function. Choose $\hat{t}$ to minimize the hold-out error

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}_t(X_i))^2.$$

Let $\hat{m} = \hat{m}_{\hat{t}}$. Then, we shall see that we have the following result (due to [4]). For any $\delta > 0$,

$$\mathbb{E}\int |\hat{m}(x) - r(x)|^2 dP(x) \leq (1+\delta)\mathbb{E}\int |\hat{m}_*(x) - r(x)|^2 dP(x) + \frac{C(1+\log(|T|)}{n}.$$

A similar result holds holds for classification. Suppose $\mathcal{H}$ is a set of classifiers indexed by t. Let $\hat{h}_*$ be the best classifier in $\mathcal{H}$. Let $\mathcal{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ denote hold-out data from which we estimate the risk by

$$\frac{1}{n}\sum_{i=1}^{n} I(Y_i \neq \hat{h}_t(X_i)).$$

We then have

$$\mathbb{E}\left[P(Y \neq \hat{h}(X))\right] \leq \mathbb{E}\left[P(Y \neq \hat{h}_*(X))\right] + \sqrt{\frac{C\log(|T|)}{n}}.$$

Notice that the rate is better for regression than for classification. This is because of the different loss functions.

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

3 Notation

Notation

If P is a probability measure and f is a function then we define

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

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

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

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

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

4 Basic Inequalities

We begin with two key results: Hoeffding's inequality and McDiarmid's inequality.

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

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

which yields *Markov's inequality*:

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

An immediate consequence is *Chebyshev's inequality*

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

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

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

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 $t > 0$,

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

We then minimize over t and conclude that:

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

Before proceeding, we need the following result.

Lemma 1. *Let Z be a mean 0 random variable such that $a \leq Z \leq b$. Then, for any t,*

$$\mathbb{E}\left[e^{tZ}\right] \leq e^{t^2(b-a)^2/8}.$$

Proof. 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

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

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

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

Hence,

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

□

Theorem 2 (Hoeffding)**.** *If $Z_1, Z_2, \ldots, Z_n$ are independent with $\mathbb{P}(a_i \leq Z_i \leq b_i) = 1$, then for any $t > 0$*

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

where $c = n^{-1}\sum_{i=1}^n (b_i - a_i)^2$ and $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$.

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

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Z_i \geq \epsilon\right) &= \mathbb{P}\left(\frac{t}{n}\sum_{i=1}^n Z_i \geq t\epsilon\right) = \mathbb{P}\left(e^{(t/n)\sum_{i=1}^n Z_i} \geq e^{t\epsilon}\right) \\
&\leq e^{-t\epsilon}\mathbb{E}\left[e^{(t/n)\sum_{i=1}^n Z_i}\right] = e^{-t\epsilon}\prod_i \mathbb{E}\left[e^{(t/n)Z_i}\right] \qquad (3) \\
&\leq e^{-t\epsilon}e^{(t^2/n^2)\sum_{i=1}^n (b_i-a_i)^2/8}
\end{aligned}$$

where the last inequality follows from Lemma 1. 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. □

Corollary 3. *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 μ, then, with probability at least $1-\delta$,*

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

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

Corollary 4. *If $Z_1, Z_2, \ldots, Z_n$ are independent Bernoulii random variables with $\mathbb{P}(Z_i = 1) = p$ then, for any $\epsilon > 0$, $\mathbb{P}\left(|\overline{Z}_n - p| > \epsilon\right) \leq 2e^{-2n\epsilon^2}$ and with probability at least $1-\delta$ we have that $|\overline{Z}_n - p| \leq \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}$.*

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 $|R(h) - \hat{R}_n(h)| \leq \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}$ with probability at least $1-\delta$.

McDiarmid's Inequality. So far we have focused on sums of random variables. The following result extends Hoeffding's inequality to more general functions $f(z_1, \ldots, z_n)$.

Theorem 5 (McDiarmid)**.** *Let $Z_1, \ldots, Z_n$ be independent random variables. Suppose that*

$$\sup_{z_1,\ldots,z_n,z_i'} \left| f(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_n) - f(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n) \right| \leq c_i, \tag{4}$$

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

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

Proof. First we write

$$\begin{aligned}
&\mathbb{P}\left(|f(Z_1, \ldots, Z_n) - \mathbb{E}(f(Z_1, \ldots, Z_n))| \geq \epsilon\right) \\
&= \mathbb{P}\left(f(Z_1, \ldots, Z_n) - \mathbb{E}(f(Z_1, \ldots, Z_n)) \geq \epsilon\right) + \mathbb{P}\left(f(Z_1, \ldots, Z_n) - \mathbb{E}(f(Z_1, \ldots, Z_n)) \leq -\epsilon\right).
\end{aligned}$$

