

NYILATKOZAT

Név: Tamás Ambrus

ELTE Természettudományi Kar, szak: alkalmazott matematikus

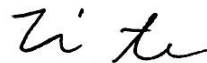
NEPTUN azonosító: G1LJXP

Szakdolgozat címe:

Kernel-Based Classification Algorithms and Their Stochastic Guarantees

A **szakdolgozat** szerzőjeként fegyelmi felelősségem tudatában kijelentem, hogy a dolgozatom önálló szellemi alkotásom, abban a hivatkozások és idézések standard szabályait következetesen alkalmaztam, mások által írt részeket a megfelelő idézés nélkül nem használtam fel.

Budapest, 2020. május 28.



a hallgató aláírása

EÖTVÖS LORÁND UNIVERSITY

FACULTY OF SCIENCE

Kernel-Based Classification Algorithms and Their Stochastic Guarantees

Thesis

Author:

Ambrus Tamás

Applied Mathematics MSc

Supervisors:

Balázs Csanád Csáji

Senior Research Fellow

Institute for Computer Science and Control (SZTAKI)

Ágnes Backhausz

Assistant Professor

Eötvös Loránd University (ELTE)



Budapest, 2020

Acknowledgements

I am very grateful to Balázs Csanád Csáji and Ágnes Backhausz for supervising me. This thesis improved a lot by their valuable guidance and useful suggestions. I really appreciate all the advice that I received during the preparation process. They helped me a lot to develop a deep understanding of this field.

I also thank my family, my girlfriend and my friends for their constant support in this period of my life and bringing daily joy.

Contents

Introduction	1
1 Foundations of Statistical Learning	3
1.1 Binary Classification	3
1.2 The Regression Function	5
1.3 Classification vs Regression	6
2 Formal Learning Models	8
2.1 Empirical Risk Minimization	8
2.2 Bias-Variance Trade-Off	9
2.3 Parametric vs Nonparametric Statistics	10
2.4 Learning Theory	11
2.5 Uniform Convergence	14
2.6 The VC Dimension	15
2.7 The Fundamental Theorem of PAC Learning	18
2.8 Structural Risk Minimization	22
2.9 Uniform Law of Large Numbers	23
2.9.1 Frequencies to Their Probabilities	24
2.9.2 Means to Their Expectations	25
2.10 Strong Uniform Laws	27
2.11 Universal Consistency	28
3 Nonparametric Methods	30
3.1 Local Averaging Estimates	30
3.2 Stone's Theorem	31
3.3 Kernel Estimates	32
3.4 Universal Consistency	32
3.5 Strong Consistency	35
3.6 The k-Nearest Neighbors Estimate	40
4 Kernel Methods	42
4.1 Ridge Regression	42
4.2 Support Vector Machines	43
4.3 Reproducing Kernel Hilbert Spaces	46

4.4	Representer Theorem	48
4.5	Kernel Mean Embedding	49
5	Confidence Regions	51
5.1	Resampling Framework	51
5.2	Non-Asymptotic Confidence Regions	54
5.3	Algorithm I (ERM Based)	56
5.4	Numerical Experiments I	62
5.5	Algorithm II (Local Averaging Based)	64
5.6	Algorithm III (Embedding Based)	66
5.7	Algorithm IV (Discrepancy Based)	68
5.8	Numerical Experiments II	70
	Conclusion	72
	Appendix	73
A	Tail and Concentration Inequalities	73
A.1	Deriving the Chernoff Bound	73
A.2	Sub-Gaussian Variables	74
A.3	Hoeffding's Inequality	76
A.4	Generalization to Martingale Differences	77
B	Proofs	79
B.1	A Bayes Optimal Classifier	79
B.2	A Uniform Exponential Bound	80
B.3	Stone's Theorem	82
B.4	Banach–Steinhaus Theorem for Integral Operators	85
B.5	Covering Lemma	86
	Bibliography	88

Introduction

Machine learning is a rapidly evolving field with a wide range of real-life applications, hence its mathematical foundation is extremely important. Since learning theory has its roots in statistics and computer science, the theoretical study of this area has a long and rich history.

Supervised learning problems are probably the most well-known examples of this field. The goal in this area is to understand observation generating mechanisms with the help of statistical samples and infer models with good generalizing property and design algorithms that can be used in practice. Besides, we would like to ensure the applicability of these learning methods with, preferably non-asymptotic and distribution-free, stochastic guarantees.

In our reasearch we studied binary classification and aimed at estimating the underlying regression function, which is the conditional expectation of the class labels given the inputs. The regression function is a key component of the Bayes optimal classifier, because it does not only provide optimal predictions, but also the risk of misclassification can be computed from it. We aimed at building non-asymptotic confidence regions for the regression function and suggested an empirical risk minimization based and three kernel-based semi-parametric resampling algorithms. Chapter 5 contains the new results of our reasearch, where the four aforementioned methods are introduced and the asymptotical analysis of the algorithms are also presented. It is proved that they are all strongly consistent in some sense.

I had two purposes in my mind throughout the writing process of this thesis. First, my goal was to introduce our new methods and present the corresponding results. These are included in Chapter 5, but since we build on many ideas from statistical learning theory I tried to provide a brief introduction to the applied concepts as well. My second goal was to give a comprehensive summary of the most important tools and theoretical results regarding the theory of supervised learning. However, since this field is huge and rapidly evolving, in most cases I chose to focus only on those materials that are related to our new findings.

Many sources were used during the preparation of this thesis. The book of Vladimir Vapnik [23], the book of Trevor Hastie et al. [14] and the book of Shai Shalev-Shwartz and Shai Ben-David [19] were good starting points. The books of László Györfi et al., [8] and [11], offered a high level analysis of nonparametric and distribution-free supervised learning concepts.

In Chapter 1 we review the most important results corresponding to the problem of classification and regression analysis. We prove that the *regression function* is a key object to examine for both problems. My main sources for this chapter were the aforementioned books of Györfi et al., [8], [11] and the book of Vapnik, [23].

In Chapter 2 we define a formal learning model and state the fundamental theorem of

probably approximately correct (PAC) learning which is closely related to the VC dimension of model classes. We present the structural risk minimization principle and two uniform laws of large numbers with the help of complexity measures. We also derive a no-free-lunch theorem which says that there is no universal PAC learner algorithm. In the end of this chapter a lighter notion, universal consistency is considered which can be reached for all regression functions with nonparametric estimates. The following books were the main sources for this part of the thesis: [8], [11], [19], [23] and [25].

The goal of Chapter 3 is to derive a strongly consistent method for regression functions with the help of nonparametric local averaging estimates. Stone's theorem is applied to show that a broad class of kernel estimates is universally consistent. Then a long reasoning is presented to show that strong consistency can be reached for many distributions of the examined sample. Beside the local averaging kernel estimates the k-nearest neighbors approach is defined. These estimation techniques are applied in our methods in Chapter 5. Mainly the books of Györfi et al. were used for this chapter, see [8] and [11].

General kernel methods are motivated and presented in Chapter 4 based on [14], [17], [18], [24] and [25]. First, the ridge regression estimate and support vector machines are introduced, then the theory of reproducing kernel Hilbert spaces and kernel mean embeddings are presented. These techniques are applied in our new algorithms in Chapter 5.

Finally, Chapter 5 contains our new results and algorithms for constructing exact, non-asymptotic confidence regions for the true underlying regression function. First, the resampling framework is introduced, which is similar to Monte Carlo tests and bootstrap methods. Second, a general construction scheme is defined and a non-asymptotic, exact guarantee is inferred for these confidence regions under mild statistical conditions. Then four concrete algorithms are introduced. The first one is based on empirical risk minimization. We provide uniform bounds on its asymptotic behaviour. The other three methods use kernel-based techniques and are strongly consistent. Algorithm II uses local averaging kernel estimates in the construction, while Algorithm III and IV are built on kernel methods and kernel mean embeddings. In the end of this chapter we illustrate our algorithms with numerical examples.

Appendix A is a short summary of tail bounds and concentration inequalities, that are used throughout the thesis. It is mainly based on [11] and [25]. In Appendix B some important proofs are presented for the sake of completeness.

Chapter 1

Foundations of Statistical Learning

1.1 Binary Classification

Classification or pattern recognition is one of the principle problems in *statistical learning theory* [23]. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space as always and let $(\mathbb{X} \times \mathbb{Y}, \mathcal{X} \otimes \mathcal{Y})$ denote a measurable sample space. We call \mathbb{X} the input space and \mathbb{Y} the output space. In classification \mathbb{X} can be an abstract high dimensional space, but in this thesis we often assume that $\mathbb{X} \subseteq \mathbb{R}^n$, while \mathbb{Y} in classification is always a finite set. In this thesis we only deal with the binary case, i.e. $\mathbb{Y} = \{+1, -1\}$. In classification a sample is given, $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$, from the $(X, Y) : \Omega \rightarrow \mathbb{X} \times \mathbb{Y}$ random pair's unknown joint distribution, which is denoted by $P_{X,Y}$ or simply by P . In the whole thesis we assume that $\{(X_i, Y_i)\}_{i=1}^n$ is an i.i.d. sample from the distribution of (X, Y) .

The joint distribution can be described in a variety of ways in this case. In classification probably the most intuitive way to define $P_{X,Y}$ is to do so by a pair of (P_X, η) , where P_X is the distribution of X - the marginal distribution of $P_{X,Y}$ - and η is the conditional probability function, i.e.

$$\begin{aligned} P_X(A) &\doteq \mathbb{P}(X \in A) \quad \text{and} \\ \eta(x) &= \mathbb{P}(Y = 1 | X = x). \end{aligned} \tag{1.1}$$

The following claims are proved based on [8].

Claim 1.1.1. *The pair (P_X, η) determines the joint distribution of (X, Y) .*

Proof. It is sufficient to see that for all measurable $C \subseteq \mathbb{X} \times \{+1, -1\}$ the probability $P_{X,Y}(C) = \mathbb{P}((X, Y) \in C)$ can be derived from P_X and η . Notice that

$$C = (C \cap (\mathbb{X} \times \{+1\})) \cup (C \cap (\mathbb{X} \times \{-1\})) = C_{+1} \times \{+1\} \cup C_{-1} \times \{-1\}, \tag{1.2}$$

where C_{+1} and C_{-1} are determined measurable parts of \mathbb{X} . With these sets we can express the

sought quantity as

$$\begin{aligned}\mathbb{P}((X, Y) \in C) &= \mathbb{P}(X \in C_{+1}, Y = +1) + \mathbb{P}(X \in C_{-1}, Y = -1) \\ &= \int_{C_{+1}} \eta(x) dP_X(x) + \int_{C_{-1}} (1 - \eta(x)) dP_X(x),\end{aligned}\tag{1.3}$$

that is P_X and η determines $P_{X,Y}$. \square

We call a $g : \mathbb{X} \rightarrow \{+1, -1\}$ measurable function *classifier* or *decision rule*. This function chooses a class for each input point. To define the problem of classification a *loss function* is needed, which is an $L : \mathbb{Y} \times \mathbb{Y} \rightarrow \mathbb{R}_+$ function measuring our error in a given output point. The most common loss function in classification is the 0/1 loss which is defined as $L(\hat{y}, y) \doteq \mathbb{I}(\hat{y} \neq y)$, where \mathbb{I} denotes the indicator function. Usually the goal of classification is to minimize the so-called *a priori risk* or *Bayes risk*, which is the expected loss, i.e. $R(g) \doteq \mathbb{E}[L(g(X), Y)]$. We say that g is *Bayes optimal* or just *optimal* if its risk is minimal, i.e. $g \in \arg \min R(f)$. From the definition we can see that the Bayes optimal classifier is not necessarily unique and sometimes it does not exist. Nevertheless the loss functions that are applied in practice usually ensure at least the existence of an optimal classifier. We will see that uniqueness does not occur in general. When the 0/1 loss is used we can see that the Bayes risk becomes simply the misclassification probability:

$$R(g) = \mathbb{E}[\mathbb{I}(g(X) \neq Y)] = \mathbb{P}(g(X) \neq Y).\tag{1.4}$$

Moreover, the Bayes optimal classifiers can be easily interpreted and expressed with the help of the η function.

For practical reasons we define the sign function as follows

$$\text{sign}(x) \doteq 2\mathbb{I}(x > 0) - 1.\tag{1.5}$$

Claim 1.1.2. *The following function*

$$g_*(x) = \text{sign}(2\eta(x) - 1) = \text{sign}\left(\mathbb{E}[Y | X = x]\right),\tag{1.6}$$

with domain \mathbb{X} , is Bayes optimal in case of the 0/1 loss.

This claim is very intuitive. It says that when we look at variable Y conditioned on $X = x$ we simply observe a Rademacher variable which takes the value $+1$ with probability $p(x)$ and the value -1 with probability $1 - p(x)$. It is reasonable that the optimal classifier always chooses the outcome which is more likely to occur, this way minimizing the misclassification error. When the two class probabilities are equal to each other then we can choose arbitrarily. In such cases our choice, g_* , prefers the $+1$ class, because of (1.5). The formal proof from [8, Theorem 2.1] can be found in the appendix, see B.1.

Notice that the conditional expected value function, $f_*(x) \doteq \mathbb{E}[Y | X = x]$, which is also called the *regression function*, contains even more information than the Bayes optimal classifier, because g_* can be easily derived from f_* . Besides, the misclassification probability for each input point is encoded in the regression function, hence it is worth examining f_* instead of g_* .

1.2 The Regression Function

We started by defining the problem of binary classification and we found that a Bayes optimal classifier can be expressed with the so-called regression function, $\mathbb{E}[Y|X = x]$. Nevertheless, as the regression word indicates it, this function is extremely important in the problem of regression which is as follows. The setup for the regression problem is similar to what was before. Again, we have a sample from an (X, Y) random pair's unknown distribution, though this time Y is real-valued, therefore we use a different loss function than in classification. There are several options here. The most common choices are the squared deviation, $L(\hat{y}, y) \doteq (\hat{y} - y)^2$, or the absolute deviation, $L(\hat{y}, y) \doteq |\hat{y} - y|$. Given a loss we can define a risk function similarly as before for any real-valued measurable $f : \mathbb{X} \rightarrow \mathbb{Y}$.

$$R(f) \doteq \mathbb{E}[L(f(X), Y)] \quad (1.7)$$

In regression our goal is to minimize this risk functional. From now on we are going to use the squared deviation to penalize the error we make, because it has many advantages. Notice that our problem becomes an L_2 risk minimization problem with this loss as

$$R(f) = \mathbb{E}[(f(X) - Y)^2]. \quad (1.8)$$

We would like to find a (measurable) function $f_* : \mathbb{X} \rightarrow \mathbb{Y}$ such that

$$\mathbb{E}[(f_*(X) - Y)^2] = \min_f \mathbb{E}[(f(X) - Y)^2] \quad (1.9)$$

Here, we assume that this integration can be carried out, therefore $\mathbb{E}Y^2 < \infty$ is required. Recall that the regression function, i.e. the conditional expected value function $f_*(x) \doteq \mathbb{E}(Y|X = x)$ has this minimizing property, indeed the conditional expectation can be viewed as an orthogonal projection in the proper L_2 space. Thus, for any (measurable) f in $L_2(P_X)$ we have that

$$\begin{aligned} \mathbb{E}[(f(X) - Y)^2] &= \mathbb{E}[(f(X) - f_*(X) + f_*(X) - Y)^2] \\ &= \mathbb{E}[(f(X) - f_*(X))^2] + \mathbb{E}[(f_*(X) - Y)^2] - 2\mathbb{E}[(f(X) - f_*(X))(f_*(X) - Y)] \\ &= \mathbb{E}[(f(X) - f_*(X))^2] + \mathbb{E}[(f_*(X) - Y)^2] - 2\mathbb{E}[\mathbb{E}((f(X) - f_*(X))(f_*(X) - Y)|X)] \\ &= \mathbb{E}[(f(X) - f_*(X))^2] + \mathbb{E}[(f_*(X) - Y)^2] - 2\mathbb{E}[(f(X) - f_*(X))(f_*(X) - f_*(X))] \\ &= \mathbb{E}[(f(X) - f_*(X))^2] + \mathbb{E}[(f_*(X) - Y)^2]. \end{aligned} \quad (1.10)$$

Therefore the risk of f can be divided into two parts, where $\mathbb{E}[(f_*(X) - Y)^2]$ is called the *Bayes risk* and it does not depend on f and $\mathbb{E}[(f(X) - f_*(X))^2]$ is called the L_2 *error* of f and it is minimized when $f = f_*$.

In theory f_* has the lowest risk, but since we do not know the distribution of (X, Y) we cannot predict Y using $f_*(X)$. Recall that our only access to the true distribution is the given i.i.d. sample, $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$. Therefore we estimate the regression function with the help of the data. In other words we try to find a data-dependent $f_n^{\mathcal{D}}$ which is close to the regression

function, f_* . This statistical estimation process is also called learning. As it is clear that an estimation $f_n^{\mathcal{D}}$ depends on the sample we leave out \mathcal{D} from the notation and think of f_n as a random function measurable to \mathcal{D} .

Similarly to (1.10), since f_n is measurable to \mathcal{D} and (X, Y) is independent of \mathcal{D}

$$\begin{aligned} \mathbb{E} \left((f_n(X) - Y)^2 \mid \mathcal{D} \right) &= \mathbb{E} \left((f_n(X) - f_*(X) + f_*(X) - Y)^2 \mid \mathcal{D} \right) \\ &= \mathbb{E} \left((f_n(X) - f_*(X))^2 \mid \mathcal{D} \right) + \mathbb{E} \left[(f_*(X) - Y)^2 \right] - 2\mathbb{E} \left[(f_n(X) - f_*(X))(f_*(X) - Y) \mid \mathcal{D} \right] \\ &= \int_{\mathbb{X}} (f_n(x) - f_*(x))^2 dP_X(x) + \mathbb{E} \left[(f_*(X) - Y)^2 \right] \end{aligned} \quad (1.11)$$

holds, so the L_2 risk of an estimate f_n is close to the optimal if and only if the random L_2 error is small. That is why it is beneficial to use the $L_2(P_X)$ metric to measure the fit of an estimate.

