Measure concentration and non-asymptotic statistics

François Portier

November 6, 2025

A Bernstein-type (with the variance) deviation inequality for suprema of empirical processes is established based on (i) Bousquet's concentration inequality, (ii) Symmetrization, and (iii) Dudley's entropy bound on the so called Rademacher complexity. The final bound is valid for general classes, called VC classes, that have a reasonably large covering number (highlighting a certain control on the complexity of the classes). This approach turns out to be useful because many functions of statistical interest have small covering number. In particular, Hausler's lemma and Vapnik's dimension will be key concepts to derive such properties. Local averaging estimate are studied and some uniform error bound are derived. Finally, empirical risk minimization with "fast rates" is studied.

Contents

1 Motivation

The aim of the course is to obtain probability bound, i.e., $\mathbb{P}(|Z| > t) \leq \delta$, for certain quantity of interest Z that results from Machine learning algorithm or statistical methods. One first estimator that will be study in this lecture notes is the kernel density estimator (KDE) which is now introduced under the following framework. Let $(X, X_1, \ldots, X_n)$ be a sequence of independent and identically distributed random elements with common distribution P_X on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. We suppose that P_X has a density with respect to Lebesgue measure, i.e., for all $B \in \mathcal{B}(\mathbb{R}^d)$, $P_X(B) = \int_B f_X \mathrm{d}\lambda$ where f_X is called the density function of X. The KDE density estimate is given by

$$\hat{f}_n(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i), \qquad x \in \mathbb{R}^d,$$

where K is a probability density function and K_h is the rescaled version so that it has standard deviation $h I_d$. Now assuming that f_X is L_X-Lipschitz, one might show easily that

$$|\mathbb{E}[\hat{f}_n(x)] - f_X(x)| \leq h L_X$$

and, for h small enough,

$$\mathrm{var}(\hat{f}_n(x)) \leq \frac{2f(x)}{nh^d}.$$

This advocates for an error decomposition of the form

$$|\hat{f}_n(x) - f(x)| \leq h L_X + \sqrt{\frac{2f(x)}{nh^d} \log(1/\delta)}$$

The previous holding true with probability $1 - \delta$. This is left as an exercise in the next section, where suitable inequalities will be given. The previous upper bound is non-asymptotic, meaning that it is valid for all $n \geq 1$ which, compared to an asymptotic bound, is more relevant for applications because the upper bound is more precise. However, the bound is only valid at a given x and we have no idea of how the upper-bound evolve when the error is measured over an interval or the full domain, such as

$$\sup_{x \in \mathbb{R}} |\hat{f}_n(x) - f(x)|.$$

The tools developed in this note will allow to obtain such a bound. Application of interest includes Nadaraya-Watson estimator which is a kernel based estimator of the regression function. Another estimator of interest is the popular k-NN regression or classification rule as well as regression tree estimator. The last estimator are well-known instances of a broad framework known as local averaging rule because they all use an averaging over small parts of the data (localized in space) to make the prediction. Other types of estimator that will be covered include empirical risk minimizer (ERM). In that case, one wishes to show that $P\ell(Z, \hat{h}_n) \leq \min_{h \in \mathcal{H}} P\ell(Z, h) + r_n$, where $\hat{h}_n$ is defined as one minimum argument of an empirical estimate of $P\ell(Z, h)$ over $\mathcal{H}$. To study such kind of estimator and obtain a non-asymptotic description of r_n, we will need more than point-wise error but uniform error on the estimator of the risk function. In all previous cases, one major issue is to be able to take advantage of small variance.

2 Basic concentration inequalities

The aim here is to recall some standard background on concentration inequalities. All stated results will be useful for the course. Most of them follows from the Chernoff bound which for any real-valued random variable asserts that (for the right tail), for all $t > 0$,

$$\mathbb{P}(Z > t) \leq \inf_{\lambda > 0} \mathbb{E}[\exp(\lambda Z)] \exp(-\lambda t)$$

A similar bound can be obtained for the left tail. The Laplace transform has therefore an important place when applying the Chernoff method.

2.1 Multiplicative Chernoff bound

The following result is known as the multiplicative Chernoff bound. It is based on a upper bound of the Laplace transform of Bernoulli random variables. The proof can be found in Hagerup and Rüb (1990).

Theorem 1. *Let* $(Z_i)_{i=1,\dots,n}$ *be a collection of independent random variables with common distribution* $\mathcal{B}(\mu)$, $\mu \in (0,1)$. *For any* $\eta > 0$ *and* $n \geq 1$, *we have*

$$\mathbb{P}\left(\sum_{i=1}^{n} Z_i < (1 - \eta)n\mu\right) \leq \exp(-\eta^2 n\mu/2)$$

$$.\mathbb{P}\left(\sum_{i=1}^{n} Z_i > (1 + \eta)n\mu\right) \leq \exp(-\eta^2 n\mu/3).$$

Proof. Consider the upper bound (second inequality). Let $Z \sim \mathcal{B}(\mu)$. The Laplace transform satisfies

$$\mathbb{E}[\exp(\lambda Z)] = \mu \exp(\lambda) + (1 - \mu) = 1 + \mu(\exp(\lambda) - 1) \leq \exp(\mu(\exp(\lambda) - 1))$$

and optimizing with respect to λ the quantity $\exp(\mu(\exp(\lambda) - 1) - t\lambda)$ minimum value is $\exp(t - \mu + t\log(\mu/t))$. Setting $t = \mu(1 + \eta)$, the Chernoff bound implies that

$$\mathbb{P}(Z > t) \leq (\exp(\eta - (1 + \eta)\log(1 + \eta)))^{\mu}.$$

One can use the following result, for all $\eta > 0$,

$$\eta - (1 + \eta)\log(\eta + 1) \leq -(1/3)\eta^2$$

which can be obtained from $\eta - (1 + \eta)\log(\eta + 1) = \int_0^{\eta} \frac{t - \eta}{1 + t} dt = -\frac{1}{2}\eta^2 + \frac{1}{2}\int_0^{\eta} \frac{(t - \eta)^2}{(1 + t)^2} dt$ and using that $1 + t \geq 1$, to get that

$$\mathbb{P}(Z > \mu(1 + \eta)) \leq \exp(-(1/3)\mu\eta^2)$$

The result can easily be extend to sums of independent Bernoulli random variables and we obtain the second inequality. The lower bound (first inequality) can be obtained similarly but separating the cases between $\eta \in (0,1)$ and $\eta \geq 1$, and using that $\eta + (1 - \eta)\log(1 - \eta) \geq \frac{1}{2}\eta^2$ when $\eta \in (0,1)$. $\qquad\square$

To compare several tail deviations, one way is to fix the probability to δ. For instance, in the lower upper bound, one may choose $\eta = \sqrt{3 \log(1/\delta)/(n\mu)}$ to get that, with probability at least $1 - \delta$,

$$\sum_{i=1}^{n} (Z_i - \mu) \leq \sqrt{3n\mu \log(1/\delta)}.$$

For the upper bound, one may choose $\eta = \sqrt{2 \log(1/\delta)/(n\mu)}$ and we obtain that, with probability at least $1 - \delta$,

$$\sum_{i=1}^{n} (Z_i - \mu) \geq -\sqrt{2n\mu \log(1/\delta)}.$$

Putting together the two previous inequalities and using the union bound, we obtain the following two-sided bound: with probability $1 - 2\delta$,

$$\left| \sum_{i=1}^{n} (Z_i - \mu) \right| \leq \sqrt{3n\mu \log(1/\delta)}.$$

Note finally that when applying Lemma 1 to $1 - Z_i$ instead of Z_i one could obtain a new upper and a new lower bound that might be advantageous compared to the initial ones depending on the value of μ (see Exercise 1 for a new upper bound).

2.2 Sub-Gaussian bound

Let us now introduce sub-Gaussian random variables.

Definition 2. *Let Z be a real-valued random variable. It is called sub-Gaussian with (variance) factor $v > 0$ if*

$$\mathbb{E}[Z] = 0, \qquad \mathbb{E}[\exp(\lambda Z)] \leq \exp(\lambda^2 v/2).$$

The factor $v > 0$ is often called variance factor because when Z is sub-Gaussian with factor v then $\mathrm{var}(Z) \leq v$. Note that any centered Gaussian random variable Z is sub-Gaussian with factor $v = \mathrm{var}(Z)$. Interestingly, sub-Gaussian is stable under convolution: summing independent sub-Gaussian random variables is still sub-Gaussian. As a consequence, we obtain the following bound for the sum of independent sub-Gaussian random variables.

Theorem 3. *Let $(Z_i)_{i=1,\ldots,n}$ be a collection of independent sub-Gaussian random variables such that each $Z_i - \mathbb{E}[Z_i]$ is sub-Gaussian with factor v_i. Then, for any $n \geq 1$ and $t > 0$, we have*

$$\mathbb{P}\left(\sum_{i=1}^{n} (Z_i - \mathbb{E}Z_i) > t \right) \leq \exp\left(-\frac{t^2}{2 \sum_{i=1}^{n} v_i} \right).$$

Proof. Use Chernoff method and the sub-Gaussian property of $\sum_{i=1}^{n} (Z_i - \mathbb{E}Z_i)$. $\qquad \square$

Note that the same holds true for the left-tail because $-Z$ is still sub-Gaussian. Important examples of sub-Gaussian random variables are Gaussian random variables and bounded random variables. In particular, one may show that whenever $Z \in [a, b]$, $Z - \mathbb{E}Z$ is sub-Gaussian with factor $(b - a)^2/4$. This is the subject of Exercise 2. Applying the previous to $Z_i \in [a_i, b_i]$ we obtain the following bound, which is known as the Hoeffding bound and might be seen as an additive version of the previous multiplicative bound.

Theorem 4. *Let $(Z_i)_{i=1,\ldots,n}$ be an independent collection of random variables such that each Z_i is valued in $[a_i, b_i]$, where $a_i \leq b_i$. Then, for any $n \geq 1$ and $t > 0$, we have*

$$\mathbb{P}\left(\sum_{i=1}^{n}\{Z_i - E[Z_1]\} > t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Proof. After checking that each $Z_i - \mathbb{E}Z_i$ is sub-Gaussian with factor $(b_i - a_i)^2/4$. The proof follows from applying Theorem 3. $\qquad\square$

Note that, in contrast with the multiplicative bound given in Theorem 1, the variance $\text{var}(Z_i)$ is not involved in the Hoeffding bound. This results in a poor scaling when the variance $\text{var}(Z_i)$ is small compared to $(b_i - a_i)^2/4$. This is the case for Bernoulli random variable $\mathcal{B}(\mu)$ with parameter μ that is small (see Exercise 3).

Let us remark that, if η is a Rademacher random variable, i.e., $P(\eta = 1) = P(\eta = -1) = 1/2$ and g is any number, the random variable ηg is sub-Gaussian with variance factor g^2. This will be useful when examining the Rademacher complexity of certain empirical processes (see Section 3). More generally, if Z is a sub-Gaussian random variable with parameter v and $a \in \mathbb{R}$, then aZ is sub-Gaussian with parameter $a^2 v$. This will be useful in our application to k-NN regression in the next section.

Finally, it is important to consider the max of sub-Gaussian random variables (instead of the sum). This is the topic of the following result where the independence of the collection Z_k is not needed any more, in contrast to previous results dealing with sums.

Lemma 5. *Let $(Z_i)_{i=1,\ldots,M}$ be a collection of sub-Gaussian random variables with common sub-Gaussian factor $v > 0$. It holds that*

$$\mathbb{E}[\max_{k=1,\ldots,M} Z_k] \leq \sqrt{2v \log(M)}$$

and with probability $1 - \delta$,

$$\max_{k=1,\ldots,M} Z_k \leq \sqrt{2v \log(M/\delta)}.$$

Proof. For any $\lambda > 0$, it holds

$$\max_{k=1,\ldots,M} Z_k = \lambda^{-1}\log(\exp(\lambda \max_{k=1,\ldots,M} Z_k)) \leq \lambda^{-1}\log(\sum_{k=1}^{M}\exp(\lambda Z_k)).$$

Jensen's inequality and the sub-Gaussian property imply that

$$\mathbb{E}[\max_{k=1,\ldots,M} Z_k] \leq \lambda^{-1}\log(\sum_{k=1}^{M}\mathbb{E}[\exp(\lambda Z_k)]) \leq \lambda^{-1}\log(M\exp(\lambda^2 v/2)) = \lambda^{-1}\log(M) + \lambda v/2.$$

Optimizing in λ gives $\lambda = \sqrt{2\log(M)/v}$ and we obtain the first statement. The upper bound in probability is obtained from the union bound. $\qquad\square$

2.3 Application: Pointwise estimation with k-nearest neighbors

Given a collection of $n \geq 1$ elements $(X_1, Y_1), \ldots, (X_n, Y_n)$, where $X_i \in \mathbb{R}^d$ is the covariate and $Y_i \in \mathbb{R}$ is the response, and a parameter $1 \leq k \leq n$, the k-nearest neighbor (k-NN for short) regression estimate of $h(x) = \mathbb{E}[Y|X = x]$ at point $x \in \mathbb{R}^d$ is defined as

$$\hat{h}(x) = k^{-1} \sum_{i \in N_{n,k}(x)} Y_i,$$

where $N_{n,k}(x)$ is the index set of the k-nearest neighbor to point x among $X_1, \ldots, X_n$ (if there are some ties use the lexicographic order). The k-NN algorithm was initially introduced in Fix and Hodges (1951); Royall (1966); Cover and Hart (1967) and has since been the subject of many studies in the statistical and machine learning literature. For an exposition of the main mathematical results on k-NN estimators, density, regression and classification, we refer to the textbooks Györfi et al. (2006) and Biau and Devroye (2015).

Our goal is to obtain a deviation inequality for the k-NN error $|\hat{h}(x) - h(x)|$, for a given $x \in \mathbb{R}^d$. The analysis is only based on the sub-Gaussian bound and the Chernoff multiplicative bound from the sections before. Let us consider the following standard framework involving independent and identically random variables.

(IID) $((X, Y), (X_1, Y_1), \ldots, (X_n, Y_n))$ is a collection of independent and identically distributed random elements with common distribution P on $\mathbb{R}^d \times \mathbb{R}$.

For the point of interest x at which we want to estimate the regression function, we suppose that the covariates distribution has some positive mass on a ball around x.

(X) The distribution of X admits a density f on $\mathbb{R}^d$. There is $\tau_0 > 0$ such that $b_0 := \inf_{y \in B(x, \tau_0)} f(y) > 0$.

The k-NN bandwidth at x is denoted by $\hat{\tau}_{n,k,x}$ and defined as the smallest radius $\tau \geq 0$ such that the ball $B(x, \tau)$ contains at least k points from the collection $\{X_1, \ldots, X_n\}$. That is,

$$\hat{\tau}_{n,k,x} := \inf\{\tau \geq 0 \, : \, \sum_{i=1}^{n} \mathbb{1}_{B(x,\tau)}(X_i) \geq k\}.$$

We have the following result to control the size of $\hat{\tau}_{n,k,x}$ with the help of $(k/n)^{1/d}$.