We will show the first inequality. The second follows similarly. Let $V_i = \mathbb{E}(g|Z_1, \ldots, Z_i) - \mathbb{E}(g|Z_1, \ldots, Z_{i-1})$. 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 1, we have

$$\mathbb{E}(e^{tV_i}|Z_1, \ldots, Z_{i-1}) \leq e^{t^2 c_i^2/8}. \tag{5}$$

Now, for any $t > 0$,

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

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

If we take $f(z_1, \ldots, z_n) = n^{-1} \sum_{i=1}^n z_i$ then we get back Hoeffding's inequality. As an example of the application of McDiarmid's inequality, we consider bounding an average of the form $n^{-1} \sum_{i=1}^n \phi(Z_i)$ where ϕ is some, possibly high-dimensional, mapping.

Theorem 6 (Shawe-Taylor and Cristianini)**.** *Suppose that $X \in \mathbb{R}^d$ and let $\phi : \mathbb{R}^d \to \mathcal{F}$ where $\mathcal{F}$ is a Hilbert space. Let $B = \sup_z ||\phi(z)||$ and assume that $B < \infty$. Then*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \phi(Z_i) - \mathbb{E}\left[\phi(Z)\right]\right\| > \frac{B}{\sqrt{n}}\left[2 + \sqrt{2\log\left(\frac{1}{\delta}\right)}\right]\right) < \delta.$$

Proof. Let $S = (Z_1, \ldots, Z_n)$ and let $S' = (Z_1, \ldots, Z_{i-1}, Z_i', Z_{i+1}, \ldots, Z_n)$. Define

$$g(S) = \left\|\frac{1}{n}\sum_{i=1}^n \phi(Z_i) - \mathbb{E}\left[\phi(Z)\right]\right\|.$$

Then $|g(S) - g(S')| \leq 2B/n$, so McDiarmid's inequality implies that

$$\mathbb{P}\left(g(S) - \mathbb{E}(g(S)) > \epsilon\right) \leq \exp\left(-\frac{2n\epsilon^2}{4B^2}\right). \tag{6}$$

It remains to bound $\mathbb{E}(g(S))$. Let $S' = (Z_1', \ldots, Z_n')$ denote a second, independent sample and let $\sigma = (\sigma_1, \ldots, \sigma_n)$ denote independent random variables satisfying $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 1/2$. (The random variables σ_i are called Rademacher variables and will be discussed in more detail in the next section.) Let $\overline{\phi}_S = n^{-1} \sum_{i=1}^n \phi(Z_i)$. Then,

$$\begin{aligned}
\mathbb{E}(g(S)) &= \mathbb{E}\left(\left|\left|\overline{\phi}_S - \mathbb{E}(\phi_S)\right|\right|\right) = \mathbb{E}\left(\left|\left|\overline{\phi}_S - \mathbb{E}(\phi_{S'})\right|\right|\right) \\
&= \mathbb{E}\left(\left|\left|\mathbb{E}(\overline{\phi}_S - (\phi_{S'}))\right|\right|\right) \leq \mathbb{E}\left(\left|\left|\overline{\phi}_S - (\phi_{S'})\right|\right|\right) \\
&= \mathbb{E}\left(\frac{1}{n}\left|\left|\sum_{i=1}^n \sigma_i(\phi(Z_i) - \phi(Z_i'))\right|\right|\right) \leq 2\mathbb{E}\left(\frac{1}{n}\left|\left|\sum_{i=1}^n \sigma_i\phi(Z_i)\right|\right|\right) \\
&= 2\mathbb{E}\left(\frac{1}{n}\sqrt{\sum_{i=1}^n \sigma_i^2\phi^2(Z_i) + \sum_{i\neq j}\sigma_i\sigma_j\phi(Z_i)\phi(Z_j)}\right) \\
&\leq \frac{2}{n}\sqrt{\mathbb{E}\left(\sum_{i=1}^n \sigma_i^2\phi^2(Z_i) + \sum_{i\neq j}\sigma_i\sigma_j\phi(Z_i)\phi(Z_j)\right)} \\
&= \frac{2}{n}\sqrt{\sum_{i=1}^n \mathbb{E}(\phi^2(Z_i))} \leq \frac{2B}{\sqrt{n}}.
\end{aligned}$$

The result follows by combining this with (6) and setting $\epsilon = B\sqrt{\frac{2}{n}\log\left(\frac{1}{\delta}\right)}$. □