Notice that $\int_{\mathbb{X}} (f_n(x) - f_*(x))^2 dP_X(x)$ is a random quantity, because f_n depends on the sample. In addition, we can see that the L_2 error has only theoretic advantages and in practice usually it is not possible to calculate it for a given f_n estimate, because we do not have access to the probability measure P_X nor we know the true regression function f_* . Because of these reasons in many cases it is useful to measure the distance between the estimator and the regression function in more traditional ways such as the pointwise distance for a fixed $x \in \mathbb{X}$, the sup-norm or the L_p norms with $p \geq 1$ with the Lebesgue measure when $\mathbb{X} \subseteq \mathbb{R}^n$ is compact.

1.3 Classification vs Regression

We introduced the two main problems of supervised learning. We defined the risk functional for each case as the expected loss and our goal became to minimize these risks. In classification we used the 0/1 loss while in regression we applied the squared error. For classification it was desirable to find the so-called Bayes optimal classifier, which can be described as the sign of the regression function. In regression we showed that the regression function has always minimal risk when the squared error is used. Consequently if we can solve the regression problem and find the regression function then we can derive the Bayes optimal classifier as well, which suggests that classification is easier than regression. In fact this is the case in some sense. In this section we are going to show that if we have a method to estimate the regression function well than we can also derive a decision rule with close to optimal risk. For any measurable function f we define $g(x) \doteq \text{sign}(f(x))$, which is the plug-in classifier of f . Then similarly to (B.4) we have

$$\begin{aligned} \mathbb{P}(g(X) \neq Y) - \mathbb{P}(g_*(X) \neq Y) &= \int_{\mathbb{X}} f_*(x) \left(\mathbb{I}(g_*(x) = 1) - \mathbb{I}(g(x) = 1) \right) dP_X(x) \\ &\leq \int_{\mathbb{X}} |f_*(x) - f(x)| dP_X(x) \leq \sqrt{\int_{\mathbb{X}} |f_*(x) - f(x)|^2 dP_X(x)}, \end{aligned} \quad (1.12)$$

where we used that $f_*(x) \left(\mathbb{I}(g_*(x) = 1) - \mathbb{I}(g(x) = 1) \right) \leq |f_*(x) - f(x)|$ for all $x \in \mathbb{X}$. It holds because when $\left(\mathbb{I}(g_*(x) = 1) - \mathbb{I}(g(x) = 1) \right) \neq 0$ then $f_*(x)$ and $f(x)$ have different signs, therefore $f_*(x) \leq |f_*(x)| \leq |f_*(x) - f(x)|$. The second inequality holds because of the

Cauchy-Schwartz inequality. Furthermore, with the same steps we can conclude that

$$\mathbb{P}(g_n(X) \neq Y | \mathcal{D}) - \mathbb{P}(g_*(X) \neq Y) \leq \sqrt{\int_{\mathbb{X}} |f_*(x) - f_n(x)|^2 dP_X(x)}. \quad (1.13)$$

As we can see an estimate f_n with small L_2 error leads to a decision rule with close to optimal misclassification probability. It is clear though that to construct an optimal classifier it is sufficient to find a function which has always the same sign as the regression function. To sum up, regression is a harder problem than classification, but it gives us a better understanding of the underlying data generating mechanism.

Chapter 2

Formal Learning Models

We have already got to know the two most important supervised learning problems, classification and regression. We defined our goals as risk minimization in both cases. In classification we wish to find the Bayes classifier in regression we search for the regression function. Since both of these are just a minimizer of a risk we call them target functions from now on. Our problem is that we do not have direct access to the risk, but only to an i.i.d. sample. That is where learning is introduced. In this section we are going to define the notion of learning precisely, but first we deal with the challenge of risk minimization.

2.1 Empirical Risk Minimization

For now, our biggest concern is that we do not have direct access to the risk which is the expected value of a loss. The first idea is to estimate the true risk with an empirical average, which can be calculated from the data

$$\hat{R}(g) \doteq \frac{1}{n} \sum_{i=1}^n L(g(X_i), Y_i). \quad (2.1)$$

This is called the *empirical risk*. The strong law of large numbers (SLLN) ensures that this quantity for any given g tends to the risk (*a.s.*) as $n \rightarrow \infty$, so this should be a good estimate for big datasets.

Our hope is that $\hat{R}(g)$ is close to $R(g)$, therefore it is a reasonable idea to minimize the empirical version instead of the true risk. Thus, we need to find a g which satisfies

$$g \in \arg \min \frac{1}{n} \sum_{i=1}^n L(g(X_i), Y_i). \quad (2.2)$$

This type of optimization is called the *empirical risk minimization* (ERM) principle. Though it is a natural approach, some issues arise when we try to solve this optimization for example for the regression problem. Since we have not restricted ourselves to a specific function class yet, i.e. f can be any measurable function, we can construct $\hat{f}(x) = \sum_{i=1}^n \mathbb{I}(x = X_i) Y_i$. It is easy to see that when all X_i are different $\hat{R}(\hat{f}) = 0$, so we found a minimum, since the empirical risk is

nonnegative. Unfortunately this does not imply that our true risk is small. Instead in this case the opposite can be true, that is for instance when P_X is absolutely continuous for the Lebesgue measure then $\hat{f} = 0$ in the $L_2(P_X)$ sense. Thus, the risk equals to the risk of the zero function which was arbitrarily chosen. Such estimates can only perform well for already seen inputs, whenever a new input is observed \hat{f} fails to predict its regression, in other words \hat{f} cannot generalize well. We can say that this estimator only memorizes the data. This phenomenon happens if we apply the empirical risk minimization principle without restricting ourselves to a function class where the minimization is considered. Our problem is that though for each function f we know that $\hat{R}(f) \xrightarrow{a.s.} R(f)$ as $n \rightarrow \infty$, it does not imply that for all function classes \mathcal{F} we have

$$\inf_{f \in \mathcal{F}} \hat{R}(f) \xrightarrow{a.s.} \inf_{f \in \mathcal{F}} R(f). \quad (2.3)$$

In fact, without restrictions on \mathcal{F} often $\inf_{f \in \mathcal{F}} \hat{R}(f) = 0$ occurs for all $n \in \mathbb{N}$. When our estimate achieves a low empirical risk without generalizing well, we say that our estimate *overfits* the data. It is something that we really want to avoid.

2.2 Bias-Variance Trade-Off

We argued that if we continue our thread of reasoning with ERM, then we must make a restriction on the function class we minimize on. As a start let \mathcal{F} denote the set where the minimization takes place. Then the ERM estimate of the target function is defined as

$$\hat{f} \in \arg \min_{f \in \mathcal{F}} \hat{R}(f). \quad (2.4)$$

We assume that this minimum exists and though we do not require unicity we think of \hat{f} as a well-defined single function. This holds for the whole thesis, whenever we write $\arg \min$ we assume that it exists and it is well-defined. Notice that when $f_* \notin \mathcal{F}$ then the ERM method cannot find the true target function. We can only hope to infer the best function in the model class, which is $\tilde{f} \in \arg \min_{f \in \mathcal{F}} R(f)$. Notice that the L_2 error can be described as the sum of two terms

$$R(\hat{f}) - R(f) = R(\hat{f}) - R(\tilde{f}) + R(\tilde{f}) - R(f), \quad (2.5)$$

where $R(\hat{f}) - R(\tilde{f})$ is called the estimation error and $R(\tilde{f}) - R(f)$ is called the approximation error. We can see that there is a trade-off here. When we choose a big \mathcal{F} which contains g_* more likely, the approximation error will become small. On the other hand in this case the chance of overfitting becomes high implying that the estimation error increases. In the other direction we may choose a small \mathcal{F} in which case it is easy to estimate \tilde{f} well, but this time the approximation error can grow large, because f_* can have much lower risk than \tilde{f} . In conclusion, it is our goal to find the right balance between these two. Doing so we should define more precisely what it means to have a big model class and find a method to control it.

2.3 Parametric vs Nonparametric Statistics

There are several ways to find the right model class we work with. The classical approach is to simply fix one in advance, which is parameterized with finitely many parameters. This way we can encode our prior knowledge about the target function's structure. For example we can assume that the regression function is in a finite dimensional linear space (linear regression) or we can assume that the Bayes classifier is linear (e.g. perceptron). This approach has many advantages and also some short-comings. When we have a priori knowledge about the target function then it sounds like a good idea to use it. In practice though, our a priori knowledge is often limited, therefore we want to make as few assumptions on the data as possible. In this perspective it is our goal to reduce the constraints that we build in our methods. Thus, when we know the structure or some property of the target function it is reasonable to exclude those models which do not coincide with our criteria, but when such solid knowledge is not available we should make as few assumptions as possible. Parametric estimators have the advantage that they can perform well even for a small data set, because of those a priori concepts that are built inside them. Besides, they are often easy to interpret. Many times the parameters have actual physical meanings. On the other hand, they are not really flexible, since usually for different problems we make different assumptions on the target function. Furthermore, it is important to keep in mind that via the parameterization we introduce an inductive biasedness to the problem, which can be misleading when we deal with real-world data. For example assuming linear dependence is reasonable many times, but in reality it is usually just an abstract simplification of the complicated data generating mechanism. In conclusion parametric methods can perform well, but it is important to keep in mind their limitations.

Nonparametric methods were developed to overcome the issues we mentioned concerning the parametric methods. These are very flexible tools that do not assume that the target function can be described by finitely many parameters. Usually the number of parameters which arise in these algorithms increases with the data size. These methods can be applied almost without any prior knowledge. The most known examples of these are the local averaging estimates, see [11] and the kernel-methods, see [18], which are going to play a central role in this thesis. In Chapter 3 and 4 we are going to analyze these techniques more deeply. For now it is sufficient to see that during these methods we do not apply the ERM principle. Instead we either construct an estimator on a smart way and hope that it will have good generalizing property and close to optimal risk or we adopt the capacity of the model class to the data size. This second paradigm is related to the bias-variance trade-off which was mentioned earlier. To make this precise, later we are going to define the capacity or complexity measure of model classes and present the *structural risk minimization* principle. To sum up nonparametric methods are very flexible, need almost no prior knowledge about the data and are computationally cheap or at least manageable, while the ERM method in practice often has a high computational burden. On the other hand their convergence to an optimal estimator is often slow, therefore they need more data in general.

2.4 Learning Theory

In this section we finally define the notion of learning. We are already familiar with all important concepts that we will use. Recall that given an i.i.d. sample, $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$, from an unknown distribution. Regarding the distribution we do not have any preconception. However we are going to assume that we have a model class \mathcal{F} . This model class or hypothesis set incorporates our prior knowledge about the target function. For now it is fixed. We are going to focus on the problem of risk minimization in this model class. Again, we assume that there exists a minimum. Let Γ denote a learning algorithm, which choses a model from \mathcal{F} based on any given sample, formally $\Gamma(\mathcal{D}) \in \mathcal{F}$. It will be clear that usually we cannot hope for an algorithm which always finds the minimum exactly. It is the case for example when we deal with absolutely continuous distributions and model classes containing infinitely many functions with high expressing capacity. Hence we are going to be satisfied even when we can reduce our risk very close to the minimum with a chosen high probability. This leads us to the framework of *probably approximately correct* learning or simply PAC learning (also called *agnostic PAC learning*), see [19]. Let R_P denote the risk with the indication that it depends on the joint distribution P .

Definition 2.4.1 (PAC learnability). *We say that a model class \mathcal{F} is PAC learnable if for all $\varepsilon > 0$ and $\delta > 0$ there exists $n_{\mathcal{F}}(\varepsilon, \delta) \in \mathbb{N}$ and a learning algorithm Γ such that when Γ learns from an i.i.d. sample with size $n \geq n_{\mathcal{F}}(\varepsilon, \delta)$ from any distribution P , then it returns a model $\Gamma(\mathcal{D}) = \hat{g}$ such that*

$$\mathbb{P}(R_P(\hat{g}) \leq \min_{g \in \mathcal{F}} R_P(g) + \varepsilon) \geq 1 - \delta. \quad (2.6)$$

Here, we have two approximation parameters, we refer to ε as accuracy parameter and to δ as the confidence parameter. The function $n_{\mathcal{F}}$ tells us how many sample points we need to infer a PAC estimate. It is easy to see that when \mathcal{F} is PAC learnable then there are many such functions. We call the minimal one the *sample complexity*. It is important that in practice when we know the sample complexity, then we can give non-asymptotic guarantees for the risk of our estimate, which is desired in practice.

This concept is very intuitive and formalizes our desires about a learning algorithm, but we will see that it is a little bit restrictive too. We should ask what model classes are PAC learnable. Now we are going to deal with only the problem of binary classification since it is our main topic. First, we are going to see that in classification the set of all measurable decision rules is not PAC learnable, hence there is no universal algorithm to solve the classification problem in the PAC sense. This statement is called the *no-free-lunch theorem* for classification. There are some stricter versions of this theorem, but trying to remain self-contained we present a basic one from [19].

Theorem 2.4.1. (*no-free-lunch*) *Consider the problem of binary classification. For a domain $\mathbb{X} \subseteq \mathbb{R}^d$ for all sample sizes, $n \in \mathbb{N}$, smaller than $|\mathbb{X}|/2$ ($|\mathbb{X}|$ can be infinite), and learning*

algorithm Γ there exists a distribution P such that there is a measurable g with $R_P(g) = 0$ and

$$\mathbb{P}\left(R_P(\Gamma(D)) \geq \frac{1}{8}\right) \geq \frac{1}{7}. \quad (2.7)$$

It says that for a learning algorithm and sample size we can find a distribution P such that Γ performs poorly on it with a significant high probability.

Proof. Fix $n \in \mathbb{N}$ and learning algorithm Γ . Let $C \subseteq \mathbb{X}$ such that $|C| = 2n$. Our idea is that if we can only observe half of the possible instances then we cannot find a classifier for the unseen part of the domain. It is clear that there are $T = 2^{2n}$ distinct classifiers on C . Denote these with $g_1, \dots, g_T : C \rightarrow \{+1, -1\}$. For all $i \in [T]$ let P_i be a distribution defined as $P_i((x, g_i(x))) = 1/2n$ for all $x \in \mathbb{X}$ and zero everywhere else. We use the $[T] \doteq \{1, \dots, T\}$ notation. We can see that we always have a uniform marginal distribution on C . Furthermore, clearly $R_{P_i}(g_i) = 0$ for $i \in [T]$. We are going to show that

$$\max_{i \in [T]} \mathbb{E}\left(R_{P_i}(\Gamma(D))\right) \geq \frac{1}{4}. \quad (2.8)$$

From this equation our theorem follows because there exists a P_i for which $\mathbb{E}(R_{P_i}(\Gamma(D))) \geq 1/4$ and we know that $R_{P_i}(\Gamma(D)) \in [0, 1]$ so the following holds

$$1 \mathbb{P}\left(R_{P_i}(\Gamma(D)) \geq \frac{1}{8}\right) + \frac{1}{8} \mathbb{P}\left(R_{P_i}(\Gamma(D)) < \frac{1}{8}\right) \geq \mathbb{E}\left(R_{P_i}(\Gamma(D))\right) \geq \frac{1}{4}. \quad (2.9)$$

Rearranging the equation yields that

$$\mathbb{P}\left(R_{P_i}(\Gamma(D)) \geq \frac{1}{8}\right) \geq \frac{1}{7}. \quad (2.10)$$

We turn to prove the inequality in (2.8). Let $K = (2n)^n$, which is the number of the possible sample sequences we can get from C . Let $\mathcal{S}_j = (x_1, \dots, x_n)$ denote these sequences for $j \in [K]$ and let $\mathcal{D}_j^i = \{(x_1, g_i(x_1)), \dots, (x_n, g_i(x_n))\}$ be the labeled version of \mathcal{S}_j with classifier g_i . We know that all sequences are equally likely so for a fixed distribution P_i we have

$$\mathbb{E}_{\mathcal{D} \sim \mathcal{D}_i^n} [R_{P_i}(\Gamma(\mathcal{D}))] = \frac{1}{K} \sum_{j=1}^K R_{P_i}(\Gamma(\mathcal{D}_j^i)). \quad (2.11)$$

The maximum is always greater than the average, which is greater than the minimum, therefore we obtain that

$$\max_{i \in [T]} \frac{1}{K} \sum_{j=1}^K R_{P_i}(\Gamma(\mathcal{D}_j^i)) \geq \frac{1}{T} \sum_{i=1}^T \frac{1}{K} \sum_{j=1}^K R_{P_i}(\Gamma(\mathcal{D}_j^i)) \quad (2.12)$$

$$= \frac{1}{K} \sum_{j=1}^K \frac{1}{T} \sum_{i=1}^T R_{P_i}(\Gamma(\mathcal{D}_j^i)) \geq \min_{j \in [K]} \frac{1}{T} \sum_{i=1}^T R_{P_i}(\Gamma(\mathcal{D}_j^i)) \quad (2.13)$$

holds. Fix a $j \in [K]$. Let v_1, \dots, v_p be those elements of C which do not appear in \mathcal{S}_j . It is clear

that $p \geq n$. For all $g : C \rightarrow \{+1, -1\}$ and for all $i \in [T]$ we know that

$$\begin{aligned} R_{P_i}(g) &= \frac{1}{2n} \sum_{x \in C} \mathbb{I}(g(x) \neq g_i(x)) \\ &\geq \frac{1}{2n} \sum_{r=1}^p \mathbb{I}(g(v_r) \neq g_i(v_r)) \geq \frac{1}{2p} \sum_{r=1}^p \mathbb{I}(g(v_r) \neq g_i(v_r)) \end{aligned} \quad (2.14)$$

holds, from which it follows that

$$\begin{aligned} \frac{1}{T} \sum_{i=1}^T R_{P_i}(\Gamma(\mathcal{D}_j^i)) &\geq \frac{1}{T} \sum_{i=1}^T \frac{1}{2p} \sum_{r=1}^p \mathbb{I}(\Gamma(\mathcal{D}_j^i)(v_r) \neq g_i(v_r)) \\ &\geq \frac{1}{2p} \sum_{r=1}^p \frac{1}{T} \sum_{i=1}^T \mathbb{I}(\Gamma(\mathcal{D}_j^i)(v_r) \neq g_i(v_r)) \geq \frac{1}{2} \min_{r \in [p]} \frac{1}{T} \sum_{i=1}^T \mathbb{I}(\Gamma(\mathcal{D}_j^i)(v_r) \neq g_i(v_r)). \end{aligned} \quad (2.15)$$