Lemma 6. *Suppose that (IID) and (X) hold true. For all $n \geq 1$, $\delta \in (0,1)$ and $1 \leq k \leq n$ such that $4 \log(1/\delta) \leq k \leq n\tau_0^d b_0 V_d / 2$, it holds, with probability at least $1 - \delta$:*

$$\hat{\tau}_{n,k,x} \leq \left(\frac{2k}{n b_0 V_d} \right)^{1/d}.$$

Proof. Define $\overline{\tau}_{n,k,x} = (2k/(n b_0 V_d))^{1/d}$. By definition, $\hat{\tau}_{n,k,x} \leq \overline{\tau}_{n,k}$ as soon as $\sum_{i=1}^{n} \mathbb{1}_{B(x,\overline{\tau}_{n,k})}(X_i) \geq k$. Therefore, we only need to show that the previous happens with probability $1 - \delta$. By muliplicative Chernoff, Lemma 1, it holds with probability $1 - \delta$,

$$\sum_{i=1}^{n} \mathbb{1}_{B(x,\overline{\tau}_{n,k})}(X_i) \geq nP(B(x,\overline{\tau}_{n,k})) - \sqrt{2nP(B(x,\overline{\tau}_{n,k})) \log(1/\delta)}$$

But we have $nP(B(x,\overline{\tau}_{n,k})) \geq n\overline{\tau}_{n,k}^d b_0 V_d = 2k$. The function $x - \sqrt{2xl}$ is increasing whenever $x \geq l/2$. Hence if $2k \geq (1/2)\log(1/\delta)$ we obtain that

$$\sum_{i=1}^n \mathbb{1}_{B(x,\overline{\tau}_{n,k})}(X_i) \geq nP(B(x,\overline{\tau}_{n,k})) - \sqrt{2nP(B(x,\overline{\tau}_{n,k}))\log(1/\delta)} \geq 2k - \sqrt{4k\log(1/\delta)}$$

The above is larger than k whenever $k \geq \sqrt{4k\log(1/\delta)}$ or $k \geq 4\log(1/\delta)$. $\qquad\square$

As a consequence of the previous Lemma, we have that, whenever h is L-Lipschitz, and under the stipulated assumption, it holds, with probability $1 - \delta$,

$$\left| k^{-1} \sum_{i \in N_{n,k}(x)} (h(X_i) - h(x)) \right| \leq L \left(\frac{2k}{nb_0 V_d} \right)^{1/d}.$$

This bound is often referred to as the bias-bound in the analysis of k-NN regression. We note that the smaller k the better the bound. This must be analysed with the variance term, which is the first one of the next well-known decomposition:

$$\hat{h}(x) - h(x) = k^{-1} \sum_{i \in N_{n,k}(x)} \epsilon_i + k^{-1} \sum_{i \in N_{n,k}(x)} (h(X_i) - h(x)).$$

While the bias term has been successfully treated with the help of the multiplicative Chernoff bound, the variance requires the use of the sub-Gaussian bound. The following assumption will allow use to use a sub-Gaussian inequality.

(E) For each $i = 1, \ldots, n$, $\epsilon_i = Y_i - h(X_i)$ is independent from X_i and is sub-Gaussian with factor σ^2.

Now we can state the main result concerning the pointwise analysis of k-NN regression.

Theorem 7. *Suppose that (IID), (X) and (E) hold true and that h is L-Lipschitz. For all $n \geq 1$, $\delta \in (0,1)$ and $1 \leq k \leq n$ such that $4\log(1/\delta) \leq k \leq n\tau_0^d b_0 V_d/2$, we have, with probability $1 - 2\delta$,*

$$\left| \hat{h}(x) - h(x) \right| \leq \sqrt{\frac{2\sigma^2}{k}\log(2/\delta)} + L\left(\frac{2k}{nb_0 V_d} \right)^{1/d}$$

Proof. First consider the conditional probability $\mathbb{P}_\epsilon$ given $(X_i)_{i=1,\ldots,n}$. Write

$$\sum_{i \in N_{n,k}(x)} \epsilon_i = \sum_{i=1}^n \epsilon_i \mathbb{1}_{B(x,\hat{\tau}_{n,k,x})}$$

Since $(\epsilon_i)_{i=1,\ldots,n}$ is sub-Gaussian with factor σ^2 and independent from $(X_i)_{i=1,\ldots,n}$, the above is $\mathbb{P}_\epsilon$ sub-Gaussian with factor $\sigma^2 \sum_{i=1}^n \mathbb{1}_{B(x,\hat{\tau}_{n,k,x})}^2 = \sigma^2 k$. It is therefore $\mathbb{P}$ sub-Gaussian. From Theorem 3, we obtain, with probability $1 - \delta$

$$\left| \sum_{i \in N_{n,k}(x)} \epsilon_i \right| \leq \sqrt{2k\sigma^2\log(2/\delta)}.$$

Using the bias-variance decomposition, the upper for the bias given in Lemma 6 and the union bound allow to conclude the proof. $\qquad\square$

2.4 Exercises

Exercise 1 (improving the multiplicative Chernoff bound)**.** *Show that with probability $1 - \delta$,*

$$\sum_{i=1}^{n}(Z_i - \mu) \leq \sqrt{2n(1 - \mu)\log(1/\delta)}.$$

Deduce that when μ is larger than $2/5$, this is preferable to the one obtained in Lemma 1. (hint: Start applying Lemma 1 to $1 - Z_i$ instead of Z_i).

Exercise 2 (examples of sub-Gaussian random variables)**.** *Show that $Z \sim \mathcal{N}(0, 1)$ is subGaussian with sub-Gaussian factor $v = 1$. Show that $Z \sim \mathcal{N}(0, v)$ is sub-Gaussian with sub-Gaussian factor v. Show that whenever $E[Z] = 0$ and $a \leq Z \leq b$ almost surely for $a \leq 0 \leq b$, then Z is sub-Gaussian with sub-Gaussian factor $(b - a)^2/4$. One may introduce $\psi(\lambda) = \log(\mathbb{E}[\exp(\lambda Z)])$ and use that the second derivative is the variance of a measure dominated by the one of Z. (hint: note that $\mathrm{var}(Z) \leq (b - a)^2/4$). Show that whenever $a \leq Z \leq b$ then $Z - \mathbb{E}[Z]$ is sub-Gaussian with sub-Gaussian factor $(b - a)^2/4$.*

Exercise 3 (multiplicative Chernoff and sub-Gaussian)**.** *Let $(Z_i)_{1 \leq i \leq n}$ be a sequence of independent Bernoulli random variables with mean μ. Compare the two upper bounds obtained (with probability $1 - \delta$) from multiplicative Chernoff and sub-Gaussian. Show that multiplicative Chernoff is better whenever $\mu \leq 1/6$.*

Exercise 4 (from tail to expectation)**.** *Show that $E[|X|] = \int_0^\infty P(|X| > t)dt$. Let $(Z_i)_{i \geq 1}$ be a collection of independent random variables such that each $Z_i - \mathbb{E}[Z_i]$ is sub-Gaussian with factor v_i. Show that*

$$\mathbb{E}|\sum_{i=1}^{n}(Z_i - \mathbb{E}Z_i)| \leq \sqrt{2\pi V}$$

where $V = \sum_{i=1}^{n} v_i$. Using that $\exp(|x|) \leq \exp(x) + \exp(-x)$ and the definition of sub-Gaussian random variables, show that the previous bound can be improved to

$$\mathbb{E}|\sum_{i=1}^{n}(Z_i - \mathbb{E}Z_i)| \leq \sqrt{2V\log(2)}.$$

Exercise 5 (KDE)**.** *Let $x \in \mathbb{R}^d$. Let $\hat{f}_n(x)$ be the KDE estimate of $f(x)$, a L_X-Lipschitz density, based on the uniform Kernel, i.e., $K(x) \propto \mathbb{1}_{\{B(0,1)\}}(x)$. Let $V_d = \int_{B(0,1)} du$. Define $f_h(x) = \mathbb{E}[\hat{f}_n(x)]$. Suppose that h is sufficiently small, i.e., $f_h(x) \in (f(x)/2, 2f(x))$. Show that with probability $1 - \delta$,*

$$|\hat{f}_n(x) - f(x)| \leq hL_X + \sqrt{\frac{6f(x)}{nh^d V_d}\log(2/\delta)}.$$

Conclude arguing that using Hoeffding inequality, one would obtain a poor scaling for the variance term in nh^{2d}.

Exercise 6 (finite ERM)**.** *Let $2, (Z_i)_{i \geq 1}$ be an independent and identically distributed sequence of random variables with common distribution P on $\mathcal{Z}$. Let $\mathcal{H} = \{h_1, \ldots, h_M\}$ be a set of prediction rules that are evaluated through the loss function $\ell : \mathcal{Z} \times \mathcal{H} \to [0, 1]$. Define $\hat{h}_n \in \arg\min_{h \in \mathcal{H}} \hat{R}_n(h)$*

where $\hat{R}_n(h) = n^{-1} \sum_{i=1}^{n} \ell(Z_i, h)$ as well as $h^* \in \arg\min_{h \in \mathcal{H}} R(h)$ where $R(h) = \mathbb{E}[\ell(Z, h)]$. Then for all $\delta \in (0, 1)$, it holds with probability $1 - \delta$,

$$R(\hat{h}_n) \leq R(h^*) + \sqrt{\frac{2}{n} \log((M+1)/\delta)}$$

One can start noting that

$$R(\hat{h}_n) - R(h^*) = R(\hat{h}_n) - \hat{R}_n(\hat{h}_n) + \hat{R}_n(\hat{h}_n) - \hat{R}_n(h^*) + \hat{R}_n(h^*) - R(h^*)$$

where $\hat{R}_n(\hat{h}_n) - \hat{R}_n(h^*) \leq 0$ by definition.

Exercise 7 (OLS). *In the fixed-design linear model with sub-Gaussian noise, i.e., $Y_i = x_i^T \theta + \epsilon_i$, where x_i are real numbers and ϵ_i are sub-Gaussian random variables with factor σ^2, show that*

$$R(\hat{\boldsymbol{\theta}}_n) \leq R(\boldsymbol{\theta}^*) + \frac{2(p+1)\sigma^2}{n} \log(2(p+1)/\delta).$$

3 Empirical processes, Bousquet's inequality, Rademacher complexity

3.1 Background

We shall rely on the following assumption.

(Ziid) $(Z, Z_1, \ldots, Z_n)$ is a collection of random variables independent and identically distributed with common distribution P on $(S, \mathcal{S})$.

We consider a class $\mathcal{G}$ of measurable real-valued functions g defined on S. The aim is to obtain upper bounds, valid under certain probability, on the random variable

$$Z_{\mathcal{G}} = \sup_{g \in \mathcal{G}} \left| \sum_{i=1}^{n} (g(Z_i) - \mathbb{E}[g(Z)]) \right|.$$

Similar bounds have been establish in previous chapter for a single function g or finitely many as in Lemma 5. Here the difficulty is to take study the case were the class $\mathcal{G}$ includes infinitely many elements.

In the sequel, the class $\mathcal{G}$ is supposed to have countably many elements $g \in \mathcal{G}$ to avoid any measurability problem related to the suprema operation. Indeed, a countable supremum of measurable functions is always measurable. In many practical cases, the measurability can be obtained using some other particular properties related to the problem of interest. For instance, the class $\mathcal{G}$ can be uncountable if one has the separability of the underlying process, see for instance Van Handel (2014), which basically ensures that there is $\mathcal{G}_0$ countable so that, almost surely,

$$\sup_{g \in \mathcal{G}} \left| \sum_{i=1}^{n} (g(Z_i) - \mathbb{E}[g(Z)]) \right| = \sup_{g \in \mathcal{G}_0} \left| \sum_{i=1}^{n} (g(Z_i) - \mathbb{E}[g(Z)]) \right|.$$

Another way around this problem is to work with the outer expectation as explained in Van Der Vaart and Wellner (2007).

A good decomposition for the analysis is given by

$$Z_{\mathcal{G}} = (Z_{\mathcal{G}} - \mathbb{E}Z_{\mathcal{G}}) + \mathbb{E}Z_{\mathcal{G}}$$

The first part, $Z_{\mathcal{G}} - \mathbb{E}Z_{\mathcal{G}}$, is examined with the help of Bousquet's inequality (which is given below). The second part, $\mathbb{E}Z_{\mathcal{G}}$, is studied using *symmetrization* or *randomization* and Dudley 's entropy bound.

Important quantities in the sequel will be the uniform bound on the class $\mathcal{G}$ and the greatest variance value among the class $\mathcal{G}$. They are introduced in the following assumption.

(B) There is $U > 0$ and $\sigma > 0$ such that for all $g \in \mathcal{G}$,

$$\sup_{x \in S} |g(x)| \leq U, \qquad \mathrm{var}(g(Z)) \leq \sigma^2.$$

3.2 Bousquet concentration inequality

The following inequality plays a crucial role in statistical learning theory and especially when dealing with functional estimators or empirical risk minimization. It allows for instance, to obtain sharp deviation bound for local regression estimators or training classifiers based on empirical risk minimization. The next statement is Theorem 2.3 from Bousquet (2002) but other versions may be found in Talagrand (1996); Massart (2000); Rio (2002).

Theorem 8. *Under (Ziid) and (B), suppose that $|g| \leq 1$ and $P(g) = 0$. We have for all $\delta \in (0,1)$,*

$$P\left(Z_\mathcal{G} - \mathbb{E}Z_\mathcal{G} \geq \sqrt{2s\log(1/\delta)} + \log(1/\delta)/3\right) \leq \delta.$$

where $s = 2\mathbb{E}Z_\mathcal{G} + n\sigma^2$.

Next theorem is just a rescaling of the previous one to consider functions bounded by U (instead of 1) and that have expected value different from 0.

Theorem 9. *Under (Ziid) and (B), we have with probability $1 - \delta$,*

$$Z_\mathcal{G} \leq 2\mathbb{E}[Z_\mathcal{G}] + \sqrt{2n\sigma^2\log(1/\delta)} + 8U\log(1/\delta)/3.$$

Proof. Using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, with probability $1 - \delta$, it holds

$$Z_\mathcal{G} - \mathbb{E}Z_\mathcal{G} \leq 2\sqrt{\mathbb{E}Z_\mathcal{G}\log(1/\delta)} + \sqrt{2n\sigma^2\log(1/\delta)} + \log(1/\delta)/3,$$

and invoking that $2ab \leq a^2 + b^2$, we obtain

$$Z_\mathcal{G} - \mathbb{E}Z_\mathcal{G} \leq \mathbb{E}Z_\mathcal{G} + \sqrt{2n\sigma^2\log(1/\delta)} + 4\log(1/\delta)/3.$$

As a consequence we have with probability $1 - \delta$,

$$Z_\mathcal{G} \leq 2\mathbb{E}Z_\mathcal{G} + \sqrt{2n\sigma^2\log(1/\delta)} + 4\log(1/\delta)/3.$$

We now apply the previous inequality to $(g - P(g))/2U$ and we get the stated result. $\square$

The previous Theorem provides an upper bound involving 3 terms. The right-hand side term is the negligible compared to the second term, for large value of n. The second term is the standard $\sqrt{n}$-rate that was already observed in previous inequalities (sub-Gaussian in Lemma 3 and multiplicative Chernoff in Lemma 1). It also involves a variance term σ^2 (just as in the Bernstein inequality) which is good for estimation problems where the variance is small. Hence it remains to study the behavior of the left-hand side term, that is, the expected value $\mathbb{E}[Z_\mathcal{G}]$ to finally derive a meaningful deviation bound on $Z_\mathcal{G}$.

3.3 Rademacher complexity

An important property in what follows is that the expected value $Z_\mathcal{G}$ is upper bounded by the so-called Rademacher complexity defined as follows. Let $(\eta_i)_{1 \leq i \leq n}$ denote a collection of independent Rademacher variables, that is, $\mathbb{P}(\eta_i = +1) = \mathbb{P}(\eta_i = -1) = 1/2$. The Rademacher complexities (the one-sided and the other) are given by

$$\mathbb{E}_\eta[\sup_{g \in \mathcal{G}} \sum_{i=1}^n \eta_i g(Z_i)] \quad \text{and} \quad \mathbb{E}_\eta[\sup_{g \in \mathcal{G}} \left|\sum_{i=1}^n \eta_i g(Z_i)\right|].$$

where $\mathbb{E}_\eta$ is over the Rademacher variables. In our context, the Rademacher complexity is useful because of the following Lemma. Depending on the situation, we might use one of the above preferably to the other.

Lemma 10 (symmetrization)**.** *Under (Ziid), suppose that $(Z_i')_{1 \leq i \leq n}$ is another collection independent from $(Z_i)_{1 \leq i \leq 1}$ with the same distribution. We have*

$$\mathbb{E}[Z_{\mathcal{G}}] \leq \mathbb{E}[\sup_{g \in \mathcal{G}} \left| \sum_{i=1}^{n} \eta_i (g(Z_i) - g(Z_i')) \right|] \leq 2\mathbb{E}[Z_{\mathcal{G}}]$$

and

$$\mathbb{E}[Z_{\mathcal{G}}] \leq 2\mathbb{E}_{\eta}[\sup_{g \in \mathcal{G}} \left| \sum_{i=1}^{n} \eta_i g(Z_i) \right|]$$

Proof. In what follows, E_W denote the expectation with respect to W. Write

$$\mathbb{E}[Z] = \mathbb{E}_Z[\sup_{g \in \mathcal{G}} \left| E_{Z'}[\sum_{i=1}^{n} (g(Z_i) - g(Z_i'))] \right|]$$

$$= \mathbb{E}_Z \mathbb{E}_{Z'}[\sup_{g \in \mathcal{G}} \left| \sum_{i=1}^{n} (g(Z_i) - g(Z_i')) \right|]$$

The first inequality follows from the remark that if W is symmetric, then ηW has the same distribution as W. The second and third inequalities are consequence of the triangle inequality. □

Note that other symmetrization inequalities exist. For instance, Gaussian symmetrization are given in Van de Geer et al. (2016). A symmetrization principle for tail probability is given in Bousquet et al. (2004) and another symmetrization result, for convex and increasing functions (different from the above absolute value function) are established in Boucheron et al. (2013). Here are some first properties on Rademacher complexity.