As another example, suppose that $Z_1, \ldots, Z_n$ are real-valued random variables with cdf F. Let $F_n(z) = n^{-1} \sum_{i=1}^n I(Z_i \leq x)$ be the empirical cdf. Let $f(Z_1, \ldots, Z_n) = \sup_z |F_n(z) - F(z)|$. If we change on Z_i then f changes by at most $1/n$. Hence,

$$\mathbb{P}\left(\sup_z |F_n(z) - F(z)| - \mathbb{E}(\sup_z |F_n(z) - F(z)|) > \epsilon\right) \leq e^{-2n\epsilon^2}. \tag{7}$$

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

$$\mathbb{P}(X > \epsilon) = \int_\epsilon^\infty \phi(s)ds \leq \frac{1}{\epsilon}\int_\epsilon^\infty s\phi(s)ds = -\frac{1}{\epsilon}\int_\epsilon^\infty \phi'(s)ds = \frac{\phi(\epsilon)}{\epsilon} \leq \frac{1}{\epsilon}e^{-\epsilon^2/2}. \tag{8}$$

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,

$$\mathbb{P}(|\overline{X}_n - \mu| > \epsilon) = \mathbb{P}\Big(\sqrt{n}|\overline{X}_n - \mu|/\sigma > \sqrt{n}\epsilon/\sigma\Big) = \mathbb{P}\Big(|Z| > \sqrt{n}\epsilon/\sigma\Big) \tag{9}$$

$$\leq \frac{2\sigma}{\epsilon\sqrt{n}} e^{-n\epsilon^2/(2\sigma^2)} \leq e^{-n\epsilon^2/(2\sigma^2)} \tag{10}$$

for all large n. This bound is very powerful because the probability on the right hand side goes to 0 exponentially fast as the sample size n increases.

Bernstein's Inequality. 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.

Lemma 7. *Suppose that $|X| \leq c$ and $\mathbb{E}(X) = 0$. For any $t > 0$,*

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

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,

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

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

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

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

Theorem 8 (Bernstein)**.** *If $\mathbb{P}(|X_i| \leq c) = 1$ and $\mathbb{E}(X_i) = 0$ then, for any $t > 0$,*

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

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

Proof. From Lemma 7,

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

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

$$\begin{aligned}\mathbb{P}\big(\overline{X}_n > \epsilon\big) &= \mathbb{P}\left(\sum_{i=1}^n X_i > n\epsilon\right) = \mathbb{P}\Big(e^{t\sum_{i=1}^n X_i} > e^{tn\epsilon}\Big) \\ &\leq e^{-tn\epsilon}\mathbb{E}(e^{t\sum_{i=1}^n X_i}) = e^{-tn\epsilon}\prod_{i=1}^n \mathbb{E}(e^{tX_i}) \\ &\leq e^{-tn\epsilon}\exp\left\{nt^2\sigma^2\frac{e^{tc} - 1 - tc}{(tc)^2}\right\}.\end{aligned}$$

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

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

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.

Lemma 9. *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 $1 - \delta$,*

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

If σ^2 is very small than the first term on the right hand side becomes negligible resulting in a very tight bound.

5 Measures of Complexity

To develop uniform bounds we need to introduce some complexity measures. More specifically, given a class of functions $\mathcal{F}$, we need some way to measure how complex the class $\mathcal{F}$ is. If $\mathcal{F} = \{f_1, \ldots, f_N\}$ is finite then an obvious measure of complexity is the size of the set, N. The more challenging case is when $\mathcal{F}$ is infinite.

Rademacher Complexity. Our first measure is rather subtle but is extremely important: the Rademacher complexity.

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

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

Some authors use a slightly different definition, namely,

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

You can use either one. They lead to essentially the same results.

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.

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 $\mathsf{Rad}_n(g \circ \mathcal{F}, Z^n) \leq 2L\mathsf{Rad}_n(\mathcal{F}, Z^n)$.

The Rademacher complexity arises naturally in many proofs. But it is hard to compute and so it is common to replace the Rademacher complexity with an upper bound. This leads us to shattering numbers.

Shattering Numbers. Let $\mathcal{Z}$ be a set and let $\mathcal{F}$ is a class of binary functions on $\mathcal{Z}$. Thus, each $f \in \mathcal{F}$ maps $\mathcal{Z}$ to $\{0, 1\}$. For any $z_1, \ldots, z_n$ define

$$\mathcal{F}_{z_1,\ldots,z_n} = \Big\{(f(z_1), \ldots, f(z_n)) : \ f \in \mathcal{F}\Big\}. \tag{11}$$

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

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

$$s(\mathcal{F}, n) = \sup_{z_1,\ldots,z_n} \left|\mathcal{F}_{z_1,\ldots,z_n}\right|. \tag{12}$$

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 $s(\mathcal{A}, F) = 2^n$.