Now fix an $r \in [p]$. Let's partite the set $\{g_1, \dots, g_T\}$ into $T/2$ disjoint pairs so that g_k and g_l belong together if $g_k(v) \neq g_l(v) \Leftrightarrow v = v_r$. Since v_r is not in sequence \mathcal{D}_j^i we have that for every pair (g_k, g_l)

$$\mathbb{I}(\Gamma(\mathcal{D}_j^i)(v_r) \neq g_k(v_r)) + \mathbb{I}(\Gamma(\mathcal{D}_j^i)(v_r) \neq g_l(v_r)) = 1. \quad (2.16)$$

Summing up these terms for all pairs and dividing by T we conclude that

$$\frac{1}{T} \sum_{i=1}^T \mathbb{I}(\Gamma(\mathcal{D}_j^i)(v_r) \neq g_i(v_r)) = \frac{1}{2}. \quad (2.17)$$

Substituting this to (2.15) the lower bound for the sought expected value is proved. \square

Corollary 2.4.1.1. *When \mathbb{X} is an infinite domain set, then the set of all measurable classifiers $\overline{\mathcal{F}}$ is not PAC learnable.*

Proof. Assume by contradiction that $\overline{\mathcal{F}}$ is PAC learnable. Then for $\varepsilon < 1/8$ and $\delta < 1/7$ there exists a learning algorithm Γ and sample size $n_{\overline{\mathcal{F}}}(\varepsilon, \delta)$ such that for all distributions P when $n \geq n_{\overline{\mathcal{F}}}(\varepsilon, \delta)$ we have

$$\mathbb{P}\left(R_P(\Gamma(\mathcal{D})) \leq \min_{g \in \overline{\mathcal{F}}} R_P(g) + \varepsilon\right) \geq 1 - \delta. \quad (2.18)$$

From the no-free-lunch theorem we also know that there is a distribution \tilde{P} such that there is a g with zero risk and

$$\mathbb{P}\left(R_{\tilde{P}}(\Gamma(\mathcal{D})) \geq \frac{1}{8}\right) = 1 - \mathbb{P}\left(R_P(\Gamma(\mathcal{D})) < 0 + \frac{1}{8}\right) \geq \frac{1}{7}. \quad (2.19)$$

It leads to a contradiction since $\varepsilon < 1/8$, $\delta < 1/7$ and

$$\mathbb{P}\left(R_{\tilde{P}}(\Gamma(\mathcal{D})) < \frac{1}{8}\right) \leq 1 - \frac{1}{7} < 1 - \delta \leq \mathbb{P}\left(R_{\tilde{P}}(\Gamma(\mathcal{D})) \leq \varepsilon\right), \quad (2.20)$$

thus the corollary follows. \square

2.5 Uniform Convergence

We present necessary and sufficient conditions for PAC learnability. In this section we concentrate on deriving a sufficient condition. Our idea to ensure PAC learnability is to apply an efficient learning algorithm Γ and determine the largest model class, \mathcal{F} , which Γ can learn. We are going to consider the ERM method. Regarding this learning algorithm our problem was that though $|\hat{R}(g) - R(g)| \rightarrow 0$ holds for all measurable g , it does not imply that $\sup_{g \in \mathcal{F}} |\hat{R}(g) - R(g)| \rightarrow 0$ for a model class \mathcal{F} , where the convergences hold almost surely. Nevertheless, recall that an asymptotical condition in general is not sufficient for PAC learning, because we want non-asymptotic guarantees. Therefore we need a stricter criterion.

Definition 2.5.1. (*ε -representative sample*) A sample \mathcal{D} is ε -representative with respect to domain \mathbb{X} , model class \mathcal{F} , loss L and distribution P if for all $g \in \mathcal{F}$

$$|\hat{R}(g) - R(g)| \leq \varepsilon. \quad (2.21)$$

Claim 2.5.1. When \mathcal{D} is $\varepsilon/2$ -representative then any output of the ERM method with respect to \mathcal{F} , that is $\hat{g} \in \arg \min_{g \in \mathcal{F}} \hat{R}(g)$, satisfies

$$R(\hat{g}) \leq \min_{g \in \mathcal{F}} R(g) + \varepsilon. \quad (2.22)$$

Proof. Let $g_* \in \arg \min_{\mathcal{F}} R(g)$. The following holds

$$\begin{aligned} R(\hat{g}) - R(g_*) &= R(\hat{g}) - \hat{R}(\hat{g}) + \hat{R}(\hat{g}) - R(g_*) \\ &\leq |R(\hat{g}) - \hat{R}(\hat{g})| + |\hat{R}(\hat{g}) - R(g_*)| \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2}, \end{aligned} \quad (2.23)$$

because both absolute deviations are lower than $\varepsilon/2$, since \mathcal{D} is $\varepsilon/2$ -representative. We used the fact that when two functions are close to each other uniformly then their minimums are also close to each other. \square

We can see that if the sample is ε -representative with probability at least $1 - \delta$, then the ERM method is a PAC learner of \mathcal{F} .

Definition 2.5.2. (*uniform convergence*) We say that a model class \mathcal{F} has the uniform convergence property w.r.t. a domain \mathbb{X} and loss L if for all $\varepsilon > 0$ and $\delta > 0$ there exists $n_{\mathcal{F}}^U(\varepsilon, \delta) \in \mathbb{N}$ such that if \mathcal{D} is a sample with at least $n \geq n_{\mathcal{F}}^U(\varepsilon, \delta)$ i.i.d. examples according to any distribution P , then \mathcal{D} is ε -representative with probability at least $1 - \delta$.

The uniform adjective refers to the fact that a PAC statement is true for any element of \mathcal{F} and for any distribution P .

An important corollary of Claim [2.5.1](#) is that whenever \mathcal{F} has the uniform convergence property then it is also PAC learnable, namely the ERM method is a PAC learner of \mathcal{F} .

2.6 The VC Dimension

In this section we still consider the problem of classification. Trying to find a necessary and sufficient condition for PAC learnability we revisit the bias-variance trade-off phenomenon, see Section 2.2. There, we showed that the “size” of the model class has a huge impact on learning, that is if we include more models in \mathcal{F} then our approximation error will decrease, but the estimation error can grow high. It can be showed that finite model classes are PAC learnable and also have the uniform convergence property, see [19]. Nevertheless, it is not hard to argue that many infinite model classes are PAC learnable as well, hence it is clear that the cardinality of \mathcal{F} is not the best measure to decide whether a model class is PAC learnable or not. Indeed, we need a new concept here, which measures the complexity or capacity of a model class. The goal of these measures is to characterize somehow what model classes are learnable. There are several notions for complexity. We are going to present the most important ones. In this section we present the celebrated theory of VC dimension or Vapnik-Chervonenkis dimension, which was proposed by Vladimir Vapnik and Alexey Chervonenkis, see [23]. Besides, other important complexity measures are going to be defined in later sections for deriving sufficient conditions for the uniform law of large numbers. In this section we are going to prove that VC dimension grants a necessary and sufficient condition for PAC learnability.

First, we take another look at the no-free-lunch theory that we proved. Recall that we showed that for every learning algorithm and sample size we can construct a distribution on which the learning algorithm will perform poorly. Our idea was that when there are too many decision rules which perform well on the data then a learning algorithm will have a very small chance to find the optimal one. It is really the problem with the case when we do not make any restriction on the model class. We can explain every dataset arbitrary well with many models. This leads us to the concept of shattering. Assume that a model class of classifiers, \mathcal{F} , is given. Let $C = \{c_1, \dots, c_n\} \subseteq \mathbb{X}$ be a finite subset of the input space. Restrict our classifiers in \mathcal{F} to set C , so that $\mathcal{F}_C = \{(g(c_1), \dots, g(c_n)) \mid g \in \mathcal{F}\}$. We see that $\mathcal{F}_C \subseteq \{+1, -1\}^n$. We say that \mathcal{F} *shatters* C if \mathcal{F}_C contains all possible outcomes on C , that is when $|\mathcal{F}_C| = 2^n$.

Definition 2.6.1. (*VC dimension*) Let \mathcal{F} be a set of classifiers. The VC dimension of \mathcal{F} is the maximal size of a set $C \subseteq \mathbb{X}$, $|C| = n$ which is shattered by \mathcal{F} . The VC dimension is infinite when the maximum does not exist.

Claim 2.6.1. If \mathcal{F} has infinite VC dimension then it is not PAC learnable.

Proof. Assume by contradiction that for $0 < \varepsilon < 1/8$ and $0 < \delta < 1/7$ there exists a learning algorithm Γ which PAC learns \mathcal{F} from any i.i.d. sample of size $n_{\mathcal{F}}(\varepsilon, \delta)$. Then consider a set $C \subseteq \mathbb{X}$ of size $2n_{\mathcal{F}}(\varepsilon, \delta)$ which is shattered by \mathcal{F} . For this set C we can apply the no-free-lunch theorem, see Theorem 2.4.1, from which it follows that there is a distribution \tilde{P} that Γ fails to PAC learn. We reached the contradiction. \square

We defined the VC dimension as the maximal size of a set which can be shattered. The maximal size of $|\mathcal{F}_C|$ taking the maximum in C with a fixed size is another useful quantity called growth, shatter coefficient or maximal effective size.

Definition 2.6.2. (*growth function*) Let \mathcal{F} be a model class. The growth function of \mathcal{F} is defined as

$$\tau_{\mathcal{F}}(n) \doteq \max_{|C|=n, C \subseteq \mathbb{X}} |\mathcal{F}_C|. \quad (2.24)$$

When \mathcal{F} has infinite VC dimension then it is clear that $\tau_{\mathcal{F}}(n) = 2^n$ for all $n \in \mathbb{N}$. The suprising fact is that when \mathcal{F} has finite VC dimension then the growth function can only increase polynomially, that is the following holds, see [11], [19] and [23].

Theorem 2.6.1. (*Sauer*) Let \mathcal{F} be a model class with VC dimension $V_{\mathcal{F}}$. Then for any $n \in \mathbb{N}$

$$\tau_{\mathcal{F}}(n) \leq \sum_{i=0}^{V_{\mathcal{F}}} \binom{n}{i}. \quad (2.25)$$

The notation of the proof becomes easier if we think of \mathcal{F} as a set system on $\mathbb{X} \times \{+1, -1\}$. Let $A(g) = \{(x, y) \mid g(x) = y\} \subseteq \mathbb{X} \times \{+1, -1\}$ for all $g \in \mathcal{F}$. Then for a model class \mathcal{F} we can construct a set system $\mathcal{A} = \{A(g) \mid g \in \mathcal{F}\}$. The VC dimension and the growth function can be defined similarly as before, i.e. for all $C \subseteq \mathbb{X} \times \{+1, -1\}$ we can define $\mathcal{A}_C \doteq \{A \cap C \mid A \in \mathcal{A}\}$ and we say that \mathcal{A} shatters C , if $|\mathcal{A}_C| = 2^n$, where $|C| = n$. The VC dimension of \mathcal{A} is the maximal size of those sets that \mathcal{A} can shatter and the growth function becomes

$$\tau_{\mathcal{A}}(n) = \max_{|C|=n, C \subseteq \mathbb{X} \times \{+1, -1\}} |\mathcal{A}_C|. \quad (2.26)$$

Then it is easy to show that $V_{\mathcal{A}} = V_{\mathcal{F}}$, and $\tau_{\mathcal{A}}(n) = \tau_{\mathcal{F}}(n)$. Later we are going to generalize this concept to arbitrary set systems.

Proof. Let \mathcal{A} be as before and $C = \{(x_1, y_1), \dots, (x_n, y_1)\} = \{z_1, \dots, z_n\} \subseteq \mathbb{X} \times \{+1, -1\}$. We are going to show that

$$|\{A \cap C \mid A \in \mathcal{A}\}| \leq \sum_{i=0}^{V_{\mathcal{A}}} \binom{n}{i}. \quad (2.27)$$

Let $K = \binom{n}{V_{\mathcal{A}}+1}$ and F_1, \dots, F_K be the subsets of C of size $V_{\mathcal{A}} + 1$. Notice that for all $i \in [K]$ there exists $H_i \subseteq F_i$ such that

$$A \cap F_i \neq H_i \text{ for all } A \in \mathcal{A} \quad (2.28)$$

because \mathcal{A} shatters no F_i . Since $F_i \subseteq C$ it is clear that $A \cap F_i = (A \cap C) \cap F_i$. Substituting this into (2.28) yields that

$$(A \cap C) \cap F_i \neq H_i \text{ for all } A \in \mathcal{A}. \quad (2.29)$$

Now let $\mathcal{C}_0 \doteq \{D \subseteq C \mid D \cap F_i \neq H_i \text{ for each } i \in [K]\}$. We just showed that $\{A \cap C \mid A \in \mathcal{A}\} \subseteq \mathcal{C}_0$. We are going to prove that $|\mathcal{C}_0| \leq \sum_{i=0}^{V_{\mathcal{A}}} \binom{n}{i}$. Notice that if $H_i = F_i$ for all $i \in [K]$ then $D \cap F_i \neq H_i \Leftrightarrow F_i \not\subseteq D$ from which it follows that \mathcal{C}_0 contains all subsets of C of size at most $V_{\mathcal{A}}$, which are $\sum_{i=0}^{V_{\mathcal{A}}} \binom{n}{i}$ elements.

We are going to reduce the general case to this special one. Let $H'_i \doteq (H_i \cup \{z_1\}) \cap F_i$ for

$i \in [K]$. It means that we link z_1 to H_i if it is in F_i . Furthermore, let

$$\mathcal{C}_1 \doteq \{D \subseteq C \mid D \cap F_i \neq H'_i \text{ for all } i \in [K]\}. \quad (2.30)$$

We claim that $|\mathcal{C}_1| \geq |\mathcal{C}_0|$. Since $\mathcal{C}_0 = (\mathcal{C}_0 \cap \mathcal{C}_1) \cup (\mathcal{C}_0 \setminus \mathcal{C}_1)$ and $\mathcal{C}_1 = (\mathcal{C}_0 \cap \mathcal{C}_1) \cup (\mathcal{C}_1 \setminus \mathcal{C}_0)$ are disjoint unions it is sufficient to prove that $|\mathcal{C}_0 \setminus \mathcal{C}_1| \leq |\mathcal{C}_1 \setminus \mathcal{C}_0|$. Let $\Delta : \mathcal{C}_0 \setminus \mathcal{C}_1 \rightarrow \mathcal{C}_1 \setminus \mathcal{C}_0$ be a function such that $\Delta(D) \rightarrow D \setminus \{z_1\}$. We are going to prove that Δ is injective from which the claim follows. For $D \in \mathcal{C}_0 \setminus \mathcal{C}_1$ we have that $D \cap F_i \neq H'_i$ for all $i \in [K]$, because $D \in \mathcal{C}_0$ and there exists an index i_0 such that $D \cap F_{i_0} = H'_{i_0}$, because $D \notin \mathcal{C}_1$. Consequently

$$H'_{i_0} = (H_{i_0} \cup \{z_1\}) \cap F_{i_0} \neq H_{i_0}, \quad (2.31)$$

hence $z_1 \notin H_{i_0}$, but z_1 is contained in F_{i_0}, H'_{i_0} and D . Since $z_1 \in D$ for all $D \in \mathcal{C}_0 \setminus \mathcal{C}_1$ then the extraction is a one-to-one mapping from $\mathcal{C}_0 \setminus \mathcal{C}_1$, i.e. Δ is injective. It remained to show that $D \setminus \{z_1\} \in \mathcal{C}_1 \setminus \mathcal{C}_0$. Since $D \cap F_{i_0} = H'_{i_0}$ and $z_1 \notin H_{i_0}$ it follows that

$$(D \setminus \{z_1\}) \cap F_{i_0} = (D \cap F_{i_0}) \setminus \{z_1\} = H'_{i_0} \setminus \{z_1\} = H_{i_0}, \quad (2.32)$$

thus $D \setminus \{z_1\} \notin \mathcal{C}_0$. We need to prove that $D \setminus \{z_1\} \in \mathcal{C}_1$. When $z_1 \notin F_i$ then we obtain that

$$(D \setminus \{z_1\}) \cap F_i = D \cap F_i \neq H_i = H'_i, \quad (2.33)$$

because of $D \in \mathcal{C}_0$. When $z_1 \in F_i$ then it is included in H'_i too, but certainly not in $D \setminus \{z_1\}$, hence $D \setminus \{z_1\} \cap F_i \neq H'_i$. Either way $(D \setminus \{z_1\}) \cap F_i \neq H'_i$ holds, i.e. $(D \setminus \{z_1\}) \in \mathcal{C}_1$. So far we proved that $|\mathcal{C}_0| \leq |\mathcal{C}_1|$. We can repeat this procedure $n - 1$ times starting from $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_{n-1}$ with elements z_2, \dots, z_n , then we get classes with

$$|\mathcal{C}_0| \leq |\mathcal{C}_1| \leq \dots \leq |\mathcal{C}_n|. \quad (2.34)$$

In the definition of \mathcal{C}_n we have $H_i = F_i$, hence the special case occurs which was argued in the beginning. \square

Corollary 2.6.1.1. *For a model class \mathcal{F} with VC dimension $V_{\mathcal{F}} < \infty$ we have that for all $n \in \mathbb{N}$*

$$\tau_{\mathcal{F}}(n) \leq (n + 1)^{V_{\mathcal{F}}} \quad (2.35)$$

and for all $n > V_{\mathcal{F}}$

$$\tau_{\mathcal{F}}(n) \leq \left(\frac{en}{V_{\mathcal{F}}}\right)^{V_{\mathcal{F}}}. \quad (2.36)$$

Proof. We can apply the Sauer theorem and the binomial theorem to obtain

$$\tau_{\mathcal{F}}(n) \leq \sum_{i=0}^{V_{\mathcal{F}}} \binom{n}{i} \leq \sum_{i=0}^{V_{\mathcal{F}}} \frac{n!}{(n-i)! i!} \leq \sum_{i=0}^{V_{\mathcal{F}}} n^i \binom{V_{\mathcal{F}}}{i} \leq (n + 1)^{V_{\mathcal{F}}}. \quad (2.37)$$

When $V_{\mathcal{F}}/n < 1$, then similarly as before

$$\begin{aligned} \left(\frac{V_{\mathcal{F}}}{n}\right)^{V_{\mathcal{F}}} \tau_{\mathcal{F}}(n) &\leq \left(\frac{V_{\mathcal{F}}}{n}\right)^{V_{\mathcal{F}}} \sum_{i=0}^{V_{\mathcal{F}}} \binom{n}{i} \leq \sum_{i=0}^{V_{\mathcal{F}}} \left(\frac{V_{\mathcal{F}}}{n}\right)^i \binom{n}{i} \\ &\leq \sum_{i=0}^n \left(\frac{V_{\mathcal{F}}}{n}\right)^i \binom{n}{i} = \left(1 + \frac{V_{\mathcal{F}}}{n}\right)^n \leq e^{V_{\mathcal{F}}}. \end{aligned} \quad (2.38)$$

Thus, the corollary follows. \square

2.7 The Fundamental Theorem of PAC Learning

In this section we state the fundamental theorem of PAC learning, which characterizes the PAC learnable model classes.