Lemma 11. *Let $\mathrm{Rad}_n(\mathcal{F}) = \mathbb{E}_{\eta}[\sup_{g \in \mathcal{F}} |\sum_{i=1}^{n} \eta_i g(Z_i)|]$. Under (Ziid), it holds*

1. *If $\mathcal{F} \subset \mathcal{G}$ then $\mathrm{Rad}_n(\mathcal{F}) \leq \mathrm{Rad}_n(\mathcal{G})$.*

2. *$\mathrm{Rad}_n(c\mathcal{G}) \leq |c|\mathrm{Rad}_n(\mathcal{G})$.*

3. *$\mathrm{Rad}_n(\mathcal{G} + g_0) \leq \mathrm{Rad}_n(\mathcal{G}) + \|g_0\|_{\infty} n^{1/2}$.*

4. *Let φ be L-Lipschitz, i.e., for all x, y, $|\varphi(x) - \varphi(y)| \leq L|x - y|$ and $\varphi(0) = 0$, then $\mathrm{Rad}_n(\varphi(\mathcal{G})) \leq 2L\mathrm{Rad}_n(\mathcal{G})$. For the one-sided Rademacher (without absolute value), the inequality is valid with L instead of $2L$.*

The last inequality is referred to as the contraction principle, due to (Ledoux and Talagrand, 1991). A proof is given in Boucheron et al. (2013), Theorem 11.6. It will be useful to obtain a bound on the empirical variance $\sum_{i=1}^{n} g(X_i)^2$.

4 Vapnik-Chervonenkis inequalities

We here focus on the Vapnik-Chervonenkis approach, initially introduced in 1971 (see Vapnik and Chervonenkis (2015) for a reprint version), to obtain deviation bound on supremas over certain class of function of sums of independent random variables. This approach relies on a combinatorial concept that is called Vapnik-Chervonenkis dimension and that allows to control suitably the size of the class of functions. We will see in the next section, that the Vapnik-Chervonenkis is in fact part of a more general approach that relies on bounded the metric entropy of the class based on covering numbers.

4.1 Finite classes

We have the following lemma which will be crucial to obtain an upper bound on the Rademacher complexity.

Lemma 12. *For each $g \in \mathcal{G}$ and any collection $(z_1, \ldots, z_n) \subset S$, the random variable $\sum_{i=1}^{n} \eta_i g(z_i)$ is sub-Gaussian with factor $\sum_{i=1}^{n} g(z_i)^2$.*

Proof. By Hoeffding Lemma (see Exercise 2), it holds that $[a, b]$-valued random variables are subGaussia with factor $(b-a)^2/4$. It follows that $\eta_i g(z_i)$ is sub-Gaussian with factor $g(z_i)^2$. The result follows using that sum of sub-Gaussian is still sub-Gaussian. $\square$

The previous sub-Gaussian property can also be stated using the notion of sub-Gaussian process as in Van Handel (2014). A random process $g \mapsto X_g$ on a metric space $(\mathcal{G}, d)$ is called sub-Gaussian if $\mathbb{E}[X_g] = 0$ and for all $\lambda > 0$ and all $(g, h) \in \mathcal{G}$, $\mathbb{E}[\exp(\lambda(X_g - X_h))] \leq \exp(\lambda^2 d(g, h)^2/2)$.

Before introducing covering numbers let us give a basic upper-bound on the Rademacher complexity that is valid when the class is made of a finite number of functions. Let us introduce

$$\|g\|_{L_2(P_n)}^2 = n^{-1} \sum_{i=1}^{n} g(Z_i)^2$$

as well as

$$\hat{\sigma}_n^2(\mathcal{G}) = \sup_{g \in \mathcal{G}} \|g\|_{L_2(P_n)}^2.$$

Lemma 13. *For any collection $(z_1, \ldots, z_n) \subset S$, and a class $\mathcal{G}$ having cardinality $M \geq 1$. It holds with probability 1,*

$$\mathbb{E}_\eta[\sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \eta_i g(Z_i)] \leq \sqrt{2n\hat{\sigma}_n^2(\mathcal{G}) \log(M)}$$

$$\mathbb{E}_\eta[\sup_{g \in \mathcal{G}} |\sum_{i=1}^{n} \eta_i g(Z_i)|] \leq \sqrt{2n\hat{\sigma}_n^2(\mathcal{G}) \log(2M)}.$$

Proof. The first inequality follows from Lemma 5 and Lemma 12. The second bound can be obtained using either that $\exp(|x|) \leq \exp(x) + \exp(-x)$ and adapting Lemma 5 or simply remarking that $|x| = \max(x, -x)$ so that

$$|R_g| = \max(R_g, -R_g) \leq \sup_{g \in \{\mathcal{G} \cup -\mathcal{G}\}} R_g$$

where $R_g = \sum_{i=1}^{n} \eta_i g(z_i)$. Then we conclude applying the first statement.

$\square$

4.2 Vapnik-Chervonenkis dimension

Let $\mathcal{A}$ be a space of sets in S. Given an arbitrary collection $z = z_1, \ldots, z_n$ of distinct points in S, consider the collection of $\mathbb{R}^n$-points

$$\mathbb{1}_{\mathcal{A}}(z) = \{(\mathbb{1}_A(z_1) \ldots, \mathbb{1}_A(z_n)) : A \in \mathcal{A}\} \subset \{0,1\}^n$$

We have that $|\mathbb{1}_{\mathcal{A}}(z)| \leq 2^n$ and when $|\mathbb{1}_{\mathcal{A}}(z)| = 2^n$ we say that z is shattered by $\mathcal{A}$. An important quantity is therefore

$$\mathbb{S}_n(\mathcal{A}) := \sup_{z \in S^n} |\mathbb{1}_{\mathcal{A}}(z)|$$

which is called the shattering coefficient. It represents loosely speaking the number of possible realization in $\mathbb{1}_{\mathcal{A}}$. The Vapnik dimension is defined as

$$vc(\mathcal{A}) = \max\{n \geq 1 : \exists z \in \mathcal{Z}^n \text{ such that } |\mathbb{1}_{\mathcal{A}}(z)| = 2^n\} = \max\{n \geq 1 : \mathbb{S}_n(\mathcal{A}) = 2^n\}.$$

As a consequence, the fact that all given $z_1, \cdots, z_{v+1}$ points cannot be shattered is equivalent to the fact that VC dimension is smaller than v. The reason why the VC dimension is appropriate to control the complexity of classes of sets is perhaps explained by the Sauer's lemma (see Lugosi (2002) for a proof) which states that

$$\mathbb{S}_n(\mathcal{A}) \leq \sum_{i=0}^{vc(\mathcal{A})} \binom{n}{i}$$

Note that when $n \leq vc(\mathcal{A})$ the bound $2^n \leq 2^{vc(\mathcal{A})}$ is better than the above. An interesting consequence is that

$$\mathbb{S}_n(\mathcal{A}) \leq (n+1)^{vc(\mathcal{A})}.$$

Hence the above bound for n large, improves upon the 2^n bound that is given from the definition. Consequently, the number of functions within $\mathbb{1}_{\mathcal{A}}(z)$ is of order $n^{vc(\mathcal{A})}$ and, loosely speaking, suprema mith be replaced by maximum. Two examples are now given to illustrate the previous. The class of cells $(-\infty, t]$ can be used to shatter any set with 1 point but fails to shatter any set with 2 points. Its Vapnik dimension is therefore 1. The class of cells $(s, t]$ can be used to shatter any set with 2 distinct points but fails to shatter any set with 3 distinct points. Its Vapnik dimension is therefore 2. An elementary preservation property is now given.

Lemma 14. *We have that $vc(\mathcal{A}) = vc(\mathcal{A}^c)$ and whenever $\mathcal{A} \subset \mathcal{B}$, $vc(\mathcal{A}) \leq vc(\mathcal{B})$.*

4.3 Main inequalities

We now give the first Vapnik's concentration bound tailored to functions valued in $\{0,1\}$ with finite Vapnik dimension.

Theorem 15 (First Vapnik inequality). *Under (Ziid), for any Borelian class $\mathcal{A} \subset \mathcal{S}$, $\delta > 0$ and $n \geq 1$, it holds with probability at least $1 - \delta$:*

$$\left| \sum_{i=1}^n (\mathbb{1}_B(Z_i) - P(B)) \right| \leq \sqrt{8n \log(2\mathbb{S}_{\mathcal{A}}(n)/\delta)}, \qquad \forall B \in \mathcal{A}$$

In Devroye et al. (2013), a proof of a similar result is given. Also in last year exam, there is a proof of the above result. The previous results can be refined to include a variance term in the bound as stated below. This is done in Theorem 2.1 in Anthony and Shawe-Taylor (1993) (see also Theorem 1.11 in Lugosi (2002)) which is now stated.

Theorem 16 (Second Vapnik inequality). *Under (Ziid), for any Borelian class $\mathcal{A} \subset \mathcal{S}$, $\delta > 0$ and $n \geq 1$, it holds with probability at least $1 - \delta$:*

$$\sum_{i=1}^{n} (\mathbb{1}_B(X_i) - P(B)) \geq -\sqrt{4nP(B)\log(4\mathbb{S}_{\mathcal{A}}(2n)/\delta)}, \qquad \forall B \in \mathcal{A},$$

The previous results are interesting when the Vapnik dimension $vc(\mathcal{A})$ (introduced in Section 6.3.3) is finite so that $\mathbb{S}_{\mathcal{A}}(n) \leq (n+1)^{vc(\mathcal{A})}$. In Devroye et al. (2013), Lemma 29.1, it shown that the previous concepts can be extended to more general function classes using the class generated by sub-graphs, a notion that will be useful in dealing with the kernel smoothing estimator of the density.

4.4 Examples

Many popular classes of sets have a known Vapnik dimension. Some relevant results might be found in Wenocur and Dudley (1981). For instance, it is shown that the class of cells made of elements as $(-\infty, b] = \prod_{k=1}^{d}(-\infty, b_k]$, where $b \in \mathbb{R}^d$ has a Vapnik dimension equal to d. This is specified in the next lemma.

Lemma 17. *We have $vc(\{(-\infty, b] : b \in \mathbb{R}^d\}) = d$ and $vc(\{(-a, b] : b \in \mathbb{R}^d, a \in \mathbb{R}^d\}) = 2d$.*

Similarly, the class of balls has Vapnik dimension equal to $d + 1$ as stated below.

Lemma 18. *We have $vc(\{B(x, r) : x \in \mathbb{R}^d, r > 0\}) = d + 1$.*

We also have the following lemma about hyperplanes.

Lemma 19. *We have $vc(\{\langle w, x \rangle > 0 : w \in \mathbb{R}^d\}) = d$.*

A proof of a slightly more general result is given in (Giné and Nickl, 2021), Proposition 2.6.6. An important consequence is for *support vector machine* classification because the previous result implies that the class of all affine hyperplanes in $\mathbb{R}^d$ is of dimension $d + 1$.

4.5 Exercises

Exercise 8. *Give conditions on n and δ to ensure that with probability at least $1 - \delta$,*

$$\sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)| \leq C\sqrt{\frac{\log(n/\delta)}{n}}.$$

Exercise 9. *Suppose that all elements $g \in \mathcal{G}$ are such that $0 \leq g(x) \leq U$. Show that*

$$\mathbb{E}[\sup_{g \in \mathcal{G}} n^{-1} \sum_{i=1}^{n} \{g(Z_i) - P(g(Z))\}] \leq 2\mathbb{E}[\sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \eta_i \mathbb{1}_{\{U_i < g(Z_i)\}}]$$

for some collection $(U_i)_{1 \leq i \leq n}$ of independent and identically distributed random variables. Give a Vapnik inequality (of the first type) for such classes.

Exercise 10. *1. Let $(Z'_1, \ldots, Z'_n)$ be a random collection with same distribution as $(Z_1, \ldots, Z_n)$ and independent from $(Z_1, \ldots, Z_n)$ and $(\eta_1, \ldots, \eta_n)$. Let $\Psi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ non-decreasing and convex. Show that*

$$\mathbb{E}_Z[\Psi(\sup_{g \in \mathcal{G}} Z_g)] \leq \mathbb{E}_Z \mathbb{E}_{Z'}[\Psi(\sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \{g(Z_i) - g(Z'_i)\})]$$

2. Deduce that

$$\mathbb{E}_Z[\Psi(\sup_{g \in \mathcal{G}} Z_g)] \leq \mathbb{E}_Z \mathbb{E}_{\eta}[\Psi(2 \sup_{g \in \mathcal{G}} R_g)]$$

3. Suppose that $g = \mathbb{1}_A$ where $A \in \mathcal{A}$, is a collection of measurable sets in S. Let $z = (z_1, \ldots, z_n) \subset S$, $\mathcal{W}(z) = \{(\mathbb{1}_A(z_1), \ldots, \mathbb{1}_A(z_n)) : A \in \mathcal{A}\}$. Show that

$$\mathbb{E}_{\eta}[\Psi(2 \sup_{g \in \mathcal{G}} R_g)] \leq \sum_{(w_1, \ldots, w_n) \in \mathcal{W}(z)} \mathbb{E}_{\eta}[\Psi(2 \sum_{i=1}^{n} \eta_i w_i)]$$

4. Let $\mathbb{S}_n(\mathcal{A}) = \max_{z_1, \ldots, z_n} |\mathcal{W}(z)|$ where for any collection of sets A, $|A|$ denotes the number of elements in A. Show that

$$\mathbb{P}(\sup_{g \in \mathcal{G}} Z_g > t) \leq \mathbb{S}_n(\mathcal{A}) \exp(-t^2/(8s^2))$$

($\Psi := \Psi_{\lambda, t}$ shall be well chosen)

5 Covering numbers and Dudley's entropy integral

5.1 Covering numbers

A standard way to control the fluctuation of the Rademacher complexity relies on the concept of covering numbers. Those numbers are useful to measure the size or the complexity of a function class. Given $\epsilon > 0$, the ϵ-covering number is the smallest number of closed balls of radius $\epsilon > 0$ required to cover $\mathcal{G}$. The smaller the covering numbers the more simple the function class and vice-versa. The next definition is formal and relies on ϵ-net.

Definition 20 (ϵ-net and covering numbers). *Given a metric space $(\mathcal{G}, d)$, a set $\mathcal{G}_\epsilon$ is called ϵ-net if for all $g \in \mathcal{G}$ there is $\pi(g) \in \mathcal{G}_\epsilon$ such that $d(g, \pi(g)) \leq \epsilon$. The ϵ-covering number, denoted $\mathcal{N}(\mathcal{G}, d, \epsilon)$, is defined as the smallest cardinality of an ϵ-net.*

We therefore focus on classes $(\mathcal{G}, d)$ that are called totally bounded, i.e., for any $\epsilon > 0$, there exist finitely many balls of radius ϵ that can be used to cover $\mathcal{G}$. In the application we have in mind the space $\mathcal{G}$ is a subsapce of $L_2(Q)$ and an ϵ-net don't have to be included in $\mathcal{G}$ but just to be elements in $L_2(Q)$. Here are some preservation properties related to covering numbers.