Theorem 10. *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. Exercise. □

Theorem 11. *Let $\mathcal{F}$ be a set of binary functions. Then, for all n,*

$$\mathsf{Rad}_n(\mathcal{F}) \leq \sqrt{\frac{2 \log s(\mathcal{F}, n)}{n}}. \tag{13}$$

Proof. Let s_n be the shattering number of $\mathcal{F}$. Recall that, If Z has mean 0 and $a \leq Z \leq b$ then $\mathbb{E}[e^{tZ}] \leq e^{t^2(b-a)^2/8}$. We have

$$\mathrm{Rad}_n(\mathcal{F}) = \mathbb{E}\left(\sup_f \frac{1}{n}\sum_i \sigma_i f(Z_i)\right) = \mathbb{E}\mathbb{E}\sup_f \left[\frac{1}{n}\sum_i \sigma_i f(Z_i) \ \middle| \ Z_1, \ldots, Z_n\right] = \mathbb{E}Q,$$

where

$$Q = \mathbb{E}\max_{v \in V}\left[\frac{1}{n}\sum_i \sigma_i v_i | Z_1, \ldots, Z_n\right],$$

where $v = (v_1, \ldots, v_n)$ and $v_i = f(Z_i)$. The vector v varies over $V = ((f(Z_1), \ldots, f(Z_n)) : \in \mathcal{F})$.

Now

$$\begin{aligned}
e^{tQ} &= \exp\left(t\mathbb{E}\max_v \left[\frac{1}{n}\sum_i \sigma_i v_i | Z_1, \ldots, Z_n\right]\right) \\
&\leq \mathbb{E}\left(\exp\left(t\max_v\left[\frac{1}{n}\sum_i \sigma_i v_i\right]\right) \ \middle| \ Z_1, \ldots, Z_n\right) \\
&= \mathbb{E}\left(\max_v \exp\left(\frac{t}{n}\frac{1}{n}\sum_i \sigma_i v_i\right) \ \middle| \ Z_1, \ldots, Z_n\right) \\
&\leq \sum_v \mathbb{E}\left(\exp\left(\frac{t}{n}\sum_i \sigma_i v_i\right) \ \middle| \ Z_1, \ldots, Z_n\right) \\
&= \sum_v \prod_i \mathbb{E}\left(e^{t\sigma_i v_i/n} | Z_1, \ldots, Z_n\right) \\
&= \sum_v \prod_i e^{t^2/(2n)} = s_n e^{t^2/(2n)}.
\end{aligned}$$

In the last step, we used the fact that, given $Z_1, \ldots, Z_n$, $\sigma_i v_i$ has mean 0 and $-1/n \leq \sigma_i v_i \leq 1/n$ and then we applied the Lemma above. Taking the log of both side gives

$$tQ \leq \log(s_n) + \frac{t^2}{2n},$$

and so

$$Q \leq \frac{\log s_n}{t} + \frac{t}{2n}.$$

Hence,

$$\mathrm{Rad}_n(\mathcal{F}) \leq \frac{\log s_n}{t} + \frac{t}{2n}.$$

Let $t = \sqrt{2n \log s_n}$. Then we get

$$\mathrm{Rad}_n(\mathcal{F}) \leq \sqrt{\frac{2 \log s_n}{n}}.$$

□

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

$$\mathsf{VC}(\mathcal{A}) = \sup\Big\{n : \ s(\mathcal{A}, n) = 2^n\Big\}. \tag{14}$$

Similarly, the *VC dimension* of a class of binary functions $\mathcal{F}$ is

$$\mathsf{VC}(\mathcal{F}) = \sup\Big\{n : \ s(\mathcal{F}, n) = 2^n\Big\}. \tag{15}$$

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.

Theorem 12. *Suppose that $\mathcal{F}$ has finite VC dimension d. Then,*

$$s(\mathcal{F}, n) \leq \sum_{i=0}^{d} \binom{n}{i}, \tag{16}$$

and for all $n \geq d$,

$$s(\mathcal{F}, n) \leq \Big(\frac{en}{d}\Big)^d. \tag{17}$$

Proof. When $n = d = 1$, (16) clearly holds. We show that now proceed by induction. Suppose that (16) 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 (16).