Theorem 2.7.1. (*Fundamental Theorem of PAC Learning*) *Consider the problem of binary classification with the 0/1 loss function. Let \mathcal{F} be a class of decision rules from domain \mathbb{X} to $\{+1, -1\}$. Then the followings are equivalent:*

1. \mathcal{F} has the uniform convergence property.
2. Any ERM method is a successful PAC learner of \mathcal{F} .
3. \mathcal{F} is PAC learnable.
4. \mathcal{F} has a finite VC dimension.

There is a quantitative version of this theorem where it turns out that the VC dimension not only characterizes PAC learnability but also determines the sample complexity, for further details see [19].

Here, we stated the theorem for classification. A similar result holds for regression with the squared loss function, however it cannot be generalized for arbitrary learning tasks. It is interesting that there are examples for learnable classes which do not possess the uniform convergence property. Furthermore, it can happen that the ERM method fails to learn a given class, but another algorithm can, see [20].

Proof. We discussed $1 \Rightarrow 2$ in Section 2.5. From 2 to 3, it is a triviality. The implication $3 \Rightarrow 4$ holds, because of Claim 2.6.1. In order to prove $4 \Rightarrow 1$ we first state Theorem 2.7.2.

We are going to use the notion of the empirical distribution. Notice that the quantity $\hat{R}(g) = \frac{1}{n} \sum_{i=1}^n L(g(X_i), Y_i)$ can be seen as an expected value with respect to the measure defined by $P_n(A) \doteq \sum_{i=1}^n \mathbb{I}(Z_i \in A)$ where $Z_i = L(g(X_i), Y_i)$. This probability measure is called the empirical distribution. Notice that it is a random measure as it depends on variables $\{Z_i\}_{i=1}^n$. The celebrated Glivenko–Cantelli theorem states that when Z_1, Z_2, \dots are i.i.d. variables then the empirical measure tends to the true distribution of Z_1 , P , uniformly on the sets of half-lines, see [10] and [4]. We need a generalization of this theorem.

Theorem 2.7.2. (Vapnik–Chervonenkis) For every probability distribution P and a class of measurable sets \mathcal{A} , where $A \subseteq \mathbb{X} \times \{+1, -1\}$ for all $A \in \mathcal{A}$, we have

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \varepsilon\right) \leq 8\tau_{\mathcal{A}}(n) e^{-n\varepsilon^2/32}. \quad (2.39)$$

Remark 2.7.3. The measurability of the supremum in Theorem 2.7.2 needs to be verified. In fact each time we consider a supremum in this thesis measurability issues can arise. The book of Giné and Nickl, see [9], handles such problems with the notion of outer probability. Since for many applications it is sufficient to consider the supremum of countable variables we do not deal with these measurability issues in this thesis.

Proof. The proof will be done in four steps as in [8]. First, we use a symmetrization with a hypothetical sample. Let Z_1, \dots, Z_n be the given i.i.d. sample from P and Z'_1, \dots, Z'_n be an alternative i.i.d. sample from the same distribution independent of the original variables, and P'_n denote the empirical distribution of the alternative sample. We claim that

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \varepsilon\right) \leq 2\mathbb{P}\left(\sup_{A \in \mathcal{A}} |P_n(A) - P'_n(A)| > \varepsilon/2\right). \quad (2.40)$$

Let A_* be a set for which $|P_n(A_*) - P(A_*)| > \varepsilon$ or if there is no such A_* , then A_* is fixed arbitrarily. The following holds

$$\begin{aligned} \mathbb{P}\left(\sup_{A \in \mathcal{A}} |P_n(A) - P'_n(A)| > \varepsilon/2\right) &\geq \mathbb{P}\left(|P_n(A_*) - P'_n(A_*)| > \varepsilon/2\right) \\ &\geq \mathbb{P}\left(|P_n(A_*) - P(A_*)| > \varepsilon, |P'_n(A_*) - P(A_*)| < \varepsilon/2\right) \\ &= \mathbb{E}\left[\mathbb{I}\left(|P_n(A_*) - P(A_*)| > \varepsilon\right) \mathbb{P}\left(|P'_n(A_*) - P(A_*)| < \varepsilon/2 \mid Z_1, \dots, Z_n\right)\right]. \end{aligned} \quad (2.41)$$

We proceed by applying Chebyshev's inequality as

$$\mathbb{P}\left(|P'_n(A_*) - P(A_*)| < \varepsilon/2 \mid Z_1, \dots, Z_n\right) \geq 1 - \frac{P(A_*)(1 - P(A_*))}{n \frac{\varepsilon^2}{4}}. \quad (2.42)$$

Since $P(A_*)(1 - P(A_*)) \leq 1/4$ and we can assume that $n\varepsilon^2 \geq 2$, otherwise the upper bound is trivial in the theorem, it follows that $\mathbb{P}\left(|P'_n(A_*) - P(A_*)| < \varepsilon/2 \mid Z_1, \dots, Z_n\right) \geq 1/2$. Substituting it back to (2.41) the inequality in (2.40) is proved.

In the second step we are going to symmetrize with an i.i.d. uniform random sign sample, that is $\sigma_1, \dots, \sigma_n$ where $\mathbb{P}(\sigma_1 = +1) = \mathbb{P}(\sigma_1 = -1) = 1/2$. Clearly as $Z_1, Z'_1, \dots, Z_n, Z'_n$ are i.i.d. $\sup_{A \in \mathcal{A}} |P_n(A) - P'_n(A)| = \sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^n \mathbb{I}(Z_i \in A) - \mathbb{I}(Z'_i \in A) \right|$ has the same distribution as $\sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i (\mathbb{I}(Z_i \in A) - \mathbb{I}(Z'_i \in A)) \right|$. Applying this remark and the union bound we conclude that

$$\begin{aligned} &\mathbb{P}\left(\sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^n (\mathbb{I}(Z_i \in A) - \mathbb{I}(Z'_i \in A)) \right| > \varepsilon/2\right) \\ &= \mathbb{P}\left(\sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i (\mathbb{I}(Z_i \in A) - \mathbb{I}(Z'_i \in A)) \right| > \varepsilon/2\right) \end{aligned}$$

$$\begin{aligned}
 &\leq \mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i(\mathbb{I}(Z_i \in A)) \right| > \varepsilon/4 \right) + \mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i(\mathbb{I}(Z'_i \in A)) \right| > \varepsilon/4 \right) \\
 &\leq 2 \mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i(\mathbb{I}(Z_i \in A)) \right| > \varepsilon/4 \right). \tag{2.43}
 \end{aligned}$$

In the third step we bound $\mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i(\mathbb{I}(Z_i \in A)) \right| > \varepsilon/4 \right)$ by conditioning on Z_1, \dots, Z_n . Fix a possible outcome z_1, \dots, z_n . Notice that there are at most $\tau_{\mathcal{A}}(n)$ different $(\mathbb{I}(z_1 \in A), \dots, \mathbb{I}(z_n \in A))$ vectors, because each such vector corresponds to an intersection of form $A \cap \{z_1, \dots, z_n\}$ for an $A \in \mathcal{A}$ and $\tau_{\mathcal{A}}(n)$ was the maximum of these. Therefore $\sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i(\mathbb{I}(Z_i \in A)) \right|$ is the maximum of at most $\tau_{\mathcal{A}}(n)$ variables, therefore we can apply the union bound to obtain

$$\begin{aligned}
 &\mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i(\mathbb{I}(Z_i \in A)) \right| > \varepsilon/4 \mid Z_1, \dots, Z_n \right) \\
 &\leq \tau_{\mathcal{A}}(n) \sup_{A \in \mathcal{A}} \mathbb{P} \left(\frac{1}{n} \left| \sum_{i=1}^n \sigma_i(\mathbb{I}(Z_i \in A)) \right| > \varepsilon/4 \mid Z_1, \dots, Z_n \right), \tag{2.44}
 \end{aligned}$$

with a supremum outside the probability.

In step four we apply Hoeffding's inequality, see Theorem [A.3.1.1](#), for bounding

$$\mathbb{P} \left(\frac{1}{n} \left| \sum_{i=1}^n \sigma_i(\mathbb{I}(Z_i \in A)) \right| > \varepsilon/4 \mid Z_1, \dots, Z_n \right). \tag{2.45}$$

For a fixed z_1, \dots, z_n variable $\sum_{i=1}^n \sigma_i(\mathbb{I}(z_i \in A))$ is the sum of n independent, zero mean variables bounded between -1 and $+1$. We apply Hoeffding's inequality to obtain

$$\mathbb{P} \left(\frac{1}{n} \left| \sum_{i=1}^n \sigma_i(\mathbb{I}(Z_i \in A)) \right| > \varepsilon/4 \mid Z_1, \dots, Z_n \right) \leq 2 e^{-n\varepsilon^2/32}, \tag{2.46}$$

hence

$$\mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i(\mathbb{I}(Z_i \in A)) \right| > \varepsilon/4 \mid Z_1, \dots, Z_n \right) \leq \tau_{\mathcal{A}}(n) 2 e^{-n\varepsilon^2/32}. \tag{2.47}$$

Taking the expected value on both sides yields

$$\mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i(\mathbb{I}(Z_i \in A)) \right| > \varepsilon/4 \right) \leq \tau_{\mathcal{A}}(n) 2 e^{-n\varepsilon^2/32} \tag{2.48}$$

The theorem is proved by putting these steps together. \square

Corollary 2.7.3.1. *Let \mathcal{F} be a class of decision rules. Let $\hat{R}(g)$ be the empirical risk and $R(g)$ the expected risk, then the following holds*

$$\mathbb{P} \left(\sup_{g \in \mathcal{F}} |\hat{R}(g) - R(g)| > \varepsilon \right) \leq 8 \tau_{\mathcal{F}}(n) e^{-n\varepsilon^2/32}. \tag{2.49}$$

Proof. Let $\mathcal{A} = \{A(g) \mid g \in \mathcal{F}\}$ for which we know that $\tau_{\mathcal{A}} = \tau_{\mathcal{F}}$ and notice that

$$\begin{aligned} & \left| \frac{1}{n} \sum_{i=1}^n \mathbb{I}(g(X_i) \neq Y_i) - \mathbb{P}(g(X) \neq Y) \right| \\ &= \left| \frac{1}{n} \sum_{i=1}^n (1 - \mathbb{I}(g(X_i) = Y_i)) - (1 - \mathbb{P}(g(X) = Y)) \right| = \left| P_n(A(g)) - P(A(g)) \right|. \end{aligned} \quad (2.50)$$

The corollary then follows from Theorem 2.7.2. \square

This corollary can be used to finish the proof of the fundamental theorem of PAC learning.

Proof of Theorem 2.7.1. We need to show that when \mathcal{F} has a finite dimension then for all $\varepsilon, \delta > 0$ there exists $n_{\mathcal{F}}^U(\varepsilon, \delta)$ such that for all distributions P whenever $|\mathcal{D}| > n_{\mathcal{F}}^U(\varepsilon, \delta)$ we have

$$\mathbb{P} \left(\sup_{g \in \mathcal{F}} |\hat{R}(g) - R(g)| \leq \varepsilon \right) \geq 1 - \delta. \quad (2.51)$$

We apply Corollary 2.7.3.1 and Corollary 2.6.1.1 to obtain

$$\mathbb{P} \left(\sup_{g \in \mathcal{F}} |\hat{R}(g) - R(g)| > \varepsilon \right) \leq 8 \tau_{\mathcal{F}}(n) e^{-n\varepsilon^2/32} \leq 8 \left(\frac{2n}{V_{\mathcal{F}}} \right)^{V_{\mathcal{F}}} e^{-n\varepsilon^2/32} \quad (2.52)$$

We know that $8 \left(\frac{2n}{V_{\mathcal{F}}} \right)^{V_{\mathcal{F}}} e^{-n\varepsilon^2/32}$ goes to 0 as $n \rightarrow \infty$, so for all given δ and ε we can find a number $n_{\mathcal{F}}^U(\varepsilon, \delta)$ such that $8 \left(\frac{2n}{V_{\mathcal{F}}} \right)^{V_{\mathcal{F}}} e^{-n\varepsilon^2/32} < \delta$ holds for any $n > n_{\mathcal{F}}^U(\varepsilon, \delta)$. As an example we show that any natural number $n_{\mathcal{F}}^U(\varepsilon, \delta)$ for which

$$n_{\mathcal{F}}^U(\varepsilon, \delta) > \frac{64 \left(\log^{8/\delta} + \log(2) \right)}{\varepsilon^2} + \left(\frac{V_{\mathcal{F}}}{64\varepsilon^2} \right)^2 \quad (2.53)$$

holds, works fine. Notice that for $n > n_{\mathcal{F}}^U(\varepsilon, \delta)$ the following holds

$$\frac{V_{\mathcal{F}}}{64\varepsilon^2} \leq \sqrt{n} = \frac{n}{n^{1/2}} \leq \frac{n}{\log n}. \quad (2.54)$$

Furthermore, using (2.54) in the second inequality yields

$$\begin{aligned} \log \left(\frac{8}{\delta} \right) + \log(2) &\leq n \left(\varepsilon^2/32 - V_{\mathcal{F}} \frac{\varepsilon^2}{V_{\mathcal{F}} 64} \right) \leq \frac{n\varepsilon^2}{32} - V_{\mathcal{F}} \log(n) \\ &\leq \frac{n\varepsilon^2}{32} - V_{\mathcal{F}} \log(n) + V_{\mathcal{F}} \log(V_{\mathcal{F}}). \end{aligned} \quad (2.55)$$

We obtain the following by subtracting $\log(2)$ from each side, then multiplying them by -1 and taking the exponential in both sides

$$\frac{\delta}{8} \geq \left(\frac{2n}{V_{\mathcal{F}}} \right)^{V_{\mathcal{F}}} e^{-n\varepsilon^2/32}. \quad (2.56)$$

It is what we wanted to see. The existence of $n_{\mathcal{F}}^U(\varepsilon, \delta)$ ensures that \mathcal{F} possesses the uniform convergence property. \square

Working through a more involved argument it can be showed that there are positive constants C_1 and C_2 such that

$$C_1 \frac{(\log 1/\delta + V_{\mathcal{F}})}{\varepsilon^2} \leq n_{\mathcal{F}}^U(\varepsilon, \delta) \leq C_2 \frac{(\log 1/\delta + V_{\mathcal{F}})}{\varepsilon^2}. \quad (2.57)$$

2.8 Structural Risk Minimization

We provided necessary and sufficient conditions for PAC learnability via uniform convergence and VC dimension for the problem of binary classification. We also showed that the ERM method is a successful PAC learner in this case. One may say that we argued that ERM is the best learning algorithm that we can imagine. It is not the whole truth for many reasons. First of all notice that we did not say anything about the feasibility of the problem of finding an ERM estimator. The sad fact is that it is often hopeless to minimize the empirical risk. In addition, it is important to mention that we only analyzed the statistical property of learning algorithms, however computational and implementational aspects should also be an important factor in learning algorithm design. For example it can happen that even though an empirical risk minimizer can be carried out for our problem in hand, still we prefer to apply a simpler or computationally lighter method, because it takes less time both to implement and run.

Besides, we mentioned above that PAC learnability is too restrictive in some sense. So far, we fixed a model class, \mathcal{F} , and tried to determine a good sample size which is sufficient to PAC learn \mathcal{F} . Notice that in real-world problems we usually do not know exactly \mathcal{F} in advance. In practice often the data size is fixed and it is part of the learning task to find a proper model class for the given sample size. Therefore it is worth switching the perspective and ask what is the lowest accuracy we can guarantee for a model class with VC dimension $V_{\mathcal{F}}$ from a fixed sample size n . Notice that when we have a fixed confidence parameter δ and data size n we can bound the accuracy parameter ε . The following theorem was deduced from (2.53) with the help of [19].

Theorem 2.8.1. *Consider the problem of binary classification with the 0/1 loss. Given a model class \mathcal{F} with VC dimension $V_{\mathcal{F}} < \infty$ and an i.i.d. sample with n elements. Then for all $g \in \mathcal{F}$, $\delta > 0$ and*

$$\varepsilon \geq \sqrt{\frac{32 \left(\log \frac{8}{\delta} + V_{\mathcal{F}} \log \left(\frac{2n}{V} + 1 \right) \right)}{n}} \quad (2.58)$$

we have that

$$\mathbb{P} \left(R(g) \leq \hat{R}(g) + \varepsilon \right) \geq 1 - \delta. \quad (2.59)$$

Proof. The following simple implications hold.

$$\varepsilon \geq \sqrt{\frac{32 \left(\log \frac{8}{\delta} + V_{\mathcal{F}} \log \left(\frac{2n}{V} + 1 \right) \right)}{n}} \Rightarrow \varepsilon^2 \geq \frac{32 \left(\log \frac{8}{\delta} + V_{\mathcal{F}} \log \left(\frac{2n}{V} + 1 \right) \right)}{n} \quad (2.60)$$

$$\left(\frac{-n\varepsilon^2}{32} \right) \leq \log \frac{\delta}{8} - V_{\mathcal{F}} \log \left(\frac{2n}{V} + 1 \right) < \log \frac{\delta}{8} - V_{\mathcal{F}} \log \left(\frac{2n}{V} \right) \quad (2.61)$$

It is easy to show that from (2.61) the inequality $8 \left(\frac{2n}{V_{\mathcal{F}}} \right)^{V_{\mathcal{F}}} e^{-n\varepsilon^2/32} < \delta$ follows. Then Corollary 2.7.3.1 says that

$$\mathbb{P} \left(\sup_{g \in \mathcal{F}} |\hat{R}(g) - R(g)| \leq \varepsilon \right) \geq 8 \tau_{\mathcal{F}}(n) e^{-n\varepsilon^2/32} \geq 1 - \delta, \quad (2.62)$$

from which it is clear that for any $g \in \mathcal{F}$ we have

$$1 - \delta \leq \mathbb{P} \left(|\hat{R}(g) - R(g)| \leq \varepsilon \right) \leq \mathbb{P} \left(R(g) \leq \hat{R}(g) + \varepsilon \right), \quad (2.63)$$

hence the theorem is proved. \square

The idea of *structural risk minimization* (SRM) is that the quantity $\hat{R}(g) + \varepsilon$ should be minimized instead of $\hat{R}(g)$ so that (2.60) holds for ε . This way we can find the tightest bound on the true risk with at least a chosen high probability. We refer to $\hat{R}(g) + \varepsilon$ as the structural risk. In SRM we consider a sequence of model classes with increasing VC dimension, usually $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots$, then we select the one which has the smallest structural risk. Later we are going to present the method of Support Vector Machines (SVM) which is an important application of this principle.