Lemma 21. *Suppose that $\mathcal{F}$ and $\mathcal{G}$ have envelop $U_\mathcal{F}$ and $U_\mathcal{G}$. Then*

$$\mathcal{N}(-\mathcal{F}, L_2(Q), \epsilon U_\mathcal{F}) \leq \mathcal{N}(\mathcal{F}, L_2(Q), \epsilon U_\mathcal{F})$$
$$\mathcal{N}(\mathcal{F} \cup \mathcal{G}, L_2(Q), \epsilon U_\mathcal{F} \vee U_\mathcal{G}) \leq \mathcal{N}(\mathcal{F}, L_2(Q), \epsilon U_\mathcal{F}) + \mathcal{N}(\mathcal{G}, L_2(Q), \epsilon U_\mathcal{G})$$
$$\mathcal{N}(\mathcal{F} + \mathcal{G}, L_2(Q), \epsilon (U_\mathcal{F} + U_\mathcal{G})) \leq \mathcal{N}(\mathcal{F}, L_2(Q), \epsilon U_\mathcal{F}) \mathcal{N}(\mathcal{G}, L_2(Q), \epsilon U_\mathcal{G})$$
$$\mathcal{N}(\mathcal{F}\mathcal{G}, L_2(Q), \epsilon U_\mathcal{F} U_\mathcal{G}) \leq \mathcal{N}(\mathcal{F}, L_2(Q), \epsilon U_\mathcal{F}/2) \mathcal{N}(\mathcal{G}, L_2(Q), \epsilon U_\mathcal{G}/2)$$

5.2 Main results

The following result, due to Dudley (see Van Handel (2014), Corollary 5.25 or Boucheron et al. (2013), Corollary 13.2, or (Zhang, 2023)) is crucial in the empirical process theory. It consists in the following upper-bound on the Rademacher complexity using the entropy-integral. The proof relies on the previous bound (for finite class) and the so-called chaining method.

Theorem 22 (Dudley). *For any collection $(Z_1, \ldots, Z_n) \subset S$, it holds with probability 1,*

$$\mathbb{E}_\eta [\sup_{g \in \mathcal{G}} \sum_{i=1}^n \eta_i g(Z_i)] \leq 12\sqrt{n} \int_0^{\hat{\sigma}_n(\mathcal{G})/2} \sqrt{\log(\mathcal{N}(\mathcal{G}, L_2(P_n), \epsilon))} \mathrm{d}\epsilon.$$

Moreover, we have, with probability 1,

$$\mathbb{E}_\eta [\sup_{g \in \mathcal{G}} \left| \sum_{i=1}^n \eta_i g(Z_i) \right|] \leq 12\sqrt{n} \int_0^{\hat{\sigma}_n(\mathcal{G})/2} \sqrt{\log(2\mathcal{N}(\mathcal{G}, L_2(P_n), \epsilon))} \mathrm{d}\epsilon.$$

Proof. The proof given here is adapted from (Zhang, 2023) and from S. Kakade's lecture notes (spring 2008). Set $\mathcal{G}_0 = \{0\}$ and $\epsilon_0 = \hat{\sigma}_n(\mathcal{G})$. Let $\epsilon_k = 2^{-k}\epsilon_0$, $k = 1, 2, \ldots$ and let $\mathcal{G}_k$ be a ϵ_k-net. Write

$$g = g - \pi_L(g) + \sum_{k=1}^L (\pi_k(g) - \pi_{k-1}(g))$$

Triangle inequality gives $\|\pi_k(g) - \pi_{k-1}(g)\|_{L_2(P_n)} \leq 3\epsilon_k$. As a consequence of Lemma 13, we have

$$\mathbb{E} \sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \eta_i (\pi_k(g) - \pi_{k-1}(g))_i \leq 3\epsilon_k \sqrt{2n \log(N_k N_{k-1})}$$

$$\leq 6(\epsilon_k - \epsilon_{k+1}) \sqrt{2n \log(N_k N_{k-1})}$$

$$\leq 12(\epsilon_k - \epsilon_{k+1}) \sqrt{n \log(N_k)}$$

$$= 12\sqrt{n \log(N_k)} \int_{\epsilon_{k+1}}^{\epsilon_k} du$$

$$\leq 12\sqrt{n} \int_{\epsilon_{k+1}}^{\epsilon_k} \sqrt{\log(N(\mathcal{G}, L_2(P_n), u))} du$$

Taking the sum over k we get that

$$\mathbb{E} \sum_{k=1}^{L} \sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \eta_i (\pi_k(g) - \pi_{k-1}(g))_i \leq 12\sqrt{n} \int_{\epsilon_L/2}^{\epsilon_1} \sqrt{\log(N(\mathcal{G}, L_2(P_n), u))} du.$$

Considering the first term, one has that (from Cauchy-Schwarz inequality)

$$|\sum_{i=1}^{n} \eta_i (g - \pi_L(g))_i| \leq n\|g - \pi_L(g)\|_{L_2(P_n)} \leq n\epsilon_L.$$

Consequently,

$$\mathbb{E}_\eta [\sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \eta_i g(Z_i)] \leq n\epsilon_L + 12\sqrt{n} \int_0^{\epsilon_0/2} \sqrt{\log(N(\mathcal{G}, L_2(P_n), u))} du$$

Taking the limit with respect to $L \to \infty$ leads to the result.

Now we consider the second statement. Set $R_g = \sum_{i=1}^{n} \eta_i g(Z_i)$. We have

$$|R_g| \leq R_g \vee -R_g \leq \sup_g R_g \vee \sup_g -R_g.$$

Noting that $\sup_g R_g$ and $\sup_g -R_g$ are both smaller than $\sup_{g \in \mathcal{G} \cup -\mathcal{G}} R_g$. We obtain that

$$\sup_g |R_g| \leq \sup_{g \in \mathcal{G} \cup -\mathcal{G}} R_g.$$

(one can also show the reverse inequality). Now remark that covering numbers of $\mathcal{G} \cup -\mathcal{G}$ are smaller than twice the covering numbers of $\mathcal{G}$. We conclude by applying Theorem 22 with the class $\mathcal{G} \cup -\mathcal{G}$. $\qquad \square$

The distance in $L_2(P_n)$ that appears in the previous statement is therefore important and we shall need at some point to have an upper bound on $\hat{\sigma}_n^2(\mathcal{G})$. The following is an easy consequence of the contraction principle given in the last statement of Lemma 11.

Lemma 23. *Under (Ziid), for any class $\mathcal{G}$ with envelop U, we have*

$$n\mathbb{E}\hat{\sigma}_n^2(\mathcal{G}) \leq n \sup_{g \in \mathcal{G}} P(g^2) + 4U\mathbb{E}[\sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \eta_i g(Z_i)]$$

The previous result can be found in Corollary 3.4 in Talagrand (1994) (with a less accurate constant 8 rather than the above 4) or in Van Handel (2014). The proof is the subject of an exercise in next section.

5.3 Exercises

Exercise 11. *Define $\tilde{\sigma}_n^2(\mathcal{G}) = n^{-1} \sup_{g \in \mathcal{G}} \sum_{i=1}^n (g(Z_i)^2 - \mathbb{P}g^2(Z))$. Under the assumption of Lemma 23 show that*

$$n\mathbb{E}\tilde{\sigma}_n^2(\mathcal{G}) \leq 4U\mathbb{E}_\eta[\sup_{g \in \mathcal{G}} \sum_{i=1}^n \eta_i g(Z_i)].$$

Conclude by showing the statement of Lemma 23.

Exercise 12. *Suppose that $\mathcal{G}$ has envelop function G. Show that for any probability measure Q,*

$$\mathcal{N}(\mathcal{G}^2, L_2(Q), \epsilon\|2G^2\|_{L_2(Q)}) \leq \mathcal{N}(\mathcal{G}, L_2(\tilde{Q}), \epsilon\|G\|_{L_2(\tilde{Q})})$$

where $d\tilde{Q} \propto (2G)^2 dQ$.

Exercise 13. *Suppose that, for all $\epsilon \in (0,1)$,*

$$\mathcal{N}(\mathcal{G}, L_2(Q), \epsilon U) \leq (A/\epsilon)^v$$

where (A, v) are two constants larger than 1. Define $\tilde{\sigma}_n^2(\mathcal{G}) = n^{-1} \sup_{g \in \mathcal{G}} \sum_{i=1}^n (g(Z_i)^2 - \mathbb{P}g^2(Z))$. Show that

$$n\mathbb{E}\tilde{\sigma}_n^2(\mathcal{G}) \leq 24\sqrt{n}\mathbb{E}\int_0^{\hat{\sigma}_n(\mathcal{G}^2)/2} \sqrt{\log(\mathcal{N}(\mathcal{G}^2, L_2(P_n), \epsilon))}d\epsilon$$

Conclude that

$$n\mathbb{E}\tilde{\sigma}_n^2(\mathcal{G}) \leq 12U^2\sqrt{2nv\log(4eA)}$$

Exercise 14. *Show that*

$$\mathbb{E}_\eta[\sup_{g \in \mathcal{G}} \sum_{i=1}^n \eta_i(g(Z_i) - g(Z_i'))]$$
$$\leq 12\sqrt{n}\int_0^{(s_n(\mathcal{G})+s_n'(\mathcal{G}))/2} \sqrt{\log(\mathcal{N}(\mathcal{G}, L_2(P_n), \epsilon/2)\mathcal{N}(\mathcal{G}, L_2(P_n'), \epsilon/2))}d\epsilon$$

where $s_n^2(\mathcal{G}) = n^{-1} \sup_{g \in \mathcal{G}} \sum_{i=1}^n (g(Z_i) - \mathbb{E}g(Z))^2$ and P_n is empirical measure on $(Z_i)_{1 \leq i \leq n}$ (s_n' and P_n' are defined accordingly with respect to Z_i').

Exercise 15. *Let $Lip = \{f : [0,1] \to \mathbb{R} : |f(x) - f(y)| \leq |x - y|\}$. Recall that $W_1(P_n, P) = \sup_{f \in Lip} P_n(f) - P(f)$ and that $\mathcal{N}(Lip, \|\cdot\|_\infty, \epsilon) \leq \exp(c/\epsilon)$. Show that $\mathbb{E}[W_1(P_n, P)] \leq 24\sqrt{2nc}$ and conclude that $W_1(P_n, P) \to 0$ almost surely.*

6 VC classes

6.1 Definition

The concept of VC class is standard to obtain uniform bounds on the empirical process. While different definitions exist we here follow the definition given in Nolan and Pollard (1987); Giné and Guillou (2002) which is based on uniform entropy numbers. This notion is more general than the one initially introduced in Vapnik and Chervonenkis (2015) as detailed for instance in (Van Der Vaart and Wellner, 1996, Theorem 2.6.4) and as shown next.

Given a probability measure Q on $(S, \mathcal{S})$, the metric space of squared-integrable functions with respect to Q is defined as

$$L_2(Q) = \left\{ g : S \mapsto \mathbb{R} \text{ such that } \|g\|^2_{L_2(Q)} := Q(g^2) < \infty \right\}.$$

Definition 24. *A class $\mathcal{G}$ of pointwise measurable (Van Der Vaart and Wellner, 1996, Example 2.3.4) real-valued functions on a measurable space $(S, \mathcal{S})$ is said to be VC with parameters $(v, A) \in (1, \infty) \times [1, \infty)$ and envelope G if for any $0 < \epsilon < 1$ and any probability measure Q on $(S, \mathcal{S})$, we have*

$$\mathcal{N}\left(\mathcal{G}, L_2(Q), \epsilon\|G\|_{L_2(Q)}\right) \leq (A/\epsilon)^v.$$

The condition on the parameter (v, A) might be alleviated at the price of more technicalities in the proof of the main results.

6.2 Main results

The previous inequality can be applied for bounding the Rademacher complexity associated with function classes that are localized in terms of variance. That is, we define

$$\mathcal{G}_r = \{g \in \mathcal{G} \, : \, P_n(g^2) \leq r\},$$

and introduce the notation

$$\text{Rad}_{n,s}(\mathcal{F}) = \mathbb{E}_\eta[\sup_{g \in \mathcal{F}} \sum_{i=1}^n \eta_i g(Z_i)],$$

$$\text{Rad}_n(\mathcal{F}) = \mathbb{E}_\eta[\sup_{g \in \mathcal{F}} \left|\sum_{i=1}^n \eta_i g(Z_i)\right|].$$

When $r > U^2$, since $P_n(g^2) \leq U^2$, we have that $\mathcal{G}_r = \mathcal{G}_{U^2}$. Therefore we assume subsequently that $r \leq U^2$.

Theorem 25. *Suppose that $\mathcal{G}$ is of VC type with envelop U and parameter (v, A). For any collection $(Z_1, \ldots, Z_n) \subset S$ and any $r \leq U^2$, it holds that, with probability 1,*

$$\text{Rad}_{n,s}(\mathcal{G}_r) \leq 6\sqrt{2nrv \log(2eAU/\sqrt{r})}.$$

In particular, when using $r^2 = \sup_{g \in \mathcal{G}} P_n(g^2)$ and $r = U$, we find, respectively,

$$\text{Rad}_{n,s}(\mathcal{G}) \leq 6\sqrt{2}\sqrt{\hat{\sigma}_n^2(\mathcal{G})nv \log(2eAU/\hat{\sigma}_n(\mathcal{G}))}$$

$$\text{Rad}_{n,s}(\mathcal{G}) \leq 6\sqrt{2}\sqrt{U^2 nv \log(2eA)}.$$

Proof. Applying Theorem 22, we find

$$\mathrm{Rad}_{n,s}(\mathcal{G}_r) \leq 12\sqrt{n} \int_0^{\sqrt{r}/2} \sqrt{\log(\mathcal{N}(\mathcal{G}, L_2(P_n), \epsilon))} \mathrm{d}\epsilon.$$

An envelop for the class is U hence

$$\mathrm{Rad}_{n,s}(\mathcal{G}_r) \leq 12U\sqrt{n} \int_0^{\sqrt{r}/(2U)} \sqrt{\log(2\mathcal{N}(\mathcal{G}, L_2(P_n), U\epsilon))} \mathrm{d}\epsilon.$$

Using Definition 24, we obtain

$$\mathrm{Rad}_{n,s}(\mathcal{G}_r) \leq 12U\sqrt{vn} \int_0^{\sqrt{r}/(2U)} \sqrt{\log(A/\epsilon)} \mathrm{d}\epsilon$$

By change of variable, it follows that

$$\mathrm{Rad}_{n,s}(\mathcal{G}_r) \leq 6\sqrt{nrv} \int_0^1 \sqrt{\log(2AU/(\epsilon\sqrt{r}))} \mathrm{d}\epsilon$$

$$= 6\sqrt{nrv} \int_0^1 \sqrt{\log(2AU/\sqrt{r}) + \log(1/\epsilon)} \mathrm{d}\epsilon$$

$$\leq 6\sqrt{nrv} \left(\sqrt{\log(2AU/\sqrt{r})} + \int_0^1 \sqrt{\log(1/\epsilon)} \mathrm{d}\epsilon \right)$$

Remarking that $\int_0^1 \sqrt{\log(1/\epsilon)} \mathrm{d}\epsilon = \sqrt{\pi}/2$, we obtain

$$\mathrm{Rad}_{n,s}(\mathcal{G}_r) \leq 6\sqrt{nrv} \left(\sqrt{\log(2AU/\sqrt{r})} + 1 \right)$$

$$\leq 6\sqrt{2nrv} \sqrt{(\log(2AU/\sqrt{r}) + 1)}$$

$$\leq 6\sqrt{2nrv \log(2eAU/\sqrt{r})}.$$

$\square$

Similar statement holds for Rad_n in place of $\mathrm{Rad}_{n,s}$. This leads to a slight increase of A which becomes $2A$.

Theorem 26. *Suppose that $\mathcal{G}$ is of VC type with envelop U and parameter (v, A). For any collection $(z_1, \ldots, z_n) \subset S$ and any $r \leq U^2$, it holds that, with probability 1,*

$$\mathrm{Rad}_n(\mathcal{G}_r) \leq 6\sqrt{2} \sqrt{rnv \log(4eAU/\sqrt{r})}.$$

In particular, when using $r^2 = \sup_{g \in \mathcal{G}} P_n(g^2)$ and $r = U$, we find

$$\mathrm{Rad}_n(\mathcal{G}) \leq 6\sqrt{2} \sqrt{\hat{\sigma}_n^2(\mathcal{G})nv \log(4eAU/\hat{\sigma}_n(\mathcal{G}))}$$

$$\mathrm{Rad}_n(\mathcal{G}) \leq 6\sqrt{2} \sqrt{U^2 nv \log(4eA)}.$$

One further result, which can be found in Giné and Guillou (2002), provides a bound on the expected Rademacher complexity. Note that the next upper bound is similar to the classic Bernstein inequality as we have a variance term and another smaller term that depends on U, the uniform bound on the random variables of interest.