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

Table 1: The VC dimension of some classes $\mathcal{A}$.

To prove (17), 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 1.

Theorem 13. *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, [3].

6 Uniform Bounds

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

Theorem 14. *Suppose that $\mathcal{F} = \{f_1, \ldots, f_N\}$ is a finite set of bounded functions. Then, with probability at least $1 - \delta$,*

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

where $c = 4 \max_j \|f_j\|_\infty^2$.

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

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

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.

Theorem 15 (Vapnik and Chervonenkis)**.** *Let $\mathcal{F}$ be a class of binary functions. For any $t > \sqrt{2/n}$,*

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

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

$$\sup_{f \in \mathcal{F}} |P_n(f) - P(f)| \leq \sqrt{\frac{8}{n} \log\left(\frac{4s(\mathcal{F}, 2n)}{\delta}\right)}. \tag{18}$$

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*.

Lemma 16 (Symmetrization)**.** *For all $t > \sqrt{2/n}$,*

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

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

$$t < |(P_n - P)f_n| = |(P_n - P'_n + P'_n - P)f_n| \leq |(P_n - P'_n)f_n| + |(P'_n - P)f_n| \leq |(P_n - P'_n)f_n| + \frac{t}{2}$$

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

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

By Chebyshev's inequality,

$$\mathbb{P}'(|(P - P'_n)f_n| > t/2) \leq \frac{4Var'(f_n)}{nt^2} \leq \frac{1}{nt^2} < \frac{1}{2}.$$

(Here we used the fact that $W \in \{0, 1\}$ implies that $Var(W) \leq 1/4$.) So

$$\mathbb{P}'(|(P - P'_n)f_n| \leq t/2) \geq \frac{1}{2}.$$

Thus,

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

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,

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

□

Recall that, for a class with finite VC dimension d, $s(\mathcal{F}, n) \le (en/d)^d$. In this case, (18) implies that

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

Now we obtain uniform bounds using Rademacher complexity. For this, we bound the symmetrization by Rademacher complexity.

Lemma 17 (Symmetrization)**.**

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |(P_n - P)f|\right] \le 2\mathsf{Rad}_n(\mathcal{F}).$$

Proof. 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,

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

□

Theorem 18. *With probability at least* $1 - \delta$*,*

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

and

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

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.
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 $|g(Z_1, \ldots, Z_n) - g(Z_1, \ldots, Z'_i, \ldots, Z_n)| \le \frac{1}{n}$. By McDiarmid's inequality,

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

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

$$g(Z_1, \ldots, Z_n) \le \mathbb{E}[g(Z_1, \ldots, Z_n)] + \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}. \tag{19}$$

Then the mean $\mathbb{E}[g(Z_1, \ldots, Z_n)]$ is bounded by Lemma 17 as

$$\mathbb{E}[g(Z_1, \ldots, Z_n)] = \mathbb{E}\left[\sup_{f\in\mathcal{F}}(P(f) - P_n(f))\right] \le 2\mathsf{Rad}_n(\mathcal{F}).$$

Combining this bound with (19) 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 $|\mathsf{Rad}_n(\mathcal{F}, Z^n) - \mathsf{Rad}_n(\mathcal{F})| \leq \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}$ with probability at least $1-\delta$. Combining this with the first result yields the second result. □

Combining this theorem some Theorem 11 and Theorem 13 we get the following result.

Corollary 19. *With probability at least* $1-\delta$,

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

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

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

7 Example: Classification

Let $\mathcal{H}$ be a set of classifiers with finite VC dimension d. The optimal classifier $h_* \in \mathcal{H}$ minimizes

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

The empirical risk minimizer is the classifier $\hat{h}$ that minimizes

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

From the VC theorem, with high probability,

$$\sup_{h\in\mathcal{H}} |\hat{R}(h) - R(h)| \preceq \sqrt{\frac{d\log n}{n}}.$$

Hence, with high probability

$$R(\hat{h}) \preceq \hat{R}(\hat{h}) + \sqrt{\frac{d\log n}{n}} \preceq \hat{R}(h_*) + \sqrt{\frac{d\log n}{n}} \preceq R(h_*) + \sqrt{\frac{d\log n}{n}}.$$

So empirical risk minimization comes close to the best risk in the class if the VC dimension is finite.

8 Example: k-means Clustering

We can use concentration of measure to prove some properties of k-means clustering. Let $C = \{c_1, \ldots, c_k\}$ and define the risk $R(C) = \mathbb{E}\,\|X - \Pi_C[X]\|^2$ where $\Pi_C[x] = \text{argmin}_{c_j} \|\|x - c_j\|^2$. Let $C^* = \{c_1^*, \ldots, c_k^*\}$ be a minimizer of $R(C)$.