It is important to see that SRM goes beyond the limitation of PAC learnability, since we do not fix the model class in this case. In fact, it can be showed that SRM is a nonuniform learner. The notion of nonuniform learnability is a strict relaxation of PAC learnability, where the sample size, which is required for the PAC statement, is allowed to depend on the model we compete against. For further readings on this interesting topic see [19] and [23].

2.9 Uniform Law of Large Numbers

Recall that the need for introducing a model class \mathcal{F} arised when we realized that

$$\inf_{g \in \mathcal{F}} \hat{R}(g) \xrightarrow{a.s.} \inf_{g \in \mathcal{F}} R(g) \quad (2.64)$$

does not hold in general. In this section we examine this phenomenon more thoroughly. The core idea of ERM was that for any fixed model g the empirical risk converges to the true risk, i.e. for i.i.d. variables $\{(X_i, Y_i)\}_{i \in \mathbb{N}}$ when $\mathbb{E}L(f(X_i), Y_i) < \infty$ we have

$$\frac{1}{n} \sum_{i=1}^n L(f(X_i), Y_i) - \mathbb{E}L(f(X_i), Y_i) \xrightarrow{a.s.} 0. \quad (2.65)$$

Observe that for any given model class \mathcal{F} the following holds

$$\left| \inf_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n L(f(X_i), Y_i) - \inf_{f \in \mathcal{F}} \mathbb{E}L(f(X_i), Y_i) \right| \leq \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n L(f(X_i), Y_i) - \mathbb{E}L(f(X_i), Y_i) \right| \quad (2.66)$$

therefore, to apply the ERM method we need to ensure that

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n L(f(X_i), Y_i) - \mathbb{E}L(f(X_i), Y_i) \right| \rightarrow 0, \quad (2.67)$$

where the convergence can be in probability or almost surely. In this section we consider a more general problem which is called the uniform law of large numbers (ULLN).

Definition 2.9.1. Let Z_1, Z_2, \dots be i.i.d. random variables taking values in \mathbf{Z} . For a class \mathcal{F} of $\mathbf{Z} \rightarrow \mathbb{R}$ functions, with property $\mathbb{E}f(Z_1) < \infty$ for all $f \in \mathcal{F}$ the **uniform law of large numbers** holds if the following convergence occurs

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(Z_i) - \mathbb{E}f(Z_1) \right| \xrightarrow{p} 0. \quad (2.68)$$

In the literature many say that \mathcal{F} is a *Glivenko-Cantelli class* for the distribution of Z_1 when the ULLN holds. We are going to derive conditions on \mathcal{F} in order to ensure the ULLN. It is easy to see that the notion of uniform convergence provides a sufficient condition for ULLN, but since uniform convergence requires universal non-asymptotic guarantees, it is not necessary.

2.9.1 Frequencies to Their Probabilities

First, we present a lighter sufficient condition for the ULLN of classifiers. Notice that in this case the expected loss is the misclassification probability and the mean is a frequency.

Recall the definition of the growth function, $\tau_{\mathcal{F}}(n) = \tau_{\mathcal{A}}(n)$. It was the maximum of $|\mathcal{A}_C|$ on sets $C = \{z_1, \dots, z_n\}$. Let $\mathcal{N}_{\mathcal{A}}(\{z_1, \dots, z_n\}) \doteq |\mathcal{A}_{\{z_1, \dots, z_n\}}|$.

Theorem 2.9.1. For every probability distribution P and a class of measurable sets \mathcal{A} where $A \subseteq \mathbb{X} \times \{+1, -1\}$ for all $A \in \mathcal{A}$, the following holds

$$\mathbb{P} \left(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \varepsilon \right) \leq 8 \mathbb{E} \mathcal{N}_{\mathcal{A}}(\{Z_1, \dots, Z_n\}) e^{-n\varepsilon^2/32}. \quad (2.69)$$

Proof. The proof from [8, Theorem 12.5] is similar to the one which was presented for Theorem 2.7.2. In fact the first two steps are identical to that. The difference is going to be in the third step. There, we bounded the quantity $\mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i(\mathbb{I}(Z_i \in A)) \right| > \varepsilon/4 \right)$ by conditioning on Z_1, \dots, Z_n . We proceed similarly, but notice that for any fixed z_1, \dots, z_n there are at most $\mathcal{N}_{\mathcal{F}}(\{z_1, \dots, z_n\})$ different vectors with form $(\mathbb{I}(z_1 \in A), \dots, \mathbb{I}(z_n \in A))$,

because such vectors correspond to intersections of form $A \cap \{z_1, \dots, z_n\}$ for $A \in \mathcal{A}$. Thus

$$\begin{aligned} & \mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i(\mathbb{I}(Z_i \in A)) \right| > \varepsilon/4 \mid Z_1, \dots, Z_n \right) \\ & \leq \mathcal{N}_{\mathcal{A}}(\{Z_1, \dots, Z_n\}) \sup_{A \in \mathcal{A}} \mathbb{P} \left(\frac{1}{n} \left| \sum_{i=1}^n \sigma_i(\mathbb{I}(Z_i \in A)) \right| > \varepsilon/4 \mid Z_1, \dots, Z_n \right) \\ & \leq \mathcal{N}_{\mathcal{A}}(\{Z_1, \dots, Z_n\}) 2 \exp \left(-\frac{n\varepsilon^2}{32} \right), \end{aligned} \quad (2.70)$$

where the second inequality is justified exactly as the fourth step in the proof of Theorem 2.7.2. Taking expectation and putting together the pieces yield the theorem. \square

Similarly as before, choosing the proper set system based on the classifiers in \mathcal{F} we obtain

$$\mathbb{P} \left(\sup_{g \in \mathcal{F}} |\hat{R}(g) - R(g)| > \varepsilon \right) \leq 8 \mathbb{E} \mathcal{N}_{\mathcal{F}}(\{Z_1, \dots, Z_n\}) e^{-n\varepsilon^2/32}. \quad (2.71)$$

Therefore a uniform law of large numbers holds whenever $\mathbb{E} \mathcal{N}_{\mathcal{F}}(\{Z_1, \dots, Z_n\}) e^{-n\varepsilon^2/32} \rightarrow 0$ as $n \rightarrow \infty$. Notice that $\mathbb{E} \mathcal{N}_{\mathcal{F}}(\{Z_1, \dots, Z_n\})$ depends on the distribution of Z_1 therefore it is harder to handle than the purely combinatorial VC dimension. With a more involved argument Vapnik and Chervonenkis proved that $\frac{\mathbb{E} \log_2 \mathcal{N}_{\mathcal{F}}(\{Z_1, \dots, Z_n\})}{n} \rightarrow 0$ is a necessary and sufficient condition for frequencies to converge to their corresponding probabilities, see [23]. They call the quantity $\mathbb{E} \log_2 \mathcal{N}_{\mathcal{F}}(\{Z_1, \dots, Z_n\})$ entropy.

2.9.2 Means to Their Expectations

In this section we present a similar sufficient condition for the general version of the ULLN via covering numbers and Rademacher complexity. In the previous section the idea was that after a symmetrization the probability $\mathbb{P} \left(\sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i(\mathbb{I}(Z_i \in A)) \right| > \varepsilon \right)$ could be bounded by a maximum of finitely many terms times a complexity measure. Now we generalize this idea for bounded $\mathbb{X} \rightarrow [0, B]$ type functions. Assume for this part that $\mathbb{X} \subseteq \mathbb{R}^d$.

Definition 2.9.2. (ε -covers) *Let \mathcal{F} be a set of bounded real-valued functions. For $\varepsilon > 0$ we say that $\mathcal{F}(\varepsilon, \|\cdot\|) = \{f_1, \dots, f_l\}$ is an ε -cover of \mathcal{F} with respect to norm $\|\cdot\|$ if for all $f \in \mathcal{F}$ there is an index $j(f)$ such that*

$$\|f - f_{j(f)}\| < \varepsilon. \quad (2.72)$$

Let $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|)$ denote the size of the smallest ε -cover of \mathcal{F} w.r.t. norm $\|\cdot\|$. Take $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|) = \infty$ if no ε -cover exists. We call $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|)$ the ε -covering number of \mathcal{F} w.r.t. the given norm. The most important examples of these covers are induced by the supremum norm or the L_p norms. We consider the L_p norms with an empirical measure. Let $\mathbf{x} = \{x_1, \dots, x_n\}$ be an arbitrary set. Then, the empirical measure with respect to \mathbf{x} is defined similarly as before

$$P_n(A) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(x_i \in A). \quad (2.73)$$

The L_p norm then becomes

$$\|f\|_{L_p(P_n)} = \left(\frac{1}{n} \sum_{i=1}^n |f(x_i)|^p \right)^{1/p}. \quad (2.74)$$

For the sake of simplicity we denote the ε -cover with respect to the $L_p(P_n)$ norm by $\mathcal{N}_p(\varepsilon, \mathcal{F}, \mathbf{x})$. Notice that when we substitute a random sample in the place of \mathbf{x} , the quantity $\mathcal{N}_p(\varepsilon, \mathcal{F}, \mathbf{x})$ becomes a random variable.

Theorem 2.9.2. *Let \mathcal{F} be a set of functions $f : \mathbb{X} \rightarrow [0, B]$ and $\mathbf{X} = \{X_i\}_{i=1}^n$ an i.i.d. sample taking values from \mathbb{X} . For any $n \in \mathbb{N}$, and any $\varepsilon > 0$,*

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}f(X_i) \right| > \varepsilon\right) \leq 8 \mathbb{E}\mathcal{N}_1(\varepsilon/8, \mathcal{F}, \mathbf{X}) e^{-n\varepsilon^2/(128 B^2)}. \quad (2.75)$$

The proof of this theorem from [11] can be found in the appendix, see [B.2]. Similarly as in (2.71), from Theorem 2.9.2 it follows easily that the ULLN holds whenever $\mathbb{E}\mathcal{N}_1(\varepsilon/8, \mathcal{F}, \mathbf{X}) e^{-n\varepsilon^2/(128 B^2)} \rightarrow 0$.

Vapnik proves in [23] that a necessary and sufficient condition for ULLN of means to their expectations for a bounded family of functions can be formulated via the so-called expected entropy or ε -entropy which is closely related to the concept of covering numbers. For further details see the final chapter of Statistical Learning Theory, [23].

In addition, in the proof the quantity $\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i(f(X_i)) \right|$ plays a crucial role and a similar quantity was extremely important in the proof of the Vapnik–Chervonenkis theorem. It is formally the maximum correlation between a vector $(f(X_1), \dots, f(X_n))$ and a random “noise” vector $(\sigma_1, \dots, \sigma_n)$, where the maximum is taken over the model class. Intuitively a function class is too rich when there is always a model which highly correlates with a random noise. The expectation of this quantity is called the Rademacher complexity of the model class, which also provides a useful concept for measuring complexity.

Finally, the notion of packing numbers is introduced, which is closely related to covering numbers and VC dimension.

Definition 2.9.3. (ε -packings) *Let \mathcal{F} be a set of real-valued functions. For $\varepsilon > 0$ we say that $\mathcal{F}(\varepsilon, \|\cdot\|) = \{f_1, \dots, f_l\}$ is an ε -packing of \mathcal{F} with respect to norm $\|\cdot\|$ if for all $1 \leq i < j \leq l$*

$$\|f_i - f_j\| > \varepsilon. \quad (2.76)$$

Let $\mathcal{M}(\varepsilon, \mathcal{F}, \|\cdot\|)$ be the size of the largest ε -packing of \mathcal{F} w.r.t. norm $\|\cdot\|$ (it can be infinity). We say that $\mathcal{M}(\varepsilon, \mathcal{F}, \|\cdot\|)$ is the ε -packing number of \mathcal{F} w.r.t. norm $\|\cdot\|$.

The following lemma, which is going to be used in Chapter 5, yields a close relationship between packing and covering numbers.

Lemma 2.9.3. *Let \mathcal{F} be a class of functions on \mathbb{R}^d and $\|\cdot\|$ be a norm on \mathcal{F} and $\varepsilon > 0$. Then*

$$\mathcal{M}(2\varepsilon, \mathcal{F}, \|\cdot\|) \leq \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|) \leq \mathcal{M}(\varepsilon, \mathcal{F}, \|\cdot\|). \quad (2.77)$$

Proof. Let $\{f_1, \dots, f_M\}$ be a 2ε -packing of \mathcal{F} w.r.t. $\|\cdot\|$ with maximal cardinality. Then for any ε -cover $\{g_1, \dots, g_N\}$ consider balls $B(g_i, \varepsilon)$ for $i \in [N]$. It is easy to see that each $B(g_i, \varepsilon)$ can contain at most one f_j . In addition, all members of the packing are covered, therefore the first inequality is proved.

For the second inequality let $\{f_1, \dots, f_M\}$ be an ε -packing of \mathcal{F} w.r.t. $\|\cdot\|$ with maximal cardinality. Then for any $h \in \mathcal{F}$ the set $\{h, f_1, \dots, f_M\}$ is not an ε -packing, therefore there exists $j \in [M]$ such that

$$\|h - f_j\| < \varepsilon, \quad (2.78)$$

i.e. $\{f_1, \dots, f_M\}$ is an ε -cover of \mathcal{F} w.r.t. $\|\cdot\|$. \square

These concepts are also strongly related to the notion of VC dimension, which can be defined for set systems similarly to Definition 2.6.1.

Definition 2.9.4. (*VC dimension*) Let \mathcal{A} be a class of subsets of $\mathbb{X} \times \mathbb{Y}$. The VC dimension ($V_{\mathcal{A}}$) of \mathcal{A} is the largest integer n such that there exists n points in $\mathbb{X} \times \mathbb{Y}$, $C = \{c_1, \dots, c_n\}$ that can be shattered by \mathcal{A} , i.e. for all $\tilde{C} \subseteq C$ there is $A \in \mathcal{A}$ such that $A \cap C = \tilde{C}$.

Let f be a real-valued function. The subgraph of f is defined as

$$f^+ \doteq \{(x, t) \in \mathbb{X} \times \mathbb{R} \mid t \leq f(x)\}. \quad (2.79)$$

The VC dimension of subgraphs can be interpreted by the definition above when $\mathbb{Y} = \mathbb{R}$. Let \mathcal{F} be a class of real-valued function, then

$$\mathcal{F}^+ \doteq \{ \{(x, t) \in \mathbb{X} \times \mathbb{R} \mid t \leq f(x)\}, f \in \mathcal{F} \} \quad (2.80)$$

contains all subgraphs of functions in \mathcal{F} . The following inequality yields a quantitative relationship between the packing numbers of \mathcal{F} w.r.t. the L_1 norm and the VC dimension of \mathcal{F}^+ .

Theorem 2.9.4. Let μ be a probability measure on \mathbb{R}^d , let \mathcal{F} be a class of μ -measurable $f : \mathbb{R}^d \rightarrow [0, 1]$ functions with $V_{\mathcal{F}^+} < \infty$, and let $\varepsilon > 0$. Then

$$\mathcal{M}(\varepsilon, \mathcal{F}, \|\cdot\|_{L_1(\mu)}) \leq e(V_{\mathcal{F}^+} + 1) \left(\frac{2e}{\varepsilon} \right)^{V_{\mathcal{F}^+}}. \quad (2.81)$$

The proof can be found in [12].

2.10 Strong Uniform Laws

So far we only proved the weak ULLN, but it is important to see that the presented results can be used easily to ensure almost sure convergence as well.

We are going to use the Borel–Cantelli lemma to provide sufficient conditions for strong ULLN based on Theorem 2.9.1 and Theorem 2.9.2, see [11, Theorem 9.1].

Theorem 2.10.1. *Let \mathcal{F} be a set of functions $f : \mathbb{X} \rightarrow [0, B]$ and $\mathbf{X} = \{X_i\}_{i=1}^n$ an i.i.d. sample taking values from \mathbb{X} . If*

$$\sum_{n=1}^{\infty} \mathbb{E} \mathcal{N}_1(\varepsilon/8, \mathcal{F}, \mathbf{X}) e^{-n\varepsilon^2/(128B^2)} < \infty \quad (2.82)$$

holds for all $\varepsilon > 0$, then the strong ULLN occurs, i.e.

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}f(X_1) \right| \xrightarrow{a.s.} 0. \quad (2.83)$$

Proof. Let $\varepsilon_k = 1/k$ and events $A_{n,k} \doteq \left\{ \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}f(X_1) \right| > 1/k \right\}$. From the Borel-Cantelli lemma and by the assumption of the theorem

$$\limsup_{n \rightarrow \infty} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}f(X_1) \right| \leq \frac{1}{k} \quad a.s., \quad (2.84)$$

and consequently almost surely

$$\limsup_{n \rightarrow \infty} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}f(X_1) \right| \leq \frac{1}{k} \quad \forall k \in \mathbb{N}, \quad (2.85)$$

which implies that the strong ULLN holds. \square

Similarly the following can be proved.