Theorem 27. *Under (Ziid), suppose that $\mathcal{G}$ is of VC type with envelop U and parameter (v, A). Suppose that $\sup_{g \in \mathcal{G}} E(g^2) \leq \sigma^2$. It holds that*

$$\mathbb{E}\mathrm{Rad}_{n,s}(\mathcal{G}) \leq C\sqrt{vn\sigma^2 \log(2eAU/\sigma)} + 4C^2Uv\log(2eAU/\sigma).$$

with $C = 6\sqrt{2}$. The same bound is valid for $\mathbb{E}\mathrm{Rad}_n(\mathcal{G})$ but with $4eAU$ in place of $2eAU$.

Proof. Recall that $C = 6\sqrt{2}$ and apply the second statement in Theorem 26. We get using (twice) Jensen inequality (functions $\sqrt{x}$ and $ax\log(b/x)$ are both concave), we get

$$\mathbb{E}[\mathrm{Rad}_{n,s}(\mathcal{G})] \leq C\sqrt{\mathbb{E}[\hat{\sigma}_n^2(\mathcal{G})nv\log(2eAU/\hat{\sigma}_n(\mathcal{G}))]}$$
$$\leq C\sqrt{\mathbb{E}[\hat{\sigma}_n^2(\mathcal{G})]nv\log((2eAU)^2/\mathbb{E}[\hat{\sigma}_n^2(\mathcal{G})])/2}$$

From Lemma 23, we obtain

$$\mathbb{E}[\hat{\sigma}_n^2(\mathcal{G})] \leq \sigma^2 + 4Un^{-1}\mathbb{E}[\mathrm{Rad}_{n,s}(\mathcal{G})].$$

Now remark that $x\log(b/x)$ is non-decreasing for $x \leq b/e$. This is always satisfied for $x = \sigma^2 + 4Un^{-1}\mathbb{E}[\mathrm{Rad}_{n,s}(\mathcal{G})]$ because this quantity is smaller than $5U^2$ which itself is smaller than $b/e = (2AU)^2e$. We therefore obtain

$$\mathbb{E}[\mathrm{Rad}_{n,s}(\mathcal{G})] \leq C\sqrt{nv(\sigma^2 + 4Un^{-1}\mathbb{E}[\mathrm{Rad}_{n,s}(\mathcal{G})])\log((2eAU)^2/(\sigma^2 + 4Un^{-1}\mathbb{E}[\mathrm{Rad}_{n,s}(\mathcal{G})]))]/2}$$
$$\leq C\sqrt{nv(\sigma^2 + 4Un^{-1}\mathbb{E}[\mathrm{Rad}_{n,s}(\mathcal{G})])\log((2eAU)^2/\sigma^2)/2}$$

It follows that

$$\mathbb{E}[\mathrm{Rad}_{n,s}(\mathcal{G})]^2 \leq C^2nv(\sigma^2 + 4Un^{-1}\mathbb{E}[\mathrm{Rad}_{n,s}(\mathcal{G})])\log(2eAU/\sigma) = a\mathbb{E}[\mathrm{Rad}_{n,s}(\mathcal{G})] + b$$

with $a = C^2 4Uv\log(2eAU/\sigma)$ and $b = C^2vn\sigma^2\log(2eAU/\sigma)$. Using Lemma 54, we get the stated result as $C^2 4 = 36 \times 2 \times 4 = 288$. The bound for $\mathbb{E}[\mathrm{Rad}_n(\mathcal{G})]$ is obtained noting that $\mathbb{E}[\mathrm{Rad}_n(\mathcal{G})] = \mathbb{E}[\mathrm{Rad}_{n,s}(\mathcal{G} \cup -\mathcal{G})]$ and that the class $\mathcal{G} \cup -\mathcal{G}$ has a covering number twice the size as the one of $\mathcal{G}$. It is then of VC type with parameter $2^{1/v}A \leq 2A$ and v. $\square$

Now we can put together Theorem 27 and Bousquet's inequality, Theorem 9, to obtain the following.

Theorem 28. *Let $(Z, Z_1, \ldots, Z_n)$ be an independent and identically distributed collection of random variables in $(S, \mathcal{S})$. Let $\mathcal{G}$ be a VC class of functions with parameters $v \geq 1$, $A \geq 1$ and uniform envelope $U \geq \sup_{g \in \mathcal{G}, x \in S} |g(x)|$. Let σ be such that $\sigma^2 \geq \sup_{g \in \mathcal{G}} Pg^2$ and $\sigma \leq U$. For any $n \geq 1$ and $\delta \in (0, 1)$, it holds, with probability at least $1 - \delta$,*

$$\sup_{g \in \mathcal{G}}\left|\sum_{i=1}^n \{g(Z_i) - P(g)\}\right| \leq \sqrt{4C^2vn\sigma^2\log(4eAU/(\delta\sigma))} + 16C^2Uv\log(4eAU/(\delta\sigma))$$

with $C = 6\sqrt{2}$.

Proof. Recall that $Z_{\mathcal{G}} = \sup_{g \in \mathcal{G}}|\sum_{i=1}^n \{g(Z_i) - P(g)\}|$. Combining Theorem 9 and symmetrization inequality we get

$$Z_{\mathcal{G}} \leq 4\mathbb{E}[\mathrm{Rad}_n(\mathcal{G})] + \sqrt{2n\sigma^2\log(1/\delta)} + 8U\log(1/\delta)/3 \tag{1}$$

Relying on Theorem 27 to bound $\mathbb{E}[\mathrm{Rad}_n(\mathcal{G})]$, we get

$$
\begin{aligned}
Z_{\mathcal{G}} &\leq 4C\sqrt{vn\sigma^2\log(4eAU/\sigma)} + 16C^2Uv\log(4eAU/\sigma) + \sqrt{2n\sigma^2\log(1/\delta)} + 8U\log(1/\delta)/3 \\
&\leq \sqrt{4C^2vn\sigma^2\log(4eAU/\sigma) + 4n\sigma^2\log(1/\delta)} + (16C^2Uv \vee (8/3)U)\log(4eAU/(\delta\sigma)) \\
&\leq \sqrt{4(C^2\vee 1)vn\sigma^2\log(4eAU/(\delta\sigma))} + Uv(16C^2 \vee 8/3)\log(4eAU/(\delta\sigma))
\end{aligned}
$$

where, in the second step, we have used that $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}$ and in the last step that $v \geq 1$. It remains to use that $C = 6\sqrt{2}$ to obtain the stated result.

$\square$

Theorem 29. *In Theorem 28, if*

$$
n\left(\frac{\sigma}{U}\right)^2 \geq (16 \times 4)C^2 v\log(4eAU/(\delta\sigma))
$$

it holds, with probability at least $1 - \delta$,

$$
\sup_{g \in \mathcal{G}}\left|\sum_{i=1}^n \{g(Z_i) - P(g)\}\right| \leq 12\sqrt{vn\sigma^2\log(4eAU/(\delta\sigma))}
$$

6.3 Examples

6.3.1 Indicator of cells

Due to its use in the definition of cumulative distribution function, one standard example the class of indicators of the cells $(-\infty, t]$, when $t \in \mathbb{R}$. There is actually two ways to deal with this class. The first one, which is presented in this paragraph, is using bracketing numbers (which is another notion of complexity). The second one is based on the shattering coefficient which will be presented next. Let us introduce bracketing numbers.

Definition 30 (bracketing numbers). *Let $\underline{f}$ and $\overline{f}$ be two functions in $L_2(Q)$. The set $[\underline{f}, \overline{f}]$, called a bracket, denotes the set of all functions g in $L_2(Q)$ such that $\underline{f} \leq g \leq \overline{f}$. A bracket $[\underline{f}, \overline{f}]$ such that $\|\underline{f} - \overline{f}\|_{L_2(Q)} \leq \epsilon$ is called an ϵ-bracket. Given $\mathcal{G} \subset L_2(Q)$, the ϵ-bracketing number, denoted $\mathcal{N}_{[]}(\mathcal{G}, L_2(Q), \epsilon)$, is defined as the smallest number of brackets of radius $\epsilon > 0$ needed to cover $\mathcal{G}$.*

The next lemma consider the bracketing numbers associated to the class of cells in $\mathbb{R}$.

$$
\mathcal{I} = \{z \mapsto \mathbb{1}_{z \leq t} \,:\, t \in \mathbb{R}\}
$$

Lemma 31. *For all $\epsilon \in (0, 1/2)$ and any probability measure Q, we have that $N_{[]}(\mathcal{I}, L_1(Q), 2\epsilon) \leq 1/\epsilon$*

Proof. Let F be the cumulative distribution function of measure F. Define $F^-(u) = \inf\{t \in \mathbb{R} \,:\, F(t) \geq u\}$ and recall that $F_- \circ F^- \leq id \leq F \circ F^-$. Let $\epsilon \in (0, 1)$. Let $u_k = \epsilon k$, $k = 0, 1 \ldots, N-1$ with N being smallest integer such that $N\epsilon \geq 1$. Consequently, $N = \lceil 1/\epsilon \rceil$. Define $t_i = F^-(u_i), i = 0, \ldots, N-1$ and set $t_N = \infty$. The associated brackets, for any $i = 1, \ldots, N$

$$
[\mathbb{1}_{(-\infty, t_{i-1}]}, \mathbb{1}_{(-\infty, t_i)}].
$$

One has, for any i,

$$Q(\mathbb{1}_{(-\infty,t_i)} - \mathbb{1}_{(-\infty,t_{i-1}]}) = F_-(t_i) - F(t_{i-1}) \leq \epsilon$$

The number of brackets is $N = \lceil 1/\epsilon \rceil$. Then for all $\epsilon \in (0,1)$, $N_{[]}(\mathcal{I}, L_1(Q), \epsilon) \leq 1/\epsilon + 1 \leq 2/\epsilon$, for all probability measure Q on $\mathbb{R}$. $\qquad\square$

Interestingly, it holds that (this is one of the exercises)

$$\mathcal{N}(\mathcal{G}, L_p(Q), \epsilon) \leq \mathcal{N}_{[]}(\mathcal{G}, L_p(Q), 2\epsilon).$$

As a consequence, we deduce the following statement, that the class of cells is VC with parameter (Van Der Vaart and Wellner, 1996, Example 2.5.4).

Lemma 32. *For any probability measure Q, we have for all $\epsilon \in (0, 1/2)$, $\mathcal{N}(\mathcal{I}, L_1(Q), \epsilon) \leq 1/\epsilon$ and, for all $\epsilon \in (0, 1/\sqrt{2})$, $\mathcal{N}(\mathcal{I}, L_2(Q), \epsilon) \leq 1/\epsilon^2$.*

The previous might be refined in $\lceil 1/(2\epsilon^2) \rceil$. Also it might be extended to the multidimensional case constructing multidimensional brackets using product of 1-dimensional bracket. We obtain that $\mathcal{N}_{[]}(\mathcal{I}, L_2(Q), \epsilon)$ is of order $(1/\epsilon^2)^d$ when the dimension is $d \geq 1$. A similar bound shall be given relying on shattering numbers introduced next.

6.3.2 Parametric classes

The following result is well-known and proved for instance in Zhang (2023) (Theorem 5.3). It relies on Packing numbers, a concept similar to covering number, which we do not introduce here.

Lemma 33. *For any norm $\| \cdot \|$ on $\mathbb{R}^d$. We have, for all $\epsilon \in (0, r]$,*

$$\mathcal{N}(B(0, r), \| \cdot \|, \epsilon) \leq (1 + 2r/\epsilon)^d$$

As a consequence, for all $\epsilon \in (0, 1]$

$$\mathcal{N}(B(0, r), \| \cdot \|, \epsilon r) \leq (3/\epsilon)^d$$

Now now consider the parametric class

$$\mathcal{F}_\Theta := \{f_\theta - f_{\theta_0} : \theta \in \Theta\}$$

Lemma 34. *Suppose that Θ is bounded and that $\theta \mapsto f_\theta(x)$ is L-Lipschitz where L depends on x. Then for all $\epsilon \in (0, 1]$,*

$$\mathcal{N}(\mathcal{F}_\Theta, L_2(Q), \epsilon r_\Theta \|L\|_{L_2(Q)}) \leq (3/\epsilon)^d$$

where r_Θ is the radius of the smallest ball with center θ_0 that contains Θ.

In the above, note that $r_\Theta L$ is a proper envelop for the class of interest.

6.3.3 Indicator of balls and hyperplanes

The Vapnik-Chervonenkis dimension turns out to be related to the VC property. Roughly speaking, indicator functions of sets in $\mathcal{A}$ with finite Vapnik-Chervonenkis dimension are of VC type with in particular $v = 2vc(\mathcal{A})$. This is stated in the next Lemma which is due to Haussler (1995); see also Lemma 13.6 in Boucheron et al. (2013).

Lemma 35 (Haussler). *For any probability measure Q, we have, for all $\epsilon \in (0, 1)$,*

$$\mathcal{N}(\mathbb{1}_{\mathcal{A}}, L_1(Q), \epsilon) \leq (e^4/\epsilon)^{vc(\mathcal{A})}$$
$$\mathcal{N}(\mathbb{1}_{\mathcal{A}}, L_2(Q), \epsilon) \leq (e^2/\epsilon)^{2vc(\mathcal{A})}$$

and the centres of the covering are elements of $\mathbb{1}_{\mathcal{A}}$.

The previous lemma is actually really useful as many popular classes of sets have a known finite Vapnik dimension as explained in Section 4. We have, for cells $\mathcal{C} = \{\prod_{k=1}^{d}(-\infty, b_k] : (b_1, \ldots, b_d) \in \mathbb{R}^d\}$ and rectangles $\mathcal{R} = \{\prod_{k=1}^{d}(-a_k, b_k]\}$, we have

$$\mathcal{N}(\mathbb{1}_{\mathcal{C}}, L_2(Q), \epsilon) \leq (e^2/\epsilon)^{2d} \qquad \mathcal{N}(\mathbb{1}_{\mathcal{R}}, L_2(Q), \epsilon) \leq (e^2/\epsilon)^{4d}.$$

Similarly, the class of balls defined as, $\mathcal{B} = \{B(x, r) : x \in \mathbb{R}^d, r > 0\}$ has Vapnik dimension equal to $d + 1$. As a consequence,

$$\mathcal{N}(\mathbb{1}_{\mathcal{B}}, L_2(Q), \epsilon) \leq (e^2/\epsilon)^{2(d+1)}.$$

We also have the following result about collections of hyperplanes $\mathcal{H} = \{\langle w, x \rangle > 0 : w \in \mathbb{R}^d\}$, that is

$$\mathcal{N}(\mathbb{1}_{\mathcal{H}}, L_2(Q), \epsilon) \leq (e^2/\epsilon)^{2d}.$$

6.3.4 Kernel functions

In the previous section, the Vapnik dimension, a property related to subsets, has been employed to deal with particular class of functions defined as indicators of sets. Here we consider general class of functions (not necessarily indicators of sets) and we show that the VC property can still be useful. The relevant property for functions is related to the Vapnik dimension of their associated subgraph.