Theorem 20. *Suppose that* $\mathbb{P}(||X_i||^2 \leq B) = 1$ *for some* $B < \infty$. *Then*

$$\mathbb{E}(R(\hat{C})) - R(C^*) \leq c\sqrt{\frac{k(d+1)\log n}{n}},$$

for some $c > 0$.

Warning! The fact that $R(\hat{C})$ is close to $R(C_*)$ does not imply that $\hat{C}$ is close to C_*.

This proof is due to [5].

Proof. Note that $R(\hat{C}) - R(C^*) = R(\hat{C}) - R_n(\hat{C}) + R_n(\hat{C}) - R(C^*) \leq R(\hat{C}) - R_n(\hat{C}) + R_n(C^*) - R(C^*) \leq 2\sup_{C\in\mathcal{C}_k} |R(\hat{C}) - R_n(\hat{C})|$. For each C define a function f_C by $f_C(x) = ||x - \Pi_C[x]||^2$. Note that $\sup_x |f_C(x)| \leq 4B$ for all C. Now, using the fact that $\mathbb{E}(Y) = \int_0^\infty \mathbb{P}(Y \geq t)dt$ whenever $Y \geq 0$, we have

$$\begin{aligned}
2\sup_{C\in\mathcal{C}_k} |R(\hat{C}) - R_n(\hat{C})| &= 2\sup_C \left| \frac{1}{n}\sum_{i=1}^n (f_C(X_i) - \mathbb{E}(f_C(X))) \right| \\
&= 2\sup_C \left| \int_0^{4B} \left(\frac{1}{n}\sum_{i=1}^n I(f_C(X_i) > u) - \mathbb{P}(f_C(Z) > u) \right) du \right| \\
&\leq 8B \sup_{C,u} \left| \frac{1}{n}\sum_{i=1}^n I(f_C(X_i) > u) - \mathbb{P}(f_C(Z) > u) \right| \\
&= 8B \sup_A \left| \frac{1}{n}\sum_{i=1}^n I(X_i \in A) - \mathbb{P}(A) \right|,
\end{aligned}$$

where A varies over all sets $\mathcal{A}$ of the form $\{f_C(x) > u\}$. The shattering number of $\mathcal{A}$ is $s(\mathcal{A}, n) \leq n^{k(d+1)}$. This follows since each set $\{f_C(x) > u\}$ is a union of the complements of k spheres. By Theorem 11,

$$\begin{aligned}
\mathbb{P}(R(\hat{C}) - R(C^*) > \epsilon) &\leq \mathbb{P}\left(8B \sup_A \left| \frac{1}{n}\sum_{i=1}^n I(X_i \in A) - \mathbb{P}(A) \right| > \epsilon \right) \\
&= \mathbb{P}\left(\sup_A \left| \frac{1}{n}\sum_{i=1}^n I(X_i \in A) - \mathbb{P}(A) \right| > \frac{\epsilon}{8B} \right) \\
&\leq 4(2n)^{k(d+1)} e^{-n\epsilon^2/(512B^2)}.
\end{aligned}$$

Now apply Theorem 24 to conclude that $\mathbb{E}(R(\hat{C}) - R(C^*)) \leq C\sqrt{k(d+1)}\sqrt{\frac{\log n}{n}}$. □

A sharper result, together with a lower bound is the following.

Theorem 21 ([1]). *Suppose that* $\mathbb{P}\left(\|X\|^2 \leq 1\right) = 1$ *and that* $n \geq k^{4/d}$, $\sqrt{dk^{1-2/d}\log n} \geq 15$, $kd \geq 8$, $n \geq 8d$ *and* $n/\log n \geq dk^{1+2/d}$. *Then,*

$$\mathbb{E}(R(\hat{C})) - R(C^*) \leq 32\sqrt{\frac{dk^{1-2/d}\log n}{n}} = O\left(\sqrt{\frac{dk\log n}{n}}\right).$$

Also, if $k \geq 3$, $n \geq 16k/(2\Phi^2(-2))$ *then, for any method* $\hat{C}$ *that selects* k *centers, there exists* P *such that*

$$\mathbb{E}(R(\hat{C})) - R(C^*) \geq c_0\sqrt{\frac{k^{1-4/d}}{n}}$$

where $c_0 = \Phi^4(-2)2^{-12}/\sqrt{6}$ *and* Φ *is the standard Gaussian distribution function.*

See [1] for a proof. It follows that k-means is risk consistent in the sense that $R(\hat{C}) - R(C^*) \xrightarrow{P} 0$, as long as $k = o(n/(d^3\log n))$. Moreover, the lower bound implies that we cannot find any other method that improves much over the k-means approach, at least with respect to this loss function.