Theorem 2.10.2. *For every probability distribution P and a class of $\mathbb{X} \rightarrow \{+1, -1\}$ type classifiers, \mathcal{F} , if for all $\varepsilon > 0$*

$$\sum_{n=1}^{\infty} \mathbb{E} \mathcal{N}_{\mathcal{F}}(\{Z_1, \dots, Z_n\}) e^{-n\varepsilon^2/32} < \infty \quad (2.86)$$

then, the strong ULLN holds, i.e.

$$\sup_{g \in \mathcal{F}} \left| \hat{R}(g) - R(g) \right| \xrightarrow{a.s.} 0. \quad (2.87)$$

2.11 Universal Consistency

We argued that the model class of all classifiers is not PAC learnable, still our hope is that a lighter asymptotic learning concept can be achieved for any measurable function. In this section we deal with the notion of *consistency*.

Recall that in Chapter 1 we defined our goals as risk minimizations both for classification and regression. For consistency we only require that our estimate's risk converges to the optimal Bayes risk in both cases, formally

Definition 2.11.1. (*consistency of classifier estimates*) *Consider the problem of binary classification with the 0/1 loss. Let $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ denote an i.i.d. sample from the (X, Y)*

random pair's distribution, P . We say that a classifier estimate sequence, g_n , is consistent for distribution P if

$$\mathbb{P}(g_n(X) \neq Y \mid \mathcal{D}_n) \xrightarrow{P} \min_g \mathbb{P}(g(X) \neq Y). \quad (2.88)$$

When the convergence holds almost surely we say that the estimate sequence, or simply the estimate is *strongly consistent*. When the convergence occurs for all possible distributions of the random pair (X, Y) , we say that the estimate sequence or learning rule is universal. Consistency is defined for regression similarly.

Definition 2.11.2. (*consistency of regression estimates*) Consider the problem of regression. Let $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ denote an i.i.d. sample from the (X, Y) random pair's distribution, P . We say that a sequence of regression function estimates, f_n , is consistent for distribution P if

$$\mathbb{E}((f_n(X) - Y)^2 \mid \mathcal{D}_n) \xrightarrow{P} \min_f \mathbb{E}[f(X) - Y]^2. \quad (2.89)$$

When the convergence holds almost surely we say that the estimate or estimate sequence is *strongly consistent*.

Definition 2.11.3. (*strong universal consistency of classifier estimates*) A sequence of regression function estimates, f_n , is *universally consistent* if it is strongly consistent for all distributions of (X, Y) with $\mathbb{E}Y^2 < \infty$.

Recall that in (1.11) we showed that

$$\mathbb{E}((f_n(X) - Y)^2 \mid \mathcal{D}) = \int_{\mathbb{X}} (f_n(x) - f_*(x))^2 dP_X(x) + \mathbb{E}[(f_*(X) - Y)^2], \quad (2.90)$$

where $\mathbb{E}[(f_*(X) - Y)^2]$ is constant. Thus, f_n is consistent if and only if the quantity $\int_{\mathbb{X}} (f_n(x) - f_*(x))^2 dP_X(x)$ tends to zero in probability and f_n is strongly consistent if and only if the quantity $\int_{\mathbb{X}} (f_n(x) - f_*(x))^2 dP_X(x)$ tends to zero almost surely. In the book of Györfi et al., [11], it is the definition of consistency. I decided to use the slightly different version to clarify the connection between the consistency of regression and classification.

In the next chapter we are going to see that there exists nonparametric regression function estimates that are strongly universally consistent. We have already proved that in case of the problem of binary classification such regression estimates induce a classifier sequence which inherits the strong universal consistency, see (1.13). In the next chapter we only deal with the problem of regression, nevertheless in the Chapter 5 we are going to use the presented methods to estimate the regression function for classification problems.

Chapter 3

Nonparametric Methods

In this chapter we introduce a simple nonparametric method which is strongly universally consistent. Namely, we present a thorough analysis of the local averaging kernel estimate. We are going to present the proof of its strong consistency. In addition, we are going to define the k -nearest neighbors estimate, which have similar properties. In Chapter 4 these kernel estimations are generalized by the theory of reproducing kernel Hilbert spaces (RKHS). This chapter is based on the books of Györfi et al., see [11], which deals with nonparametric regression, and [8], which examines classification.

3.1 Local Averaging Estimates

In the book of Györfi et al. four paradigms are introduced. Local averaging is the simplest approach. Local modeling, global modeling and regularized modeling are three more elaborated versions of nonparametric regression estimation. In this thesis we limit ourselves to the local averaging estimates, because they are good enough examples for illustrating strong universal consistency.

Often we view regression as a function approximation problem from noisy observations, that is we assume that the following holds for the datapoints

$$Y_i = f_*(X_i) + \varepsilon_i \quad (3.1)$$

for $i \in [n]$, where X_i and ε_i are random and it is easy to see that $\mathbb{E}(\varepsilon_i | X_i) = 0$. Notice that Y_i can be viewed as the sum of the regression function in X_i plus some noise with zero expectation. It motivates us to try to locally average out the noise term. To define locality we require \mathbb{X} to be a metric space. For simplicity for this whole chapter we assume that $\mathbb{X} = \mathbb{R}^d$ endowed with the standard euclidean metric. We usually define a local averaging estimate by

$$f_n(x) = \sum_{i=1}^n w_{n,i}(x) Y_i, \quad (3.2)$$

where $w_{n,i}(x)$ are nonnegative weights often sum up to 1, dependent of the inputs, $\{X_i\}_{i=1}^n$.

3.2 Stone's Theorem

We are going to apply Stone's theorem which provides sufficient conditions on the weights for universal consistency of local averaging estimates. The proof of Stone's theorem can be found in the appendix, see [B.3](#), where we take advantage of the fact that the set of continuous functions of bounded support is dense in $L_p(\mu)$ for any $p \geq 1$ and probability measure μ . For the proof of this auxiliary statement see [\[11\]](#), Theorem A.1]. We refer to this theorem as the denseness result later on.

Theorem 3.2.1. (*Stone's theorem*) *Let X be a variable identically distributed as X_1 and independent of the given sample. Assume that for all possible distributions of X :*

i There exists a constant c such that for every (measurable) nonnegative function f satisfying $\mathbb{E}f(X) < \infty$ and for any $n \in \mathbb{N}$ the following holds

$$\mathbb{E}\left(\sum_{i=1}^n |w_{n,i}(X)| f(X_i)\right) \leq c \mathbb{E}f(X). \quad (3.3)$$

ii There is $D \geq 1$ such that for all $n \in \mathbb{N}$

$$\mathbb{P}\left(\sum_{i=1}^n |w_{n,i}(X)| \leq D\right) = 1. \quad (3.4)$$

iii For all $a > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{E}\left(\sum_{i=1}^n |w_{n,i}(X)| \mathbb{I}(\|X_i - X\| > a)\right) = 0. \quad (3.5)$$

iv

$$\sum_{i=1}^n w_{n,i}(X) \xrightarrow{p} 0 \quad (3.6)$$

v

$$\lim_{n \rightarrow \infty} \mathbb{E}\left(\sum_{i=1}^n w_{n,i}^2(X)\right) = 0 \quad (3.7)$$

Then $f_n(x) = \sum_{i=1}^n w_{n,i}(x) Y_i$ is universally consistent.

Assumption *ii* and *iv* ensures that the sum of the weights is bounded and tends to 1. Assumption *iii* says that $f_n(x)$ is only influenced in the long run by the sample points that are in the neighborhood of x and assumption *v* says that the weights are asymptotically vanishing. Assumption *i* is rather technical. For noiseless regression it says that the expected value of the estimate is at most a constant times the expected value of the regression function. The proof from [\[11\]](#), Theorem 4.1] can be found in the appendix, see [B.3](#).

3.3 Kernel Estimates

In this section we define the local averaging kernel estimate. It is an example for nonparametric regression estimation. The k-nearest neighbors algorithm is a very similar technique, which is also presented in the end of this section.

Let function $K : \mathbb{X} \rightarrow \mathbb{R}^+$ be the so-called user-chosen kernel function. For example in practice the naive kernel $K(x) = \mathbb{I}(\|x\| \leq 1)$, the Epanechnikov kernel $K(x) = (1 - x^2)_+$, where $f_+(x) \doteq \max(f(x), 0)$ and the Gaussian kernel $K(x) = \exp(-x^2/2)$ are often used. Let h_n be the *bandwidth* which is positive for all $n \in \mathbb{N}$. We define the kernel estimate for all $n \in \mathbb{N}$ as

$$f_n(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x - X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right)} \mathbb{I}\left(\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) \neq 0\right). \quad (3.8)$$

3.4 Universal Consistency

We prove the universal consistency of such estimates. For a start we prove a simple lemma which is going to be applied several times, see [11, Lemma 4.1].

Lemma 3.4.1. *Let $B(n, p)$ be a binomial random variable with parameters n and p . Then*

$$\mathbb{E}\left(\frac{1}{1 + B(n, p)}\right) \leq \frac{1}{(n + 1)p} \quad \text{and} \quad (3.9)$$

$$\mathbb{E}\left(\frac{1}{B(n, p)} \mathbb{I}(B(n, p) > 0)\right) \leq \frac{2}{(n + 1)p}. \quad (3.10)$$

Proof. The following calculation yields (3.9).

$$\begin{aligned} \mathbb{E}\left(\frac{1}{1 + B(n, p)}\right) &\leq \sum_{i=1}^n \frac{1}{i + 1} \binom{n}{i} p^i (1 - p)^{n-i} = \frac{1}{(n + 1)p} \sum_{i=1}^n \binom{n + 1}{i + 1} p^{i+1} (1 - p)^{n-i} \\ &\leq \frac{1}{(n + 1)p} \sum_{k=0}^{n+1} \binom{n + 1}{k} p^k (1 - p)^{n-k+1} = \frac{1}{(n + 1)p} (p + (1 - p))^{n+1} = \frac{1}{(n + 1)p} \end{aligned} \quad (3.11)$$

Using this result, we obtain (3.10) by

$$\mathbb{E}\left(\frac{1}{B(n, p)} \mathbb{I}(B(n, p) > 0)\right) \leq \mathbb{E}\left(\frac{2}{1 + B(n, p)}\right) \leq \frac{2}{(n + 1)p}. \quad (3.12)$$

□

Theorem 3.4.2. *Assume that there are numbers R and r with relation $0 < r \leq R$ and $b > 0$ such that*

$$\mathbb{I}(x \in B(0, R)) \geq K(x) \geq b \mathbb{I}(x \in B(0, r)) \quad (3.13)$$

holds. Let f_n be as in [3.8]. If $h_n \rightarrow 0$ and $nh_n^d \rightarrow \infty$, then f_n is universally consistent.

Proof. The proof is from [11, Theorem 5.1]. It is sufficient to check the conditions of Stone's theorem. Let $K_h(x) \doteq K(x/h)$. We see that

$$w_{n,i}(x) = \frac{K_{h_n}(x - X_i)}{\sum_{i=1}^n K_{h_n}(x - X_i)} \mathbb{I}\left(\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) \neq 0\right). \quad (3.14)$$

For condition i let $h = h_n$. Then

$$\begin{aligned} & \mathbb{E}\left(\frac{\sum_{i=1}^n K_h(X - X_i)f(X_i)}{\sum_{i=1}^n K_h(X - X_i)} \mathbb{I}\left(\sum_{i=1}^n K_h(X - X_i) \neq 0\right)\right) \\ &= n\mathbb{E}\left(\frac{K_h(X - X_1)f(X_1)}{\sum_{i=1}^n K_h(X - X_i)} \mathbb{I}\left(\sum_{i=1}^n K_h(X - X_i) \neq 0\right)\right) \\ &= n\mathbb{E}\left(\frac{K_h(X - X_1)f(X_1)}{K_h(X - X_1) + \sum_{i=2}^n K_h(X - X_i)} \mathbb{I}\left(\sum_{i=1}^n K_h(X - X_i) \neq 0\right)\right) \\ &= n \int_{\mathbb{X}} f(u) \mathbb{E}\left[\frac{K_h(x - u) \mathbb{I}\left(\sum_{i=1}^n K_h(X - X_i) \neq 0\right)}{K_h(x - u) + \sum_{i=2}^n K_h(x - X_i)} dP_X(x)\right] P_X(u). \end{aligned} \quad (3.15)$$

We are going to show that

$$\mathbb{E}\left[\int_{\mathbb{X}} \frac{K_h(x - u) \mathbb{I}\left(\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right) \neq 0\right)}{K_h(x - u) + \sum_{i=2}^n K_h(x - X_i)} dP_X(x)\right] \leq \frac{c}{n}. \quad (3.16)$$

Notice that K has a compact support, which can be covered by finitely many balls with radius $r/2$. Let M denote the number of these balls and x_k , $k \in [M]$ denote the centers of these covering balls. Then for all $\frac{x-u}{h} \in \mathbb{X}$ there exists x_k such that $\frac{x-u}{h} \in B(x_k, r/2)$, which is equivalent to $x \in B(u + hx_k, rh/2)$. By the covering property for all x and u

$$K_h(x - u) \leq \sum_{k=1}^M \mathbb{I}(x \in B(u + hx_k, rh/2)). \quad (3.17)$$

Furthermore, if $x \in B(u + hx_k, rh/2)$ then $B(u + hx_k, rh/2) \subseteq B(x, rh)$ equivalently $B(\frac{u-x}{h} + x_k, r/2) \subseteq B(0, r)$. Applying these two observations, the assumptions on K and Lemma 3.4.1 yields the following

$$\begin{aligned} & \mathbb{E}\left[\int_{\mathbb{X}} \frac{K_h(x - u) \mathbb{I}\left(\sum_{i=1}^n K_h(x - X_i) \neq 0\right)}{K_h(x - u) + \sum_{i=2}^n K_h(x - X_i)} dP_X(x)\right] \\ &\leq \sum_{k=1}^M \mathbb{E}\left[\int_{B(u + hx_k, rh/2)} \frac{K_h(x - u) \mathbb{I}\left(\sum_{i=1}^n K_h(x - X_i) \neq 0\right)}{K_h(x - u) + \sum_{i=2}^n K_h(x - X_i)} dP_X(x)\right] \\ &\leq \sum_{k=1}^M \mathbb{E}\left[\int_{B(u + hx_k, rh/2)} \frac{1}{b + \sum_{i=2}^n K_h(x - X_i)} dP_X(x)\right] \\ &\leq \frac{1}{b} \sum_{k=1}^M \mathbb{E}\left[\int_{B(u + hx_k, rh/2)} \frac{1}{1 + \sum_{i=2}^n \mathbb{I}(X_i \in B(x, rh))} dP_X(x)\right] \end{aligned}$$

$$\begin{aligned}
 &\leq \frac{1}{b} \sum_{k=1}^M \mathbb{E} \left[\int_{B(u+hx_k, rh/2)} \frac{1}{1 + \sum_{i=2}^n \mathbb{I}(X_i \in B(u+hx_k, rh/2))} dP_X(x) \right] \\
 &= \frac{1}{b} \sum_{k=1}^M \mathbb{E} \left[\frac{P_X(B(u+hx_k, rh/2))}{1 + \sum_{i=2}^n \mathbb{I}(X_i \in B(u+hx_k, rh/2))} \right] \\
 &\leq \frac{1}{b} \sum_{k=1}^M \mathbb{E} \left[\frac{P_X(B(u+hx_k, rh/2))}{nP_X(B(u+hx_k, rh/2))} \right] \leq \frac{M}{nb}.
 \end{aligned} \tag{3.18}$$

For condition *ii* it is easy to see that

$$\left| \frac{K_h(X - X_i)}{\sum_{i=1}^n K_h(X - X_i)} \right| \mathbb{I} \left(\sum_{i=1}^n K_h(X - X_i) \neq 0 \right) \leq 1 \tag{3.19}$$

with probability one. For condition *iii* let $h_n R < a$, then

$$\begin{aligned}
 &\sum_{i=1}^n |w_{n,i}(X)| \mathbb{I}(\|X_i - X\| > a) \\
 &= \frac{\sum_{i=1}^n K_{h_n}(X - X_i) \mathbb{I}(\|X_i - X\| > a)}{\sum_{i=1}^n K_{h_n}(X - X_i)} \mathbb{I} \left(\sum_{i=1}^n K_{h_n}(X - X_i) \neq 0 \right) = 0.
 \end{aligned} \tag{3.20}$$

For condition *iv* we need to prove that $\sum_{i=1}^n w_{n,i}(X) \rightarrow 1$ in probability. Notice that

$$1 - w_{n,i}(X) = \mathbb{I} \left(\sum_{i=1}^n K_{h_n}(X - X_i) = 0 \right). \tag{3.21}$$

Then, we proceed as

$$\begin{aligned}
 &\mathbb{P} \left(1 \neq \sum_{i=1}^n w_{n,i}(X) \right) = \mathbb{P} \left(\sum_{i=1}^n K_{h_n}(X - X_i) = 0 \right) \\
 &\leq \mathbb{P} \left(\sum_{i=1}^n \mathbb{I}(X_i \notin B(X, rh_n)) = 0 \right) = \mathbb{P} \left(P_X^n(B(X, rh_n)) = 0 \right) \\
 &= \int (1 - P_X(B(X, rh_n)))^n dP_X(x),
 \end{aligned} \tag{3.22}$$

where P_X^n denotes the empirical version of P_X . For an arbitrary $\delta > 0$ let $B = B(0, L)$ such that $P_X(\bar{B}) < \delta$, where $\bar{B} = \mathbb{X} - B$. Then

$$\begin{aligned}
 &\mathbb{P} \left(1 \neq \sum_{i=1}^n w_{n,i}(X) \right) \leq \int_B e^{-nP_X(B(X, rh_n))} dP_X(x) + P_X(\bar{B}) \\
 &= \int_B nP_X(B(X, rh_n)) e^{-nP_X(B(X, rh_n))} \frac{1}{nP_X(B(X, rh_n))} dP_X(x) + P_X(\bar{B}) \\
 &\leq \max_u u e^{-u} \int_B \frac{1}{nP_X(B(X, rh_n))} dP_X(x) + \delta.
 \end{aligned} \tag{3.23}$$