Definition 36. *A function class $\mathcal{F}$ is called VC-subgraph whenever the class $\mathcal{C}$ of all its subgraphs $\mathcal{C}_f = \{(x, t) : t < f(x)\}$ has a finite Vapnik dimension $vc(\mathcal{C})$.*

Lemma 37. *Let $\mathcal{F}$ be a VC subgraph class, we have*

$$\mathcal{N}(\mathcal{F}, L_2(Q), \epsilon\|F\|_{L_2(Q)}) \leq (2e^2/\epsilon)^{2vc(\mathcal{C})}$$

Proof. The proof is adapted from Van Der Vaart and Wellner (1996) Theorem 2.6.7. We have

$$Q(|f - g|) = Q((f - g)1_f > g) + Q((g - f)1_f < g) = Q \times \lambda(1_{g \leq t < f} + 1_{f \leq t < g})$$

But the function $1_{g \leq t < f} + 1_{f \leq t < g}$ is equal to $1_{\mathcal{C}_f \Delta \mathcal{C}_g}$. Let $S = \mathcal{C}_F \Delta \mathcal{C}_{-F}$. One may check that $P(A) = Q \times \lambda(A \cap S)/2QF$ is a probability measure on the set $S = \{-F(x) \leq t < F(x)\}$. From

Lemma 35, there exists $\mathbb{1}_{A_k}$, $k = 1 \ldots N_\epsilon$, where A_k is a subgraph, a ϵ cover of $\mathcal{C}$. Let f_k be the associated function. We have

$$Q(|f - f_k|) = P(\mathcal{C}_f \Delta \mathcal{C}_k) 2QF \leq 2QF\epsilon$$

As a consequence,

$$\mathcal{N}(\mathcal{F}, L_1(Q), \epsilon 2QF) \leq \mathcal{N}(\mathbb{1}_\mathcal{C}, L_1(P), \epsilon) \leq (e^4/\epsilon)^{vc(\mathcal{A})}.$$

We have a result for $L_1(Q)$. To obtain it for $L_2(Q)$, note that

$$Q(f^2) = Q(2F)\tilde{Q}(f)$$

where $\tilde{Q} = Q/Q(2F)$ is another probability measure. Take f_k a $L_1(\tilde{Q})$ cover of size $\epsilon^2 Q(2F)$ of $\mathcal{F}$. As a consequence, $Q(f - f_k)^2 = \tilde{Q}(|f - f_k|)Q(2F) \leq \epsilon^2 Q(2F)^2 \leq \epsilon^2 Q((2F^2)$. Therefore,

$$\mathcal{N}(\mathcal{F}, L_2(Q), \epsilon\|2F\|_{L_2(Q)}) \leq \mathcal{N}(\mathcal{F}, L_1(\tilde{Q}), \epsilon^2 Q(2F)) \leq (e^4/\epsilon^2)^{vc(\mathcal{A})}.$$

$\square$

We conclude this section by another result which will be useful for kernel estimate. A more general result is in Giné and Nickl (2021) Proposition 3.6.12. This type of ideas were developed in Nolan and Pollard (1987).

Lemma 38. *Suppose that $K : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a decreasing function. Then the subgraphs of*

$$\mathcal{K} = \{x \mapsto K(\|h^{-1}(x - s)\|) : h > 0, s \in \mathbb{R}^d\}$$

have a Vapnik dimension $d + 3$. As a consequence

$$\mathcal{N}(\mathcal{K}, L_2(Q), \epsilon\|K\|_{L_2(Q)}) \leq (2e^2/\epsilon)^{2(d+3)}$$

Proof. The proof use that subgraphs of the previous might be written as

$$\|h^{-1}(x - s)\|^2 - K^{-1}(t)^2 > 0$$

but $\|h^{-1}(x - s)\|^2 = h^{-2}\|x\|^2 - 2h^{-2}\langle x, s \rangle - K^{-1}(t)^2 + h^{-1}\|s\|^2$ is included the space generated by $(x, t) \mapsto (\|x\|^2, x, K^{-1}(t)^2, 1)$. The dimension is $d + 3$ based on Lemma 19. The conclusion follows from Theorem 37.

$\square$

When h is considered fixed, the approach can be simplified in the real case as indicated in Giné and Nickl (2009).

Lemma 39. *Suppose that $K : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a decreasing function. Then the subgraphs of*

$$\mathcal{K} = \{x \mapsto K(x - s) : s \in \mathbb{R}\}$$

have a Vapnik dimension 2.

The previous results will be useful to obtain a uniform bound with respect to x for the KDE estimate.

6.4 Preservation properties

We next call an envelope for $\mathcal{G}$ any function $G : S \mapsto \mathbb{R}$ that satisfies $|g(x)| \leq G(x)$ for all $x \in S$ and $g \in \mathcal{G}$. We first describe some basic preservation property related to VC classes.

Lemma 40. *Suppose that $\mathcal{F}$ and $\mathcal{G}$ are both VC. Then $-\mathcal{F}$, $\mathcal{F} \cup \mathcal{G}$, $\mathcal{F} + \mathcal{G}$, $\mathcal{F}\mathcal{G}$ are VC*

Lemma 41. *Suppose that $\mathcal{F}$ (resp. $\mathcal{G}$) defined on $(S, \mathcal{S})$ is of VC-type with envelope U and parameter (v, A) and let $E \in \mathcal{S}$. The following holds:*

1. *$\{f I_E : f \in \mathcal{F}\}$ is of VC-type with envelope U and parameter (v, A),*

2. *$\mathcal{F} - \mathcal{G} = \{f - g : f \in \mathcal{F}, g \in \mathcal{G}\}$ is of VC-type with envelope $2U$ and parameter $(2v, 2A)$,*

3. *$\{f - P(f|E) : f \in \mathcal{F}\}$ is of VC-type with envelope $2U$ and parameter $(2v, A)$,*

4. *$\{qf + (1 - q)g : f \in \mathcal{F}, g \in \mathcal{G}, q \in [0, 1]\}$ is of VC-type with envelope U and parameter $(2v + 1, 3A)$.*

Proof. Let Q be a probability measure on $(S, \mathcal{S})$. Let $(f_k)_{k=1,\ldots,K}$ be the center of an ϵU-covering of $(\mathcal{F}, Q)$. The first statement follow from the fact that $\|f 1_E - f_k 1_E\|_{L_2(Q)} \leq \|f - f_k\|_{L_2(Q)}$. For the second statement consider $U\epsilon$-covers $(f_k, k \leq K)$ and $(g_j, j \leq J)$ respectively of $\mathcal{F}$ and $\mathcal{G}$. Then the triangle inequality shows that $(f_k - g_j), k \leq K, j \leq J$ forms a $2U\epsilon$-cover of $\mathcal{F} + \mathcal{G}$ and the result follows. Now let $(\tilde{f}_k)_{k=1,\ldots,K}$ be the center of an ϵU-covering of $(\mathcal{F}, P_E)$ with $P_E(\cdot) = P(\cdot|E)$. Consider the covering induces by the centers $(f_k - P_E(\tilde{f}_j))_{1 \leq k, j \leq K}$ made of K^2 elements. Suppose that $f \in \mathcal{F}$. Then there is k and j such that

$$\|(f - P_E(f)) - (f_k - P_E(\tilde{f}_j))\|_{L_2(Q)} \leq \|f - f_k\|_{L_2(Q)} + P_E(f - \tilde{f}_j)$$
$$\leq \|f - f_k\|_{L_2(Q)} + \|f - \tilde{f}_j\|_{L_2(P)}$$
$$\leq 2U\epsilon.$$

Hence we have found a $2U\epsilon$-covering of size K^2 which by assumption is smaller than $(A/\epsilon)^{2v}$. This implies the third statement of the lemma. For the last statement, let $\mathcal{H} = \{qf + (1 - q)g : f \in \mathcal{F}, g \in \mathcal{G}, q \in [0, 1]\}$. Let $(f_k)_{k=1,\ldots,K}$ (resp. $g_\ell, (g_\ell)_{\ell=1,\ldots,L}$) be the center of an ϵU-covering of $(\mathcal{F}, Q)$ (resp. $(\mathcal{G}, Q)$). Let q_i, $i = 1, \ldots, \lfloor 1/\epsilon \rfloor$ be an ϵ-covering of $[0, 1]$. Let $h = qf + (1 - q)g$ be such that $q \in [0, 1]$, $f \in \mathcal{F}$ and $g \in \mathcal{G}$. There is f_k, g_ℓ, q_i such that

$$\|qf + (1 - q)g - (q_i f_k + (1 - q_i)g_\ell)\|_{L_2(Q)}$$
$$\leq \|q(f - f_k) + (1 - q)(g - g_\ell)\|_{L_2(Q)} + \|(q - q_i)f_k + (q_i - q)g_\ell\|_{L_2(Q)}$$
$$\leq qU\epsilon + (1 - q)U\epsilon + \epsilon U + \epsilon U = 3\epsilon U.$$

Hence the element $(q_i f_k + (1 - q_i)g_\ell)$ form an $3\epsilon U$-covering in the space $L_2(Q)$. There are $KL\lceil 1/\epsilon \rceil \leq KL/\epsilon$ such elements. As a consequence, since $\mathcal{F}$ and $\mathcal{G}$ are are of VC-type, it follows that

$$\mathcal{N}(\mathcal{H}, L_2(Q), 3\epsilon U) \leq (A/\epsilon)^{2v}(1/\epsilon) \leq (A/\epsilon)^{2v+1}$$

which implies the stated result.

$\square$

Define the set

$$Star(\mathcal{G}) = \{\alpha g : \alpha \in [0, 1], g \in \mathcal{G}\}.$$

Proposition 42. *If $\mathcal{G}$ is of VC type with envelop $\|G\|$ and parameters (v, A) then $Star(\mathcal{G})$ is of VC type with envelop $\|G\|$ and parameter $(v + 1, A)$.*

Proof. Let Q be a probability measure on $(S, \mathcal{S})$ and $\epsilon \in (0, 1)$. Let $g_1, \ldots, g_J$ be a $\|G\|_{L_2(Q)} \epsilon/2$-cover of $\mathcal{G}$. By definition $J \leq (2A/\epsilon)^v$. Let $K = \lceil 1/\epsilon \rceil$. The following set made of K balls $[0, \epsilon], [\epsilon, 2\epsilon], \ldots, [(K-1)\epsilon, K\epsilon]$ is an $\epsilon/2$-cover of the space $[0, 1]$. The centres of these balls are $\alpha_k = (2k-1)\epsilon/2$, $k = 1, \ldots, K$. Consider the collection of functions $\{\alpha_k g_j, \ k = 1, \ldots, K, \ j = 1, \ldots, J\}$. Let $\alpha g \in Star(\mathcal{G})$. Then there is $\alpha_k g_j$ such that

$$\|\alpha g - \alpha_k g_j\|_{L_2(Q)} \leq \|(\alpha - \alpha_k)g\|_{L_2(Q)} + \|\alpha_k(g - g_j)\|_{L_2(Q)} \leq \epsilon \|G\|_{L_2(Q)}.$$

We have then found an $\epsilon \|G\|_{L_2(Q)}$-net of $(Star(\mathcal{G}), L_2(Q))$. The number of elements in the net is $(2A/\epsilon)^v K \leq (2A/\epsilon)^v(2/\epsilon) \leq (2A/\epsilon)^{v+1}$.

$\square$

6.5 Exercises

Exercise 16. *1. Show that*

$$\mathcal{N}(\mathcal{G}, L_2(Q), \epsilon) \leq \mathcal{N}_{[]}(\mathcal{G}, L_2(Q), 2\epsilon).$$

2. Suppose that $\mathcal{F}$ and $\mathcal{G}$ both are valued in $[0, 1]$. Show that

$$\mathcal{N}_{[]}(\mathcal{G}\mathcal{F}, L_2(Q), 2\epsilon) \leq \mathcal{N}_{[]}(\mathcal{F}, L_2(Q), \epsilon)\mathcal{N}_{[]}(\mathcal{G}, L_2(Q), \epsilon)$$

3. If $\mathcal{I}$ is the class of cells in dimension 2. Show that

$$\mathcal{N}_{[]}(\mathcal{I}, L_2(Q), 2\epsilon) \leq (2/\epsilon^2)^2$$

7 Local averaging rules

Local averaging is a common statistical approach which consists in "averaging" data restricted to some (often small) region of the space. This is contrasting with standard approach that would use the full data to estimate some unknown quantities depending on the underlying probability measure. Local averaging around a given point x is often useful in building estimator of "local" quantities that depends on the conditional probability measure given x, μ_x, instead of the common underlying measure μ. For instance, if one which to estimate the regression function $E[Y|X = x] = \int y \mathrm{d}\mu_x(y)$, then one could use the points among the initial sample that are nearby x by averaging them. Another example different from regression is density estimation as introduce in the next section.

7.1 Density estimation by kernel smoothing

In this section, we provide a uniform deviation result on the kernel estimate of the density $\hat{f}_n$ introduced in the first section. Results of this type can be found in Giné and Guillou (2001); Giné and Guillou (2002); Giné and Nickl (2009). We consider the following standard framework for density estimation.

(KDE) Let $(X, X_1, \ldots, X_n)$ be a sequence of independent and identically distributed random elements with common distribution P_X on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. We suppose that P_X has a density f with respect to Lebesgue measure, i.e., for all $B \in \mathcal{B}(\mathbb{R}^d)$, $P_X(B) = \int_B f \mathrm{d}\lambda$.

We wish to estimate f uniformly in $x \in \mathbb{R}^d$ using KDE $\hat{f}_n$ which has been shown to be a good point-wise estimator of f. Relying on the same decomposition as before, we have

$$\hat{f}_n(x) - f(x) = \hat{f}_n(x) - f_h(x) + f_h(x) - f(x).$$

To take care of the bias term, we consider the following regularity condition.

(Reg2) The function f is L smooth; that is, f is differentiable and $\nabla_x f(x)$ is L-Lipschitz

$$\|\nabla_x f(x) - \nabla_x f(y)\| \leq L\|x - y\|$$

The bias term has already been considered in a previous exercise but under a Lipschitz condition. Here we hope that the L-smoothness (which is stronger) will imply better convergence rate.

Theorem 43. *Suppose that (KDE) and (Reg2) are fulfilled. If*

$$n \geq nh^d V_d U_f \geq 16C^2(d+1)\log(4e^2/(\delta h^d V_d U_f))$$

then with probability $1 - \delta$,

$$\sup_{x \in \mathbb{R}^d} |\hat{f}_n(x) - f(x)| \leq 2\sqrt{\frac{16C^2(d+1)U_f \log(4e^2 n/\delta)}{nh^d V_d}} + \frac{1}{2}h^2 L$$

Proof. Lemma 53 is useful to obtain appropriate rate of convergence for the bias term. Because $\int_{B(0,1)} u du = 0$, we have that

$$f_h(x) - f(x) = V_d^{-1} \int_{B(0,1)} f(x-hu) - f(x) dx = V_d^{-1} \int_{B(0,1)} f(x-hu) - f(x) - hu^T \nabla_x f(x) dx$$

Therefore

$$|f_h(x) - f(x)| \leq V_d^{-1} \int_{B(0,1)} |f(x-hu) - f(x) - hu^T \nabla_x f(x)| dx \leq \frac{1}{2} V_d^{-1} h^2 L \int_{B(0,1)} \|u\|^2 dx$$

which gives

$$\sup_{x \in \mathbb{R}^d} |f_h(x) - f(x)| \leq \frac{1}{2} h^2 L$$

Now let us consider the variance part. We have

$$\hat{f}_n(x) - f_h(x) = n^{-1} h^{-d} V_d^{-1} \sum_{i=1}^n \{g_{x,h}(X_i) - P(g_{x,h})\}$$

with $g_{x,h}(X) = \mathbb{1}_{B(x,h)}(X)$. The class $\{g_{x,h} \ \ x \in \mathbb{R}^d, \ h > 0\}$ is VC because it is included in the set of indicator of balls. The parameter are $(e^2, 2(d+1))$. The class $\{g_{x,h} \ \ x \in \mathbb{R}^d\}$ is also VC because it is included in the previous class. Note that the VC parameter do not depend on h. Applying our main result, Theorem 29, we need to choose the variance value σ^2. We have

$$\sup_x P(g_{x,h}^2) = \sup_x P(g_{x,h}) \leq h^d V_d U_f = \sigma^2$$

so we need to assume that $h^d V_d U_f \leq 1$. If in addition

$$n h^d V_d U_f \geq 16 C^2 (d+1) \log(4e^2 / (\delta h^d V_d U_f))$$

we obtain that, with probability $1 - \delta$,

$$\sup_{x \in \mathbb{R}^d} |\hat{f}_n(x) - f_h(x)| \leq 2 \sqrt{\frac{16 C^2 (d+1) U_f \log(4e^2 / (\delta h^d V_d U_f))}{n h^d V_d}}$$

Note that since $h^d V_d U_f \leq 1$ it holds that $\log(4e^2 / (\delta h^d V_d U_f)) \geq 1$ and in particular $n h^d V_d U_f \geq 1$. Hence using this lower bound on h in previous upper bound we obtain the stated result.