The previous results depend on the dimension d. It is possible to get a dimension-free result at the expense of replacing $\sqrt{k}$ with k. In fact, the following result even applies to functions instead of vectors. In that case, we interpret $||\cdot||$ to be the norm in a Hilbert space. The proof uses Radaemacher variables instead of VC arguments.

Theorem 22 ([2]). *Suppose that* $\mathbb{P}(||X_i|| \leq B) = 1$. *Then*

$$\mathbb{E}(R(\hat{C})) - R(C^*) \leq \frac{12B^2k}{\sqrt{n}}.$$

Proof. Define $W(C, P) = \mathbb{E}_P\left(\min_{1\leq j\leq k}\left[-2\langle X, c_j\rangle + ||c_j||^2\right]\right)$. Minimizing $R(C)$ is equivalent to minimizing $W(C, P)$ and minimizing $R_n(C)$ is equivalent to minimizing $W(C, P_n)$ where P_n is the empirical measure that puts mass $1/n$ at each X_i. Arguing as in the proof of Theorem 20,

$$\mathbb{E}(W(\hat{C}, P)) - W(C^*, P) \leq 2\mathbb{E}\Big(\sup_C W(C, P) - W(C, P_n)\Big).$$

Let $\sigma_1, \ldots, \sigma_n$ be Rademacher random variables. That is, $\sigma_1, \ldots, \sigma_n$ are iid and $\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = 1/2$. Let $X_1', \ldots, X_n'$ be a second independent sample. Let $\ell_c(x) = -2\langle x, c\rangle + \|c\|^2$. Then,

$$\begin{aligned}
\mathbb{E}\Big(\sup_C W(C,P) - W(C,P_n)\Big) &\leq \mathbb{E}\left(\sup_C \frac{1}{n}\sum_{i=1}^n \sigma_i \left[\min_{1\leq j\leq n} \ell_{c_j}(X_i) - \min_{1\leq j\leq n} \ell_{c_j}(X_i')\right]\right) \\
&\leq \mathbb{E}\left(\sup_C \frac{1}{n}\sum_{i=1}^n \sigma_i \left[\min_{1\leq j\leq n} \ell_{c_j}(X_i)\right]\right) \\
&\quad + \mathbb{E}\left(\sup_C \frac{1}{n}\sum_{i=1}^n (-\sigma_i) \left[\min_{1\leq j\leq n} \ell_{c_j}(X_i)\right]\right) \\
&= 2\mathbb{E}\left(\sup_C \frac{1}{n}\sum_{i=1}^n \sigma_i \left[\min_{1\leq j\leq n} \ell_{c_j}(X_i)\right]\right).
\end{aligned}$$

An inductive argument shows that

$$2\mathbb{E}\left(\sup_C \frac{1}{n}\sum_{i=1}^n \sigma_i \left[\min_{1\leq j\leq n} \ell_{c_j}(X_i)\right]\right) \leq 4k\left[\mathbb{E}\sup_{c\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \sigma_i\langle X_i, c\rangle + \frac{B^2}{2\sqrt{n}}\right]$$

Also,

$$\begin{aligned}
\mathbb{E}\left(\sup_{c\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \sigma_i\langle X_i, c\rangle\right) &= \mathbb{E}\left(\sup_{c\in\mathbb{R}^d} \frac{1}{n}\left\langle \sum_{i=1}^n \sigma_i X_i, c\right\rangle\right) = \frac{B}{n}\mathbb{E}\left|\left|\sum_{i=1}^n \sigma_i X_i\right|\right| \\
&\leq \frac{B}{n}\sqrt{\mathbb{E}\left|\left|\sum_{i=1}^n \sigma_i X_i\right|\right|^2} = B\sqrt{\frac{\mathbb{E}||X||^2}{n}} \leq \frac{B^2}{\sqrt{n}}.
\end{aligned}$$

□

9 Example: Density Estimation

Let $X_1, \ldots, X_n \sim P$ where P has density p and $X_i \in \mathbb{R}^d$. Let

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

be the kernel density estimator. Let $p_h(x) = \mathbb{E}[\hat{p}_h(x)]$. Then

$$\|\hat{p}_h - p\|_\infty \leq \|\hat{p}_h - p_h\|_\infty + \|p_h - p\|_\infty.$$

The second term, the bias, is bounded by Ch^2 under standard smoothness conditions. What above the first term? Let us first focus on a single point x.