Since $B = B(0, L)$ there are z_1, \dots, z_{M_n} such that $B \subseteq \cup_{i=1}^{M_n} B(z_i, rh_n/2)$ and $M_n \leq \frac{\tilde{c}(L)}{h_n^d}$ from

which it follows that

$$\begin{aligned} \int_B \frac{1}{nP_X(B(X, rh_n))} dP_X(x) &\leq \sum_{j=1}^{M_n} \int \frac{\mathbb{I}(X \in B(z_j, rh_n/2))}{nP_X(B_X, rh_n)} dP_X(x) \\ &\leq \sum_{j=1}^{M_n} \int \frac{\mathbb{I}(X \in B(z_j, rh_n/2))}{nP_X(B_{z_j}, rh_n/2)} dP_X(x) \leq \frac{M_n}{n} \leq \frac{\tilde{c}(L)}{nh_n^d} \rightarrow 0. \end{aligned} \quad (3.24)$$

Concerning condition v we use that $K(x) \leq 1$. Then, for all $\varepsilon > 0$ we have

$$\begin{aligned} \sum_{i=1}^n w_{n,i}^2(X) &= \frac{\sum_{i=1}^n K_{h_n}^2(X - X_i)}{(\sum_{i=1}^n K_{h_n}(X - X_i))^2} \mathbb{I}\left(\sum_{i=1}^n K_{h_n}(X - X_i) \neq 0\right) \\ &\leq \frac{\sum_{i=1}^n K_{h_n}(X - X_i)}{(\sum_{i=1}^n K_{h_n}(X - X_i))^2} \mathbb{I}\left(\sum_{i=1}^n K_{h_n}(X - X_i) \neq 0\right) \\ &\leq \min\left(\varepsilon, \frac{\mathbb{I}(\sum_{i=1}^n K_{h_n}(X - X_i) \neq 0)}{\sum_{i=1}^n K_{h_n}(X - X_i)}\right) \leq \min\left(\varepsilon, \frac{1}{\sum_{i=1}^n b \mathbb{I}(X_i \in B(X, rh_n))}\right) \\ &\leq \varepsilon + \frac{1}{\sum_{i=1}^n b \mathbb{I}(X_i \in B(X, rh_n))} \mathbb{I}\left(\sum_{i=1}^n \mathbb{I}(X_i \in B(X, rh_n)) > 0\right). \end{aligned} \quad (3.25)$$

It is sufficient to prove that

$$\mathbb{E}\left(\frac{1}{\sum_{i=1}^n \mathbb{I}(X_i \in B(X, rh_n))} \mathbb{I}\left(\sum_{i=1}^n \mathbb{I}(X_i \in B(X, rh_n)) > 0\right)\right) \rightarrow 0. \quad (3.26)$$

Let B be similarly defined as before (3.23). Applying Lemma 3.4.1 yields

$$\begin{aligned} &\mathbb{E}\left(\frac{1}{\sum_{i=1}^n \mathbb{I}(X_i \in B(X, rh_n))} \mathbb{I}\left(\sum_{i=1}^n \mathbb{I}(X_i \in B(X, rh_n)) > 0\right)\right) \\ &\leq \mathbb{E}\left(\frac{1}{\sum_{i=1}^n \mathbb{I}(X_i \in B(X, rh_n))} \mathbb{I}\left(\sum_{i=1}^n \mathbb{I}(X_i \in B(X, rh_n)) > 0\right) \mathbb{I}(X \in B)\right) + P_X(\bar{B}) \\ &\leq 2\mathbb{E}\left(\frac{1}{(n+1)P_X(B(X, rh_n))} \mathbb{I}(X \in B)\right) + \delta, \end{aligned} \quad (3.27)$$

where $\mathbb{E}\left(\frac{1}{(n+1)P_X(B(X, rh_n))} \mathbb{I}(X \in B)\right)$ goes to zero similarly as in (3.24).

Since all five conditions hold Theorem 3.2.1 can be used. This proves the universal consistency of our kernel estimate. \square

3.5 Strong Consistency

We turn our attention to prove strong consistency for a broad class of kernel estimates in case of bounded variable Y .

In order to prove strong consistency a quite involved argument is needed, see [11, Theorem 23.5]. The Banach–Steinhaus theorem for integral operators is going to be one of our tool to proceed in the proof. The proof from [11, Theorem 23.2] can be found in the appendix, see B.4

Theorem 3.5.1. *Let K_n be $\mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ type functions for $n \in \mathbb{N}$ and μ be a probability measure on \mathbb{R}^d . Assume the followings:*

i There exists $c > 0$ such that for all $n \in \mathbb{N}$ the following holds

$$\int |K_n(x, z)| d\mu(x) \leq c \quad (3.28)$$

for μ -almost every z .

ii There exists $D \geq 1$ such that for all $x \in \mathbb{R}^d$ and for all $n \in \mathbb{N}$

$$\int |K_n(x, z)| d\mu(z) \leq D. \quad (3.29)$$

iii For all $a > 0$

$$\lim_{n \rightarrow \infty} \int \int |K_n(x, z)| \mathbb{I}(\|x - z\| > a) d\mu(z) d\mu(x) = 0. \quad (3.30)$$

iv

$$\lim_{n \rightarrow \infty} \text{ess sup}_x \left| \int K_n(x, z) d\mu(z) - 1 \right| = 0. \quad (3.31)$$

Then for all $f \in L_1(\mu)$

$$\lim_{n \rightarrow \infty} \int \left| f(x) - \int K_n(x, z) f(z) d\mu(z) \right| d\mu(x) = 0. \quad (3.32)$$

Strong consistency will be proved for a broad class of kernel functions.

Definition 3.5.1. (regular kernel) *The kernel function K is regular, if $K(x) \geq 0$ and there exists $r > 0$ and $b > 0$ such that*

$$b\mathbb{I}(x \in B(0, r)) \leq K(x) \leq 1 \quad \text{and} \\ \int \sup_{u \in B(x, r)} K(u) dx < \infty.$$

Theorem 3.5.2. *Let f_n be a kernel estimate of f_* with a regular kernel function K . Assume that there exists $L < \infty$ such that $\mathbb{P}(|Y| \leq L) = 1$. If $h_n \rightarrow 0$ and $nh_n^d \rightarrow \infty$ then the kernel estimate is strongly consistent.*

The proof will be presented in many steps as in [11, Theorem 23.5]. Before proving this general theorem four lemmas are presented. Let $K_h(x) = K(x/h)$ as before.

Lemma 3.5.3. (covering lemma) *Let K be a regular kernel. Then there exists a finite constant $\varrho = \varrho(K)$ such that for all $u \in \mathbb{R}^d$, $h > 0$ and probability measure μ*

$$\int \frac{K_h(x - u)}{\int K_h(x - z) d\mu(z)} d\mu(x) \leq \varrho. \quad (3.33)$$

In addition, for all $\delta > 0$

$$\lim_{n \rightarrow \infty} \sup_u \int \frac{K_h(x-u) \mathbb{I}(\|x-u\| > \delta)}{\int K_h(x-z) d\mu(z)} d\mu(x) = 0. \quad (3.34)$$

The proof can be found in the appendix, see [B.5](#).

Lemma 3.5.4. *Let h and R be real numbers such that $0 < h \leq R < \infty$ and $B = B(0, R)$. Then for all probability measures μ we have that*

$$\int_B \frac{1}{\sqrt{\mu(B(x, h))}} d\mu(x) \leq \left(1 + \frac{R}{h}\right)^{d/2} c(d) \quad (3.35)$$

Proof. The proof is similar to the argument in [\(3.24\)](#). Let $B = B(0, R)$ such that $P_x(\bar{B}) < \delta$. Take the balls $B(z_k, h/2)$, where $z_k = (-R + k_1 h/2, \dots, -R + k_d h/2)$ for $k = (k_1, \dots, k_d) \in \left[\left\lceil \frac{2R}{h/2} \right\rceil\right]^d$. The union of these balls covers B . Let the number of these balls be M . Notice that

$$M = \left\lceil \frac{2R}{h/2} \right\rceil^d \leq \left(\frac{2R}{h/2} + 1\right)^d \leq \left(1 + \frac{R}{h}\right)^d c(d). \quad (3.36)$$

Applying Jensen's inequality for integrals and the covering property yields

$$\begin{aligned} \left(\int_B \frac{1}{\sqrt{\mu(B(x, h))}} d\mu(x)\right)^2 &\leq \int_B \frac{1}{\mu(B(x, h))} d\mu(x) \\ &\leq \sum_{i=1}^M \int \frac{\mathbb{I}(x \in B(z_j, h/2))}{\mu(B(x, h))} d\mu(x) \leq \sum_{i=1}^M \int \frac{\mathbb{I}(x \in B(z_j, h/2))}{\mu(B(z_j, h/2))} d\mu(x) \leq M, \end{aligned} \quad (3.37)$$

which together with [\(3.36\)](#) proves the lemma. \square

In the proof of Theorem [3.5.2](#) the following auxiliary function

$$f_n^*(x) \doteq \frac{\sum_{i=1}^n Y_i K_{h_n}(x - X_i)}{n \mathbb{E} K_{h_n}(x - X)} \quad (3.38)$$

plays an important role. Lemma [3.5.5](#) is an important observation about this function.

Lemma 3.5.5. *Under the conditions of Theorem [3.5.2](#) the following holds*

$$\lim_{n \rightarrow \infty} \int \mathbb{E} |f_*(x) - f_n^*(x)| dP_X(x) = 0. \quad (3.39)$$

Proof. By the triangle inequality

$$\begin{aligned} \int \mathbb{E} |f_*(x) - f_n^*(x)| dP_X(x) &\leq \int |f_*(x) - \mathbb{E} f_n^*(x)| dP_X(x) \\ &\quad + \int \mathbb{E} |f_n^*(x) - \mathbb{E} f_n^*(x)| dP_X(x) = I_1 + I_2. \end{aligned} \quad (3.40)$$

We show that Theorem 3.5.1 can be used for I_1 . Let

$$K_n(x, z) \doteq \frac{K_{h_n}(x - z)}{\int K_{h_n}(x - u) dP_X(u)}. \quad (3.41)$$

For condition *i* by Lemma 3.5.3 with $c = \varrho$ for all $z \in \mathbb{R}^d$ and $n \in \mathbb{N}$

$$\int |K_n(x, z)| dP_X(x) \leq c. \quad (3.42)$$

Condition *ii* and *iv* trivially hold, since for all $n \in \mathbb{N}$ we have

$$\int K_n(x, z) dP_X(z) = 1. \quad (3.43)$$

For condition *iii* we can apply Fubini's theorem and the second part of Lemma 3.5.3, that is

$$\begin{aligned} & \int \int K_n(x, z) \mathbb{I}(\|x - z\| > a) dP_X(z) dP_X(x) \\ &= \int \int \frac{K\left(\frac{x-z}{h_n}\right) \mathbb{I}(\|x - z\| > a)}{\int K\left(\frac{x-u}{h_n}\right) dP_X(u)} dP_X(x) dP_X(z) \\ &\leq \int \sup_z \left(\int \frac{K\left(\frac{x-z}{h_n}\right) \mathbb{I}(\|x - z\| > a)}{\int K\left(\frac{x-u}{h_n}\right) dP_X(u)} dP_X(x) \right) dP_X(z) \\ &= \sup_z \left(\int \frac{K\left(\frac{x-z}{h_n}\right) \mathbb{I}(\|x - z\| > a)}{\int K\left(\frac{x-u}{h_n}\right) dP_X(u)} dP_X(x) \right) \rightarrow 0 \end{aligned} \quad (3.44)$$

as $n \rightarrow \infty$. Therefore by Theorem 3.5.1 we proceed as

$$\begin{aligned} & \int \left| f_*(x) - \mathbb{E} \left(\frac{Y_i K_{h_n}(x - X_i)}{n \mathbb{E} K_{h_n}(x - X)} \right) \right| dP_X(x) \\ &= \int \left| f_*(x) - \sum_{i=1}^n \mathbb{E} \left(\frac{\mathbb{E}(Y_i | X_i) K_n(x, X_i)}{n} \right) \right| dP_X(x) \\ &= \int \left| f_*(x) - \mathbb{E}(\mathbb{E}(Y | X) K_n(x, X)) \right| dP_X(x) \\ &= \int \left| f_*(x) - \int f_*(z) K_n(x, z) dP_X(z) \right| dP_X(x) \rightarrow 0. \end{aligned} \quad (3.45)$$

For I_2 let $h = h_n$. Apply the Cauchy–Schwarz inequality and that $\mathbb{P}(|Y| \leq L) = 1$ to obtain

$$\begin{aligned} \mathbb{E} |f_n^*(x) - \mathbb{E} f_n^*(x)| &\leq \sqrt{\mathbb{E} |f_n^*(x) - \mathbb{E} f_n^*(x)|^2} \\ &= \sqrt{\frac{\mathbb{E} \left[\left(\sum_{i=1}^n Y_i K_h(x - X_i) - \mathbb{E}(Y K_h(x - X)) \right)^2 \right]}{n^2 (\mathbb{E} K_h(x - X))^2}} \\ &\leq \sqrt{\frac{\mathbb{E} \left[\left(Y K_h(x - X) - \mathbb{E}(Y K_h(x - X)) \right)^2 \right]}{n (\mathbb{E} K_h(x - X))^2}} \leq \sqrt{\frac{\mathbb{E} \left[\left(Y K_h(x - X) \right)^2 \right]}{n (\mathbb{E} K_h(x - X))^2}} \end{aligned}$$

$$\begin{aligned}
 &\leq L \sqrt{\frac{\mathbb{E} \left[K^2 \left(\frac{x-X}{h} \right) \right]}{n(\mathbb{E} K_h(x-X))^2}} \leq L \sqrt{\frac{\mathbb{E} [K_h(x-X)] \sup_{x \in \mathbb{R}^d} K(x)}{n(\mathbb{E} K_h(x-X))^2}} \\
 &\leq L \frac{1}{\sqrt{b}} \frac{1}{\sqrt{n P_X(B(x, h))}}.
 \end{aligned} \tag{3.46}$$

For $\varepsilon > 0$ let $B \doteq B(0, R)$ such that $P_X(\bar{B}) < \varepsilon/(2L)$. Then

$$\int_{\bar{B}} \mathbb{E} |f_n^*(x) - \mathbb{E} f_n^*(x)| \, dP_X(x) \leq 2 \int_{\bar{B}} \mathbb{E} |f_n^*(x)| \, dP_X(x) \leq 2L P_X(\bar{B}) < \varepsilon. \tag{3.47}$$

Furthermore, applying Lemma 3.5.4 yields

$$\begin{aligned}
 \int_B \mathbb{E} |f_n^*(x) - \mathbb{E} f_n^*(x)| \, dP_X(x) &\leq L \sqrt{\frac{1}{bn}} \int_B \frac{1}{\sqrt{P_X(B(x, h))}} \, dP_X(x) \\
 &\leq L \sqrt{\frac{1}{bn}} \left(1 + \frac{R}{h}\right)^{d/2} c(d) \rightarrow 0
 \end{aligned} \tag{3.48}$$

as $n \rightarrow \infty$, since $nh_n^d \rightarrow \infty$. Putting these together proves the lemma. \square

Lemma 3.5.6. *For all $\varepsilon > 0$ there exists $n_0 \in \mathbb{N}$ such that for all $n > n_0$ we have*

$$\mathbb{P} \left(\int |f_*(x) - f_n^*(x)| \, dP_X(x) \leq \varepsilon \right) \leq \exp \left(- \frac{n\varepsilon^2}{8L^2\varrho^2} \right) \tag{3.49}$$

Proof. We add and subtract the mean of the examined variable

$$\begin{aligned}
 \int |f_*(x) - f_n^*(x)| \, dP_X(x) &= \int \mathbb{E} |f_*(x) - f_n^*(x)| \, dP_X(x) \\
 &+ \int (|f_*(x) - f_n^*(x)| - \mathbb{E} |f_*(x) - f_n^*(x)|) \, dP_X(x) = I_n + J_n.
 \end{aligned} \tag{3.50}$$

By Lemma 3.5.5 $I_n \rightarrow 0$ so there exists n_0 such that for all $n > n_0$: $|I_n| < \varepsilon/2$. For J_n we are going to apply McDiarmid's inequality (Theorem A.4.2) for function

$$\kappa(\{(X_i, Y_i)\}_{i=1}^n) = \int |f_*(x) - f_n^*(x)| \, dP_X(x). \tag{3.51}$$

Let $f_n^*(x)$ be the estimate function defined by a fixed $(x_1, y_1), \dots, (x_n, y_n)$ sample and let $f_{ni}(x)$ be defined by a sample which is distinct from the one above only in the i^{th} coordinate. Then by Lemma 3.5.4

$$\begin{aligned}
 &\left| \int |f_*(x) - f_n^*(x)| \, dP_X(x) - \int |f_*(x) - f_{ni}^*(x)| \, dP_X(x) \right| \\
 &\leq \int |f_n^*(x) - f_{ni}^*(x)| \, dP_X(x) \leq \sup_{y \in \mathbb{R}^d} \int \frac{2LK_h(x-y)}{n\mathbb{E}K_h(x-X)} \, dP_X(x) \leq \frac{2L\varrho}{n}
 \end{aligned} \tag{3.52}$$

By McDiarmid's inequality we obtain that for $n > n_0$:

$$\mathbb{P} \left(\int |f_*(x) - f_n^*(x)| \, dP_X(x) > \varepsilon \right)$$

$$\begin{aligned}
 &\leq \mathbb{P} \left(\left(\int |f_*(x) - f_n^*(x)| \, dP_X(x) - \mathbb{E} \int |f_*(x) - f_n^*(x)| \, dP_X(x) \right) > \varepsilon/2 \right) \\
 &\leq \exp \left(- \frac{n\varepsilon^2}{8L^2\varrho^2} \right).
 \end{aligned} \tag{3.53}$$