□

To emphasize the strength of previous result, one may use it when $\delta = n^{-2}$ to obtain that whenever

$$n h^d / \log(n) \to \infty, \qquad h \to 0$$

we have, with probability 1, as $n \to \infty$,

$$\sup_{x \in \mathbb{R}^d} |\hat{f}_n(x) - f_h(x)| = O\left(\sqrt{\frac{\log(n)}{n h^d}} + h^2 \right).$$

Optimizing the above with respect to h gives $h = (\log(n)/n)^{1/(d+4)}$ which is the optimal choice of h for twice differentiable functions. If the density would have been only Lipschitz then we would

have obtained $h = (\log(n)/n)^{1/(d+2)}$. Each regularity scenario leads to different rates, $n^{-2/(d+4)}$ and $n^{-1/(d+2)}$, which are the optimal ones in the minimax sense Tsybakov (2008). More generally, when f is $\mathcal{C}^s$, the optimal rate of convergence is $(\log(n)/n)^{s/(d+2s)}$. The latter facts emphasizes the difficulty to select the bandwidth in practice as we do not know in general the regularity of the function. For choosing the bandwidth, adaptive methods aim to achieve the optimal choice without knowing the smoothness of the function. Such methods rely on bandwidth selecting procedure for which $\hat{h}_n$ depends on the sample. The results proposed before might be useful in that case because of the VC property of balls we have to show that $\hat{h}$ lies in the appropriate range so that the above bound goes to 0. For instance, as soon as the $\hat{h}_n$ is such that

$$n\hat{h}_n^d \to \infty \qquad \hat{h}_n \to 0$$

One might obtain the consistency of adaptive methods. This does not imply that they would be optimal. More involved results are that adaptive choices can be optimal with respect to the regularity of the associated function f. The previous estimator can be modified to estimate the regression function. Define

$$\hat{g}_n(x) = \frac{\sum_{i=1}^n Y_i \mathbb{1}_{B(x,h)}(X_i)}{\sum_{i=1}^n \mathbb{1}_{B(x,h)}(X_i)}$$

7.2 Nearest neighbor algorithm

Consider the following k-NN estimate $\mu_x(g) = \mathbb{E}[g(Y)|X = x]$ given by

$$\hat{\mu}_n(g) = k^{-1} \sum_{i \in N_{n,k}(x)} g(Y_i).$$

Some additional assumptions will be needed to control the size of the k-NN neighborhood.

(X1) There is $c > 0$ and $T > 0$ such that

$$\lambda(S_X \cap B(x,\tau)) \geq c\lambda(B(x,\tau)), \qquad \forall \tau \in (0,T], \forall x \in S_X,$$

where λ is the Lebesgue measure.

(X2) There is $0 < b_X \leq U_X < +\infty$ such that $b_X \leq f(x) \leq U_X$, for all $x \in S_X$.

It is useful to derive an upper bound on the k-NN k-NN radius. Define

$$\overline{\tau}_{n,k} = \left(\frac{2k}{nb_X cV_d}\right)^{1/d}.$$

The first above condition is interesting for small $\tau > 0$ and hence deals with the boundary of the set S_X while the second means that the measure P_X charges uniformly the set S_X (no region with no point).

The following Lemma controls the size of the k-NN balls uniformly over all $x \in S_X$. In this way, it extends the multiplicative Chernoff bound which focuses on a particular x. The proof is based on an auxiliary result, Theorem 52 given in the Appendix. We stress that a similar result (with lass advantageous constants) could be obtained using Theorem 29 instead of Theorem 52.

Lemma 44. *Suppose that (IID), (X1) and (X2) hold true. Then, for all $n \geq 1$, $\delta \in (0,1)$ and $1 \leq k \leq n$ such that $16d\log(12n/\delta) \leq k \leq T^d n b_X c V_d/2$, it holds, with probability at least $1 - \delta$:*

$$\sup_{x \in S_X} \hat{\tau}_{n,k,x} \leq \overline{\tau}_{n,k}.$$

Proof. The proof is similar the proof of Lemma 6. The difference is that we use Lemma 52, a consequence of second Vapnik inequality, instead of Lemma 1 so that the reasoning is uniform over x. We have that, for all $x \in S_X$,

$$P(B(x, \overline{\tau}_{n,k})) = \int_{S_X \cap B(x, \overline{\tau}_{n,k})} f_X(x) dx \geq bc\lambda(B(x, \overline{\tau}_{n,k})) = 2\frac{k}{n}.$$

Lemma 52 indicates that, with probability $1 - \delta$, for all $x \in S_X$,

$$n\mathbb{P}_n(B(x, \overline{\tau}_{n,k})) \geq nP(B(x, \overline{\tau}_{n,k})) - \sqrt{nP(B(x, \overline{\tau}_{n,k}))8d\log(12n/\delta)}$$

Noting that $x \mapsto x - \sqrt{x\ell}$ is increasing whenever $x \geq \ell/4$, we obtain, since $2k \geq (16/4)kd\log(12n/\delta)$,

$$n\mathbb{P}_n(B(x, \overline{\tau}_{n,k})) \geq 2k - \sqrt{16kd\log(12n/\delta)}.$$

The above quantity is larger than k whenever $k \geq \sqrt{16kd\log(12n/\delta)}$, equivalently, $k \geq 16d\log(12n/\delta)$. Now it remains to use that, for each x, $\hat{\tau}_{n,k,x}$ is the smaller such τ so it needs to be smaller than $\overline{\tau}_{n,k}$. □

In the standard regression setup, $g(Y) = Y$, as in second chapter and under assumption (E). The variance is given by

$$v_n(x) = k^{-1} \sum_{i=1}^n \epsilon_i \mathbb{1}_{B(x, \hat{\tau}_{n,k,x})}(X_i)$$

Conditionally on $X_1, \ldots, X_n$, the previous quantity is sub-Gaussian with factor $\sigma^2 \sum_{i=1}^n \epsilon_i \mathbb{1}_{B(x, \hat{\tau}_{n,k,x})}(X_i) = \sigma^2 k$. Since $B(x, \hat{\tau}_{n,k,x})$ is included in the set of balls in $\mathbb{R}^d$. Conditionally on $X_1, \ldots, X_n$, there is at most $(n+1)^v$, $v = d+1$, elements in the sets $(\mathbb{1}_{B(x, \hat{\tau}_{n,k,x})}(X_i))_{i=1,\ldots,n}$. Therefore we can apply Vapnik type inequality, Theorem 15, to obtain that

$$\mathbb{P}(\sup_{x \in S_X} |v_n(x)| > t | X_1, \ldots, X_n) \leq 2(n+1)^v \exp(-t^2 k/(8\sigma^2))$$

Taking the expectation and changing variables gives with probability at least $1 - \delta$,

$$\sup_{x \in S_X} |v_n(x)| \leq \sqrt{\frac{8\sigma^2 \log(2(n+1)^v)}{k}}.$$

We now state our main result, a non-asymptotic bound on the error associated to the nearest neighbor regression estimator. The proposed bound on the k-NN process should hold uniformly in $x \in S_X$ where the set S_X stands for the support of P_X.

Theorem 45. *Suppose that (IID), (X1), (X2) are fulfilled and that $x \mapsto \mu_x(g)$ is L-Lipschitz. Let $\gamma = U_g U_f / b_f$. Then whenever*

$$n \geq 2\gamma k \geq 16C^2(d+1)\log(8e^2n/\delta)$$

it holds with probability $1 - \delta$,

$$|\hat{\mu}_x(g) - \mu_x(g)| \leq 2\sqrt{\frac{16C^2(d+1)2\gamma}{k}\log(8e^2n/\delta)} + L\overline{\tau}_{n,k}$$

Proof. Note that the estimator of interest may be written as

$$\hat{\mu}_x(g) = k^{-1} \sum_{i=1}^{n} g(Y_i)\mathbb{1}_{B(x,\hat{\tau}_{n,k,x})}(X_i).$$

Write

$$\hat{\mu}_x(g) - \mu_x(g) = k^{-1} \sum_{i\in N_{n,k}(x)} (g(Y_i) - \mu_{X_i}(g)) + k^{-1} \sum_{i\in N_{n,k}(x)} (\mu_{X_i}(g) - \mu_x(g))$$

The second term in the left-hand side is smaller than

$$Lk^{-1} \sum_{i\in N_{n,k}(x)} \|X_i - x\| \le L\hat{\tau}_{n,k,x}$$

Using Lemma 44, we have that, with probability $1 - \delta$,

$$\hat{\tau}_{n,k,x} \le \overline{\tau}_{n,k}$$

For the first term in the left-hand side, we have that, with probability $1 - \delta$,

$$\sum_{i\in N_{n,k}(x)} (Y_i - g(X_i)) \le \sup_{x\in S_X, \tau \le \tau_{n,k}} \sum_{i=1}^{n} (g(Y_i) - \mu_{X_i}(g))\mathbb{1}_{B(x,\tau)}(X_i)$$

The above is a centred empirical process to which we apply the Theorem 29 with

$$\mathbb{E}[(g(Y) - \mu_X(g))^2 \mathbb{1}_{B(x,\tau)}(X)] \le \|\operatorname{var}(g(Y)|x)\|_\infty \tau^d V_d U_f \le U_g^2 U_f \overline{\tau}_{n,k}^d V_d = \sigma^2$$

Whenever $U_g^2 U_f \overline{\tau}_{n,k}^d V_d \le 1$ and

$$nU_g^2 U_f \overline{\tau}_{n,k}^d V_d \ge 16C^2(d+1)\log(4e^2/(\delta U_g^2 U_f \overline{\tau}_{n,k}^d V_d))$$

we find, with probability $1 - \delta$,

$$\sum_{i\in N_{n,k}(x)} (Y_i - g(X_i)) \le 2\sqrt{16C^2(d+1)nU_g^2 U_f \overline{\tau}_{n,k}^d V_d \log(4e^2/(\delta U_g U_f \overline{\tau}_{n,k}^d V_d))}$$

Using that $U_g^2 U_f \overline{\tau}_{n,k}^d V_d \le 1$ we get that $nU_g^2 U_f \overline{\tau}_{n,k}^d V_d \ge 1$ and therefore, with probability $1 - \delta$,

$$\log(4e^2/(\delta U_g^2 U_f \overline{\tau}_{n,k}^d V_d)) \le \log(4e^2 n/\delta)).$$

Conclude by taking $\delta/2$ in the above so that the two previous event are together realized with probability $1 - \delta$. $\square$

Similar to before we have the following corollary that whenever $k/n \to 0$ and $k/\log(n) \to \infty$, with probability 1,

$$|\hat{\mu}_x(g) - \mu_x(g)| = O\left(\sqrt{\frac{\log(n)}{k}} + \overline{\tau}_{n,k}\right)$$

Choice of k   The optimal choice of k is therefore of order $n^{2/d+2}$.

The variance   In low density region, the variance term is smaller than the variance term of Nadaraya-Watson. As a consequence, if a non-asymptotic bound would be obtained for Nadaraya-Watson, it would be interested to compare it with the one given before.

7.3 Choice of k by hold-out

The Hold-out procedure is based on choosing the optimal k as follows:

$$\hat{k}'_m \in \operatorname*{arg\,min}_{k=1,\ldots,n} P'_m (Y - \hat{h}_k)^2$$

where P'_m is the empirical measure based on another sample $(X'_1, Y'_1), \ldots, (X'_m, Y'_m))$ which is independent from the initial sample. The objective is that the performance $P(h - \hat{h}_{\hat{k}'_m})^2$ is similar to that of k^* given by $\min_k P(h - \hat{h}_k)^2$.

Theorem 46. *Suppose that Y is bounded by U. Then with probability* 1,

$$\mathbb{E}' P(h - \hat{h}_{\hat{k}'_m})^2 \le P(h - \hat{h}_{k^*})^2 + \sqrt{\frac{8U^4 \log(n)}{m}}$$

Proof. By definition of the conditional expectation

$$P(h - \hat{h}_{\hat{k}'_m})^2 - P(h - \hat{h}_{k^*})^2 = P(Y - \hat{h}_{\hat{k}'_m})^2 - P(Y - \hat{h}_{k^*})^2$$

and, in addition, we have $k^* \in \operatorname*{arg\,min}_k P(Y - \hat{h}_k)^2$. Therefore, using that $P'_m(Y - \hat{h}_{\hat{k}'_m})^2 \le P'_m(Y - \hat{h}_{k^*})^2$, we obtain

$$P(h - \hat{h}_{\hat{k}'_m})^2 - P(h - \hat{h}_{k^*})^2 \le (P - P'_m)(Y - \hat{h}_{\hat{k}'_m})^2 - (P'_m - P)(Y - \hat{h}_{k^*})^2$$

Take the expectation with respect to P'_m-sample, we get

$$\mathbb{E}'[P(h - \hat{h}_{\hat{k}'_m})^2] - P(h - \hat{h}_{k^*})^2 \le \mathbb{E}'[(P - P'_m)(Y - \hat{h}_{\hat{k}'_m})^2] \le \mathbb{E}'[\sup_{k=1,\ldots,n} (P - P'_m)(Y - \hat{h}_k)^2]$$

We can now use Lemma 13 to obtain that

$$\mathbb{E}'[\sup_k (P - P'_m)(Y - \hat{h}_k)^2] \le \sqrt{\frac{2s^2 \log(n)}{m}}$$

where s^2 stands for the sub-Gaussian factor of $Y - \hat{h}_k$ that we now derive. Since Y is bounded by U we have that

$$|Y - \hat{h}_k| \le 2U$$

which implies that $(h - \hat{h}_k)^2 \le 4U^2$ is sub-Gaussian with factor $4U^4$. Hence we obtain the stated result.

$\square$

Note that the previous result provides a slight improvement in the constant compared to Theorem 16.1 in Biau and Devroye (2015). This result is interesting because whenever m is large enough (m larger than $n^{4/(d+2)} \log(n)$), using that $P(h - \hat{h}_{k^*})^2 = O_P(n^{-2/d+2})$, we obtain that

$$\mathbb{E}' P(h - \hat{h}_{\hat{k}'_m})^2 = O_P(n^{-2/(d+2)})$$

which is optimal for Lipschitz function.

7.4 Tree construction

In this section, we consider partitioning estimate based on data driven and greedy construction of the partitioning. This type of construction is sequential and at each step, the partition is refined. Those estimates are refereed to as regression tree. Any tree construction provide a set of leaf $\mathcal{L} \subset S_X$ that forms a partition of S_X. The regression estimate is then given by, for each $x \in S_X$,

$$\frac{\sum_{i=1}^{n} Y_i \mathbb{1}_{L(x)}(X_i)}{\sum_{i=1}^{n} \mathbb{1}_{L(x)}(X_i)}$$

where $L(x)$ denotes the unique element of the partition that contain x.

A popular tree construction consists in the following: for each leaf L:

- choose (k, t), k is the variable to split (among the relevant candidate) and t is the threshold (among the relevant value), such that the resulting MSE is minimized

- Check for the stopping criterion

To apply the tools developed in previous sections, we need to have a particular stopping rule. This is important to obtain in the end the property that a leaf has enough points to control the variance and is small enough diameter to allow the bias to be small. The stopping rule that we consider is that the L threshold should be selected among thresholds that give a leaf diameter d_k larger than ℓh. The parameter ℓ is selected by the user. As a consequence, direction cannot be selected whenever its size d_k (size of the leaf along direction k) is smaller than $2\ell h$. A tree fully grown according to this rule is such that for all k, $\ell h \leq d_k \leq 2\ell h$. One can stop growing the tree whenever $\max_k d_k(L) \leq hL$. Given the previous rules, any direction and any threshold can be selected at each step.

Theorem 47. *For Lipschitz function, the rate of convergence of regression tree is of order* $1/\sqrt{n(\ell h)^d} + Lh$.

8 Fast rates in empirical risk minimization

In this section, we consider the following empirical risk minimization problem

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(Y_i, f(X_i)).$$

where the loss function is L-Lipschitz in g uniformly in y. That is we shall consider the following assumption.