Theorem 23. *Suppose that* $(\log n/n)^{1/d} \leq h \leq C$ *for some* $C > 0$. *Then*

$$\mathbb{P}(|\hat{p}_h(x) - p_h(x)| > \epsilon) \leq c_1 e^{-nc_2\epsilon^2 h^d}.$$

Proof. This can be proved by Bernstein's inequality. We leave the proof as an exercise. If you use Hoeffding's inequality you will not get a sharp bound. □

The more general result is the following.

Theorem. *Suppose that* P *has compact support and that* $(\log n/n)^{1/d} \leq h \leq C$ *for some* $C > 0$. *Under weak conditions on* K, *we have*

$$\mathbb{P}\left(\|\hat{p}_h - p_h\|_\infty > \epsilon\right) \leq c_1 e^{-nc_2\epsilon^2 h^d}.$$

The proof is essentially the same except that one has to replace Bernstein's inequality by Talagrand's inequality. You can think of Talagrand's inequality as an extension of Benstein's inequality over infinite sets of functions. The theorm can also be proved used a bracketing argument combined with Bernstein's inequality. (Bracketing is discussed in a later section.)

10 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.

Theorem 24. *Suppose that $X_n \geq 0$ and that for every $\epsilon > 0$,*

$$\mathbb{P}(X_n > \epsilon) \leq c_1 e^{-c_2 n \epsilon^2}, \tag{20}$$

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

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

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.

Theorem 25. *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*

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

Proof. By Jensen's inequality,

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

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 $t = \sqrt{2 \log n}/\sigma$. □

11 Covering Numbers and Bracketing Numbers

Often the VC dimension is infinite. In such cases we need other measures of complexity.

If Q is a measure and $p \geq 1$ we 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 *ϵ-cover of $\mathcal{F}$* if, for every $f \in \mathcal{F}$ there exists a $f_j \in \mathcal{C}$ such that $\|f - f_j\|_{L_p(Q)} < \epsilon$.

Definition. The size of the smallest ϵ-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.

Theorem 26. *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 such that $B_\infty(f_1, \epsilon/3), \ldots, B_\infty(f_N, \epsilon/3)$ is 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,

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

from the union bound and Hoeffding's inequality. □

When the VC is finite, it can be used to bound covering numbers.

Theorem 27. *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*

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

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. A collection of pairs of functions $(\ell_1, u_1), \ldots, (\ell_N, u_N)$ is a *bracketing* of $\mathcal{F}$ if, for each $f \in \mathcal{F}$ there exists a pair (ℓ_j, u_j) such that $\ell_j(x) \leq f(x) \leq u_j(x)$ for all x. The collection is an ϵ-bracketing if it is a bracketing and $(\int |u_j(x) - \ell_j(x)|^p dQ(x))^{1/p} \leq \epsilon$ for $j = 1, \ldots, N$. The *bracketing number* $N_{[]}(\epsilon, \mathcal{F}, Q, p)$ is the size of the smallest ϵ bracketing. Bracketing number are a little larger than covering numbers but provide stronger control of the class $\mathcal{F}$.

Theorem 28. *1.* $N_p(\epsilon, \mathcal{F}, P) \leq N_{[]}(2\epsilon, \mathcal{F}, P, p)$.

2. There are positive constants c_1, c_2, c_3 *such that, for any* $\epsilon > 0$,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| > \epsilon\right) \leq c_1\, N_{[]}(c_2\epsilon, \mathcal{F}, P, 1) e^{-nc_3\epsilon^2}.$$

3. Let $X_1, \ldots, X_n \sim P$. *If Suppose that* $N_{[]}(\epsilon, \mathcal{F}, P, 1) < \infty$ *for all* $\epsilon > 0$. *Then, for every* $\delta > 0$,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| > \delta\right) \to 0,$$

as $n \to \infty$.

12 Summary

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

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

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.

References

[1] Peter L. Bartlett, Tamás Linder, and Gábor Lugosi. The minimax distortion redundancy in empirical quantizer design. *IEEE Trans. Inform. Theory*, 44(5):1802–1813, 1998.

[2] Gérard Biau, Luc Devroye, and Gábor Lugosi. Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.*, 9:2015–2033, 2008.

[3] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer-Verlag, New York, 2001.

[4] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

[5] Tamás Linder, Gábor Lugosi, and Kenneth Zeger. Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Trans. Inform. Theory*, 40(6):1728–1740, 1994.