(LIP) There is $L > 0$ such that for all $y \in \mathbb{R}$, $f \in \mathbb{R}$ and $f' \in \mathbb{R}$,

$$|\ell(y, f) - \ell(y, f')| \leq L|f - f'|$$

Note that the contraction principle (stated in Proposition 11), we have that

$$\mathrm{Rad}_n(\ell_{\mathcal{F}}) \leq 2L\mathrm{Rad}_n(\mathcal{F})$$

Our starting point is standard rate of CV in ERM. We have that, with probability 1,

$$P(\ell_{\hat{f}_n}) - P(\ell_{f^*}) \leq 2 \sup_{f \in \mathcal{F}} |(P_n - P)(\ell_f)|$$

In light of the results given previously, if the class $\mathcal{F}$ is VC, we have that

$$P(\ell_{\hat{f}_n}) - P(\ell_{f^*}) \simeq \sqrt{v\sigma^2 \log(4AU/\sigma^2)/n}$$

Next we consider fast rates under the so-called Bernstein's condition that $P(g^2) \leq BPg$ within the class of interest. We follow the paper by Bartlett, Bousquet and Mendelson (Bartlett et al., 2005) which is a modification of Massart's approach (Massart and Nédélec, 2006). I will first recall Theorem 3.3 in (Bartlett et al., 2005) and then show that whenever the covering numbers are suitably behaved we can obtain a bound on r^* (a fixed point that appear in the statement).

The following is taken from the main result in (Bartlett et al., 2005), Theorem 3.3. Next the point that $T(g)$ is between $\mathrm{var}(g)$ and $BP(g)$ is important.

Theorem 48. *Let $\mathcal{G}$ be a class of function with range $[a, b]$. Suppose that $\mathrm{var}(g) \leq T(g) \leq BPg$ holds for all g. Then with probability $1 - \delta$:*

$$\forall g \in \mathcal{G}, \qquad Pg \leq \frac{K}{K-1}P_n g + c_1 K r^*/B + \log(1/\delta)(11(b-a) + c_2 BK)/n$$

where r^* is the fixed point of $\psi(r)$ a sub-root function such that for all $r \geq r^*$,

$$\psi(r) \geq \frac{B}{n}\mathbb{E}\mathrm{Rad}_n(\mathcal{G} : T(g) \leq r)$$

Aim is to apply this result using the tools developed previously. We can apply the previous result to the ERM problem. We consider $\mathcal{G}$ made with functions $g(x, y) = \ell(y, f(x))$. We will consider the next assumption that is related to some convexity property.

(CONV) There is $B > 0$ such that for all $f \in \mathcal{F}$

$$P((f - f^*)^2) \leq BP((\ell_f - \ell_{f^*}))$$

As a consequence of the Lipschitz condition, the Bernstein condition is met:

$$P((\ell_f - \ell_{f^*})^2) \leq L^2 P((f - f^*)^2) \leq BL^2 P((\ell_f - \ell_{f^*}))$$

In this case, we can choose $T(\ell_f) = L^2 P((f - f^*)^2)$ to apply Theorem 48. It gives the following statement.

Theorem 49. *Under (CONV) and (LIP),*

$$P\ell_{\hat{f}} - \ell_{f^*} \leq (c_1 K/B)r^* + \log(1/\delta)(22U + c_2 BK)/n$$

where r^* is the fixed point of $\psi(r)$ a sub-root function such that for all $r \geq r^*$,

$$\psi(r) \geq \frac{B}{n}\mathbb{E}\mathrm{Rad}_n(\ell_{\mathcal{F}} \; : \; L^2 P((f - f^*)^2) \leq r).$$

Now it remains to give a suitable bound on the function $\mathbb{E}\mathrm{Rad}_n(\ell_{\mathcal{F}} \; : \; L^2 P((f - f^*)^2) \leq r)$. By contraction, with $\varphi(f) = \ell_f - \ell_{f^*}$, we have

$$\mathbb{E}\mathrm{Rad}_n(\ell_{\mathcal{F}} - \ell_{f^*} \; : \; L^2 P((f - f^*)^2) \leq r) \leq L\mathbb{E}\mathrm{Rad}_n(\mathcal{F} \; : \; L^2 P((f - f^*)^2) \leq r)$$

and note also that

$$\mathbb{E}\mathrm{Rad}_n(\mathcal{F} \; : \; L^2 P((f - f^*)^2) \leq r) = \mathbb{E}\mathrm{Rad}_n(\mathcal{F} - f^* \; : \; L^2 P((f - f^*)^2) \leq r)$$

In light of the previous results on Rademacher with VC class ($\mathcal{F} - f^*$ is still VC) we would like to have P_n instead of P in the above restriction. This is the topic of next lemma.

Lemma 50. *Suppose that $\mathcal{F}$ is VC class with parameter (v, A) and envelope U. If $rn \geq \gamma U^2 \log(4An)$ (for some γ large enough), we have, with probability $1 - (1/n)$,*

$$\{f \in \mathcal{F} \; \; P(f^2) \leq r\} \subset \{f \in \mathcal{F} \; \; P_n(f^2) \leq 2r\}$$

As a consequence,

$$\mathbb{E}\mathrm{Rad}_n\{f \in \mathcal{F} \; \; P(f^2) \leq r\} \leq C\sqrt{nrv \log(4An)} + U$$

where C is an absolute constant.

Proof. Write $P_n(f^2) = P(f) + (P_n - P)f \leq r + \sup_{f \in \mathcal{F}}(P_n - P)(f^2)$. By Bousquet's inequality (here the variance of f^2 is bounded by U^2 times r)

$$\sup_{f \in \mathcal{F}}(P_n - P)(f) \leq (4/n)\mathbb{E}\mathrm{Rad}_n(\mathcal{F}^2) + \sqrt{2nU^2 r \log(1/\delta)} + 8U^2 \log(1/\delta)/3.$$

Contraction principle gives

$$\sup_{f \in \mathcal{F}}(P_n - P)(f) \leq (8U/n)\mathbb{E}\mathrm{Rad}_n(\mathcal{F}) + \sqrt{2U^2 r \log(1/\delta)/n} + 8U^2 \log(1/\delta)/(3n)$$

which because $\mathcal{F}$ is VC, is bounded by (Theorem 26),

$$8C\sqrt{U^2 rv \log(4AU/\sqrt{r})/n} + \sqrt{2U^2 r \log(1/\delta)/n} + 8U^2 \log(1/\delta)/(3n)$$

but since $r \geq U^2/n$, and taking $\delta = 1/n$, we find the upper bound

$$8C\sqrt{U^2 rv \log(4A\sqrt{n})/n} + \sqrt{2U^2 r \log(n)/n} + 8U^2 \log(n)/(3n)$$
$$\leq (8C + \sqrt{2})\sqrt{U^2 rv \log(4An)/n} + 8U^2 \log(4An)/(3n)$$

Now use that $r \geq \gamma U^2 \log(4An)/n$ to finally obtain that the latter upper bound is smaller than $2r$. The first result follows. For the second result, we have that

$$\mathbb{E}\mathrm{Rad}_n\{f \in \mathcal{F} \; P(f^2) \leq r\} \leq \mathbb{E}\mathrm{Rad}_n\{f \in \mathcal{F} \; P_n(f^2) \leq 2r\} + (nU)(1/n)$$

and we can use the bound for Rademacher over VC class. $\qquad\square$

Based on the previous lemma, we find that whenever $rn \geq U^2\gamma \log(4An)$,

$$\mathbb{E}\mathrm{Rad}_n(\mathcal{F} - f^* : L^2 P((f - f^*)^2) \leq r) \leq CL\sqrt{nrv \log(4An)} + U$$
$$\leq CL\sqrt{nrv \log(4An)} + U + \alpha$$

with a different constant C. The function ψ from above theorem is $\psi(r) = CL\sqrt{n^{-1}rv \log(4An)} + (U + \alpha)/n$ is sub-root. We check that the fixed point r^* is unique and given by

$$\sqrt{r^*} = (B + \sqrt{B^2 + 4(U + \alpha)/n}))/2$$

with $B = CL\sqrt{n^{-1}v \log(4An)}$. Note also that $r^* \geq \alpha/n$ so we can choose $\alpha = U^2\gamma \log(4An)$. Then whenever $r \geq r^*$ we have that $r \geq U^2\gamma \log(4An)/n$ so that

$$n^{-1}\mathbb{E}\mathrm{Rad}_n(\mathcal{F} - f^* : L^2 P((f - f^*)^2) \leq r) \leq \psi(r)$$

Therefore, we obtain the following statement.

Theorem 51. *Suppose that $\mathcal{F}$ is VC class and ℓ_f is L-Lipschitz and B-convex. We have with probability $1 - \delta$,*

$$P\ell_{\hat{f}} - \ell_{f^*} \leq Cn^{-1}v \log(4An/\delta)$$

where C depends on B and L.

Proof. Applying the previous results we get

$$P\ell_{\hat{f}} - \ell_{f^*} \leq (C/B)n^{-1}v \log(4An) + U^2 \log(4An)/n) + \log(1/\delta)(22U + c_2 BK)/n$$

which may be factorized to obtain the stated result.

$\qquad\square$

9 Auxiliary results

Let us conclude the section with an extended version of the multiplicative Chernoff inequality which holds uniformly over B when restricted to the collection of closed balls

$$\mathcal{B} = \{B(x, \tau) \; : \; x \in \mathbb{R}^d, \; \tau > 0\},$$

where $B(x, \tau)$ is the closed ball with center x and radius τ. It takes different forms in the literature such as Theorem 2.1 in Anthony and Shawe-Taylor (1993); Theorem 15 in Chaudhuri and Dasgupta (2010); Theorem 1 in Goix et al. (2015) and more recently Corollary 4.4 in Lhaut et al. (2022). Using the bound in Anthony and Shawe-Taylor (1993), combined with the complexity results on the set of closed balls provided in Wenocur and Dudley (1981), we obtain the following statement.

Lemma 52. *Let $(X_i)_{i \geq 1}$ be a sequence of independent and identically distributed random variables valued in $\mathbb{R}^d$ with common distribution P. For any $\delta > 0$ and $n \geq 1$, with probability at least $1 - \delta$:*

$$n^{-1} \sum_{i=1}^{n} \mathbb{1}_B(X_i) \geq P(B) \left(1 - \sqrt{\frac{8d \log(12n/\delta)}{nP(B)}} \right), \qquad \forall B \in \mathcal{B}.$$

Proof. Using Theorem 16 and that $S_\mathcal{A}(n) \leq (n+1)^V$ where V is the Vapnik dimension of $\mathcal{A}$ (see for instance Corollary 1.3 in Lugosi (2002)) and the fact that the set of closed balls has a Vapnik dimension equal to $d + 1$ as shown before, we find that $S_\mathcal{B}(2n) \leq (2n+1)^{d+1}$. We obtain that with probability at least $1 - \delta$

$$n^{-1} \sum_{i=1}^{n} \mathbb{1}_B(X_i) \geq P(B) \left(1 - \sqrt{\frac{4(d+1) \log(4(2n+1)/\delta)}{nP(B)}} \right), \qquad \forall B \in \mathcal{B}.$$

Using that $4(2n+1) \leq 12n$ and $4(d+1) \leq 8d$ leads to the result. $\qquad\square$

Lemma 53. *Under (Reg2), we have*

$$\|f(x) - f(y) - (x-y)^T \nabla f(y)\| \leq \frac{1}{2} L \|x - y\|^2$$

Proof. Consider the function $F(t) = f(tx + (1-t)y)$, $t \in [0, 1]$. We have that

$$f(x) - f(y) = \int_0^1 F'(t)\mathrm{d}t = (y-x)^T \int_0^1 \nabla f(tx + (1-t)y)\mathrm{d}t$$

Consequently

$$f(x) - f(y) - (x-y)^T \nabla f(y) = (y-x)^T \int_0^1 (\nabla f(tx + (1-t)y) - \nabla f(y))\mathrm{d}t$$

and we conclude using the L-smoothness. $\qquad\square$

Lemma 54. *Let $a > 0$ and $b > 0$. Suppose that $z^2 \leq az + b$, then $z \leq a + \sqrt{b}$.*

Proof. Solve $z^2 - az - b = 0$. There is two roots and only the positive one, $1/2(a + \sqrt{a^2 + 4b})$ is of interest. We have that $z \leq 1/2(a + \sqrt{a^2 + 4b}) \leq a + \sqrt{b}$. $\qquad\square$

References

Anthony, M. and J. Shawe-Taylor (1993). A result of Vapnik with applications. *Discrete Appl. Math. 47*(3), 207–217.

Bartlett, P., O. Bousquet, and S. Mendelson (2005). Local rademacher complexities. *Ann. Statist. 33*(4), 1497–1537.

Biau, G. and L. Devroye (2015). *Lectures on the nearest neighbor method.* Springer Series in the Data Sciences. Springer, Cham.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities. A nonasymptotic theory of independence.* Oxford University Press, Oxford.

Bousquet, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris 334*(6), 495–500.

Bousquet, O., S. Boucheron, and G. Lugosi (2004). *Introduction to Statistical Learning Theory*, Volume Lecture Notes in Artificial Intelligence 3176, pp. 169–207. Heidelberg, Germany: Springer.

Chaudhuri, K. and S. Dasgupta (2010). Rates of convergence for the cluster tree. In *NeurIPS proceedings*, Volume 23, pp. 343–351.

Cover, T. and P. Hart (1967). Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory 13*(1), 21–27.

Devroye, L., L. Györfi, and G. Lugosi (2013). *A probabilistic theory of pattern recognition*, Volume 31. Springer Science & Business Media.

Fix, E. and J. L. Hodges (1951). Discriminatory analysis. nonparametric discrimination: Consistency properties. *Int. Stat. Rev. 57*(3), 238–247.

Giné, E. and A. Guillou (2001). On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. *Ann. Inst. H. Poincaré Probab. Statist. 37*(4), 503–522.

Giné, E. and A. Guillou (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. Henri Poincaré Probab. Stat. 38*(6), 907–921. En l'honneur de J. Bretagnolle, D. Dacunha-Castelle, I. Ibragimov.

Giné, E. and R. Nickl (2009). An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation. *Probab. Theory Related Fields 143*(3-4), 569–596.

Giné, E. and R. Nickl (2021). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press.

Goix, N., A. Sabourin, S. Clémen, et al. (2015). Learning the dependence structure of rare events: a non-asymptotic study. In *Conference on Learning Theory*, pp. 843–860. PMLR.

Györfi, L., M. Kohler, A. Krzyzak, and H. Walk (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Hagerup, T. and C. Rüb (1990). A guided tour of chernoff bounds. *Information processing letters 33*(6), 305–308.

Haussler, D. (1995). Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A 69*(2), 217–232.

Ledoux, M. and M. Talagrand (1991). *Probability in Banach Spaces: isoperimetry and processes*, Volume 23. Springer Science & Business Media.

Lhaut, S., A. Sabourin, and J. Segers (2022). Uniform concentration bounds for frequencies of rare events. *Statistics & Probability Letters 189*, 109610.

Lugosi, G. (2002). Pattern classification and learning theory. In *Principles of nonparametric learning*, pp. 1–56. Springer.

Massart, P. (2000). About the constants in talagrand's concentration inequalities for empirical processes. *The Annals of Probability 28*(2), 863–884.

Massart, P. and É. Nédélec (2006). Risk bounds for statistical learning. *The Annals of Statistics*, 2326–2366.

Nolan, D. and D. Pollard (1987). U-processes: rates of convergence. *Ann. Statist. 15*(2), 780–799.

Rio, E. (2002). Une inégalité de bennett pour les maxima de processus empiriques. In *Annales de l'IHP Probabilités et statistiques*, Volume 38, pp. 1053–1057.

Royall, R. M. (1966). *A class of non-parametric estimate of a smooth regression.* ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–Stanford University.

Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *Ann. Probab. 22*(1), 28–76.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math. 126*(3), 505–563.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation* (1st ed.). Springer Publishing Company, Incorporated.

Van de Geer, S. A. et al. (2016). *Estimation and testing under sparsity.* Springer.

Van Der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics.* Springer Series in Statistics. New York: Springer-Verlag.

Van Der Vaart, A. W. and J. A. Wellner (2007). Empirical processes indexed by estimated functions. In *Asymptotics: particles, processes and inverse problems*, Volume 55 of *IMS Lecture Notes Monogr. Ser.*, pp. 234–252. Inst. Math. Statist.

Van Handel, R. (2014). Probability in high dimension. *Lecture Notes (Princeton University).*

Vapnik, V. N. and A. Y. Chervonenkis (2015). On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pp. 11–30. Springer, Cham. Reprint of Theor. Probability Appl. 16 (1971), 264–280.

Wenocur, R. S. and R. M. Dudley (1981). Some special Vapnik-Chervonenkis classes. *Discrete Math. 33*(3), 313–318.

Zhang, T. (2023). *Mathematical analysis of machine learning algorithms.* Cambridge University Press.