Thus, the lemma is proved. \square

Proof of Theorem 3.5.2. It is easy to see that f_n and f_* are bounded by L , because $|Y| \leq L$. Notice that

$$\int |f_n(x) - f_*(x)|^2 \, dP_X(x) \leq 2L \int |f_n(x) - f_*(x)| \, dP_X(x), \tag{3.54}$$

therefore it is sufficient to prove $\int |f_n(x) - f_*(x)| \, dP_X(x) \xrightarrow{a.s.} 0$. By the triangle inequality

$$\begin{aligned}
 &\int |f_n(x) - f_n^*(x)| \, dP_X(x) \\
 &\leq \int |f_n(x) - f_n^*(x)| \, dP_X(x) + \int |f_n^*(x) - f_*(x)| \, dP_X(x)
 \end{aligned} \tag{3.55}$$

holds. The quantity $\int |f_n^*(x) - f_*(x)| \, dP_X(x)$ goes to zero almost surely by Lemma 3.5.6 and the Borel-Cantelli lemma. Moreover

$$\begin{aligned}
 |f_n^*(x) - f_n(x)| &= \left| \frac{\sum_{i=1}^n Y_i K_{h_n}(x - X_i)}{n\mathbb{E}K_{h_n}(x - X)} - \frac{\sum_{i=1}^n Y_i K_{h_n}(x - X_i)}{\sum_{i=1}^n K_{h_n}(x - X_i)} \right| \\
 &\leq \left| \sum_{i=1}^n Y_i K_{h_n}(x - X_i) \right| \left| \frac{1}{n\mathbb{E}K_{h_n}(x - X)} - \frac{1}{\sum_{i=1}^n K_{h_n}(x - X_i)} \right| \\
 &\leq L \left| \sum_{i=1}^n Y_i K_{h_n}(x - X_i) \right| \left| \frac{1}{n\mathbb{E}K_{h_n}(x - X)} - \frac{1}{\sum_{i=1}^n K_{h_n}(x - X_i)} \right| \\
 &= L |F_n^*(x) - 1|,
 \end{aligned} \tag{3.56}$$

where $F_n^*(x) = f_n^*(x)$ when $Y = 1$ with probability 1. Then by Lemma 3.5.5

$$\int |f_n^*(x) - f_n(x)| \, dP_X(x) \leq L \int |F_n^*(x) - 1| \, dP_X(x) \rightarrow 0 \tag{3.57}$$

almost surely, which finishes the proof. \square

3.6 The k-Nearest Neighbors Estimate

Another simple and efficient technique for nonparametric regression is the k-nearest neighbors estimate (kNN). It can be viewed as a kernel estimate using the naive kernel with data-dependent adaptive bandwidth. Consider the space $\mathbb{X} \subseteq \mathbb{R}^d$ with the euclidean metric. Fix $x \in \mathbb{X}$, then we can define an order of the variables $\|x - X_i\|$ for all $i = 1, \dots, n$. We extend these random variables with the different elements of a random permutation $\pi[n] \rightarrow [n]$ to decide in case of ties, so we define the total order \prec_π as $\|x - X_i\| \prec_\pi \|x - X_j\|$ if and only if $\|x - X_i\| < \|x - X_j\|$ or $\|x - X_i\| = \|x - X_j\|$ and $\pi(i) < \pi(j)$. Then for all $i \in [n]$ we consider variables $\mathbb{I}(X_i \in N(x, k_n))$, where $N(x, k_n)$ is the first k_n elements in the defined total order,

that is the k_n closest points in $\{X_i\}_{i=1}^n$ to x . With the help of these variables we can define the kNN estimate for all $n \in \mathbb{N}$ as

$$f_n^{kNN}(x) \doteq \frac{1}{k_n} \sum_{i=1}^n \mathbb{I}(X_i \in N(x, k_n)). \quad (3.58)$$

Stone's theorem can be applied to show that if $k_n \rightarrow \infty$ and $k_n/n \rightarrow 0$ as $n \rightarrow \infty$ then the kNN estimate is universally consistent. Besides, for the strong consistency of kNN estimates, similarly as for kernel estimates (Theorem [3.5.2](#)), the following holds:

Theorem 3.6.1. *Assume that there exists $L < \infty$ such that $\mathbb{P}(|Y| \leq L) = 1$ and that for each random variable $\|x - X\|$ is absolutely continuous. If $k_n \rightarrow \infty$ and $k_n/n \rightarrow 0$, then the kNN regression function estimate is strongly consistent.*

The proof is similarly involved as the argument for Theorem [3.5.2](#), see [\[11\]](#), Theorem 23.7].

Chapter 4

Kernel Methods

In this chapter we present the theory of kernel methods. It is important to distinguish these methods from kernel estimates which were introduced in Chapter 3, therefore we are not going to use the word kernel estimate here. Kernel methods are generalizations of algorithms which only depend on the data through an inner product.

Before we formally dive into the details of the theory of kernel methods, we show two very important examples to illustrate that several algorithms can be expressed via an inner product.

4.1 Ridge Regression

Least squares estimates are probably the most widely applied tools in statistics. Ridge regression (RR) is a generalization of the least squares method. We consider a regression problem with data $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ for $i \in [n]$. Let $X \doteq [x_1, \dots, x_n]^T$ and $Y = [y_1, \dots, y_n]^T$. We want to find a model from a linear space parameterized by $\theta \in \mathbb{R}^d$ which minimizes the following cost function

$$\text{minimize} \quad \frac{1}{2} \|Y - X\theta\|^2 + \frac{\lambda}{2} \|\theta\|^2$$

for a user-chosen $\lambda > 0$. The minimizer of this cost function is called the *ridge regression* estimate or regularized least squares estimate, because the term $\|\theta\|^2$ is included in the cost for regularization, which is an extremely important concept in machine learning. Without delving into details, it is easy to see that in this case regularization ensures the existence and the uniqueness of the minimizer, thus it can help to transform an ill-posed problem to a well-posed one. Besides, it often happens that matrix $X^T X$, which plays an important role in the analytic least squares solution, has a very large condition number implying that it is numerically problematic to compute $(X^T X)^{-1}$. In such cases the regularizer helps us to reduce the condition number, so we can transform an ill-conditioned problem to a well-conditioned one. For further readings on regularization see the books [14] and [18].

Notice that we can reduce the regularized problem to a least squares problem by substituting in $\tilde{X} = [X^T \sqrt{\lambda}]^T$ and $\tilde{Y} = [Y^T 0]^T$. Let I denote the identity matrix, then applying the analytic

least squares formula we can express the ridge regression estimate as

$$\hat{\theta}_\lambda = (X^T X + \lambda I)^{-1} X^T Y, \quad (4.1)$$

which is well-defined, because matrix $X^T X + \lambda I$ is always positive definite. Notice that

$$(X^T X + \lambda I) \hat{\theta}_\lambda = X^T Y \Rightarrow \hat{\theta}_\lambda = \lambda^{-1} (X^T Y - X^T X \hat{\theta}_\lambda). \quad (4.2)$$

We introduce variable α as

$$\alpha \doteq \lambda^{-1} (Y - X \hat{\theta}_\lambda) \quad (4.3)$$

then we obtain that $\hat{\theta}_\lambda = X^T \alpha$. Substituting it back to the definition of α yields the following linear equation system

$$(X X^T + \lambda I) \alpha = Y. \quad (4.4)$$

Notice that our model estimate $\hat{f} : \mathbb{R}^d \rightarrow \mathbb{R}$ on a given input can be evaluated as

$$\hat{f}(x) = x^T \hat{\theta}_\lambda = x^T X^T \alpha = \sum_{k=1}^n \alpha_k \langle x, x_k \rangle, \quad (4.5)$$

where from (4.4) we can see that α and the prediction depends on the inputs only through the inner product.

4.2 Support Vector Machines

Our second example is the so-called Support Vector Machines (SVM) for classification, see [18], [21] and [23].

Given a sample $\{(x_i, y_i)\}_{i=1}^n$ for binary classification. Consider the model class of linear classifiers i.e. $\mathcal{H} = \{g : g(x) = \text{sign}(w^T x + b), w \in \mathbb{R}^d, b \in \mathbb{R}\}$ parameterized with normal vector w and a bias term b . We assume that our data is linearly separable so there exists $g \in \mathcal{H}$ with zero empirical risk. We say that (w, b) corresponds to a δ -margin separating hyperplane if

$$y_i(w^T x_i + b) \geq \delta \quad \forall i \in [n], \|w\| = 1 \quad (4.6)$$

holds. We call δ the *margin*, that is the minimal distance of the input points from the hyperplane. In many books 2δ is called the margin, because if we take the convex hull of those x_i 's which have label $+1$ and the convex hull of the rest of the inputs the distance of these two hulls is 2δ .

Claim 4.2.1. *The VC dimension of \mathcal{H} in \mathbb{R}^d is $d + 1$.*

Proof. The proof is based on [19]. First we show that the VC dimension of the homogenous hyperplanes in \mathbb{R}^d is d . It is easy to see that d points can be shattered. Let e_1, \dots, e_d be the usual basis in \mathbb{R}^d , where all coordinates of e_i are zero except the i^{th} , which is 1. In fact, for all labelings y_1, \dots, y_d let the hyperplane be $w = (y_1, \dots, y_d)^T$. Then $\langle w, e_k \rangle = y_k$. Assume by contradiction that we can shatter $d + 1$ points. Let z_1, \dots, z_{d+1} be such points in \mathbb{R}^d . Then there exists $a_1, \dots, a_{d+1} \in \mathbb{R}$ not all zeros, such that $\sum_{i=1}^{d+1} a_i x_i = 0$. Let $I = \{i : a_i > 0\}$ and

$J = \{i : a_i \leq 0\}$. First, assume that both I and J are nonempty. Then

$$\sum_{i \in I} a_i x_i = \sum_{j \in J} |a_j| x_j. \quad (4.7)$$

Since we assumed that x_1, \dots, x_{d+1} can be shattered, there exists w for which $\langle w, x_i \rangle \geq 0$ for all $i \in I$ and $\langle w, x_j \rangle < 0$ for all $j \in J$. Then the contradiction follows since

$$0 \leq \sum_{i \in I} a_i \langle x_i, w \rangle = \left\langle \sum_{i \in I} a_i x_i, w \right\rangle = \left\langle \sum_{j \in J} |a_j| x_j, w \right\rangle = \sum_{j \in J} |a_j| \langle x_j, w \rangle < 0. \quad (4.8)$$

When J is empty then the contradiction follows from the fact that $\sum_{i \in I} a_i x_i = 0$, similarly when I is empty then $\sum_{j \in J} |a_j| x_j = 0$. For inhomogenous hyperplanes notice that the points $0, e_1, \dots, e_d$ can be shattered. For an arbitrary labeling y_0, \dots, y_d let $w = (y_1, \dots, y_d)^T$ as before and let the bias $b = y_0/2$. Then $\text{sign}(w^T e_k + b) = y_k$ and $\text{sign}(w^T 0 + b) = y_0$. Furthermore, notice that $\tilde{w} = (w^T, b)^T$ defines a homogeneous hyperplane in \mathbb{R}^{d+1} therefore by the first part the VC dimension of the inhomogeneous hyperplanes, \mathcal{H} , cannot be more than $d + 1$. \square

Theorem 4.2.1. *Let $x_i \in \mathbb{X} \subseteq \mathbb{R}^d$ belong to a ball with radius R for all $i \in [n]$. The set of δ -margin separating hyperplanes has VC dimension V bounded by*

$$V \leq \min \left\{ \left\lceil \frac{R^2}{\delta^2} \right\rceil, d \right\} + 1. \quad (4.9)$$

Proof. The proof is only partially presented here based on [3], [13] and [24].

From Claim 4.2.1 it follows that $V \leq d + 1$. Let $h \leq d$ and let x_1, \dots, x_h be h points that can be shattered by δ -margin separating hyperplanes. We can think of a labeling by a δ -margin hyperplane as a division of the points into two subsets such that the distance between the convex hulls of the two sets is greater than 2δ . Let these divisions be T_1, \dots, T_{2^h} and let $\varrho(T_i)$ denote the distance between the two convex hulls induced by the two subsets in T_i . Since x_1, \dots, x_h can be shattered $\min_{i \in [2^h]} \varrho(T_i) \geq 2\delta$. Let

$$H(h) \doteq \max_{x_1, \dots, x_h \in \mathbb{R}^d} \min_i \varrho(T_i). \quad (4.10)$$

The maximum is attained in a $h - 1$ dimensional regular simplex on the h dimensional sphere and can be calculated as

$$H(h) = \begin{cases} \frac{2R}{\sqrt{h-1}} & \text{if } h \text{ is even} \\ \frac{2Rh}{(h-1)\sqrt{h-1}} & \text{if } h \text{ is odd.} \end{cases} \quad (4.11)$$

For the detailed arguments see [3] and [13].

Then because of $H(h) \geq 2\delta$ if h is even it is easy to see that

$$h \leq \frac{R^2}{\delta^2} + 1 \leq \left\lceil \frac{R^2}{\delta^2} \right\rceil + 1. \quad (4.12)$$

Furthermore, if h is odd we can proceed as

$$\begin{aligned} \frac{R^2}{\delta^2} &= \frac{(h+1)(h-1)^2}{h^2} \geq (h-1)\frac{h-1}{h} \\ \left\lceil \frac{R^2}{\delta^2} \right\rceil + 1 &\geq \left\lceil \frac{(h-1)^2}{h} \right\rceil + 1 \geq \left\lceil h-2 + \frac{1}{h} \right\rceil + 1 \geq h. \end{aligned} \quad (4.13)$$

Therefore the bound $\left\lceil \frac{R^2}{\delta^2} \right\rceil + 1$ holds in both cases. \square

It is useful to parameterize the linear classifiers such way that $\min_{i=1,\dots,n} |w^T x_i + b| = 1$. It is called the *canonical form*. According to the SRM principle, by maximizing the margin we minimize the VC dimension of the appropriate model class, because of Theorem 4.2.1. Therefore the bound for the true risk of a model in Theorem 2.8.1 is also minimized. It can be shown that the margin maximization yields the following convex quadratic program

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2} \|w\|^2 \\ &\text{subject to} \quad y_i(w^T x_i + b) \geq 1, \quad i = 1, 2, \dots, n, \end{aligned} \quad (4.14)$$

which is called the primal problem and has a unique minimizer which can be found efficiently by a quadratic solver. We are going to show that the solution of this optimization problem has an inner product representation. Let

$$L(w, b, \alpha) \doteq \|w\|^2 + \sum_{k=1}^n (1 - y_k(w^T x_k + b)) \quad (4.15)$$

be the Lagrange dual function, where $\alpha_k \in \mathbb{R}^+$ are the Lagrange multipliers. By the Karush-Kuhn-Tucker (KKT) conditions, see [2], the optimum of (4.14) occurs if

$$\begin{aligned} \frac{\partial L}{\partial w}(w, b, \alpha) = 0 &\Rightarrow w = \sum_{k=1}^n \alpha_k y_k x_k \quad \text{and} \\ \frac{\partial L}{\partial b}(w, b, \alpha) = 0 &\Rightarrow 0 = \sum_{k=1}^n \alpha_k y_k. \end{aligned} \quad (4.16)$$

Substituting these identities back to the Lagrange dual function yields the following Wolfe-dual problem

$$\begin{aligned} &\text{maximize} \quad \sum_{k=1}^n \alpha_k - \frac{1}{2} \sum_{k=1}^n \sum_{l=1}^n \alpha_k \alpha_l y_k y_l \langle x_k, x_l \rangle \\ &\text{subject to} \quad \alpha_k \geq 0, \quad i = 1, 2, \dots, n, \end{aligned} \quad (4.17)$$

$$\sum_{k=1}^n \alpha_k y_k = 0. \quad (4.18)$$

It is a convex quadratic problem for the Lagrange multipliers. Our prediction on a given input takes the form of $g(x) = \text{sign}(\sum_{k=1}^n \alpha_k^* y_k \langle x, x_k \rangle + b^*)$. In order to calculate b^* we can use the KKT conditions (complementary slackness) from which for every $\alpha_k \neq 0$ we have $\langle w^*, x_k \rangle + b^* = y_k$.

The name of the method originated from these points, because we say that x_k is a support vector when $\alpha_k > 0$. It can be shown that α usually admits a sparse representation with at least $d + 1$ nonzero coordinates. Notice that our prediction only depends on the inputs via the inner products.

In practice the capacity of linear hyperplanes is limited. Therefore a good idea is to project the inputs to a higher dimensional space and search for the optimal linear classifier in the feature space. This leads to the notion of positive definite kernels that can be interpreted as an inner product in a higher dimensional feature space.

4.3 Reproducing Kernel Hilbert Spaces

Definition 4.3.1. Let \mathbb{X} be a set. The symmetric bivariate function $k : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ is a positive definite kernel if for all $n \geq 1$ and $x_1, \dots, x_n \in \mathbb{X}$ the matrix $K \in \mathbb{R}^{n \times n}$ defined elementwise as $K_{i,j} \doteq k(x_i, x_j)$ is positive semidefinite.

The data dependent positive semidefinite matrix K is called the Gram matrix. If we want to generalize the method of RR and the linear SVM we need to find a reasonably broad family of functions where the optimization can be carried out. Reproducing kernel Hilbert spaces have good properties both in the statistical and computational aspects.

Definition 4.3.2. Given a Hilbert space \mathcal{H} of $f : \mathbb{X} \rightarrow \mathbb{R}$ type functions, with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, we say that it is a reproducing kernel Hilbert space (RKHS) if the point evaluation function $\delta_x : f \rightarrow f(x)$ is bounded (or equivalently continuous) for all $x \in \mathbb{X}$.

In this case, by the Riesz representation theorem, there uniquely exists $k(\cdot, \cdot)$, such that for all $x \in \mathbb{X}$, $k(\cdot, x) \in \mathcal{H}$ and $f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$. This is called the *reproducing property*, and the function $k : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ is called the *reproducing kernel*. In particular $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}} = k(x, y)$ thus k is symmetric. The following claims are from [1] and [25].

Claim 4.3.1. The reproducing kernel k is positive semidefinite.

Proof. For any $n \geq 1$ and x_1, \dots, x_n consider the Gram matrix K . For all $a = (a_1, \dots, a_n)^T$ by the reproducing property we have

$$a^T K a = \sum_{i=1}^n \sum_{j=1}^n a_i a_j K_{i,j} = \left\langle \sum_{i=1}^n a_i k(\cdot, x_i), \sum_{j=1}^n a_j k(\cdot, x_j) \right\rangle = \left\| \sum_{i=1}^n a_i k(\cdot, x_i) \right\|_{\mathcal{H}}^2 \geq 0. \quad (4.19)$$

□

Claim 4.3.2. For a RKHS the positive definite reproducing kernel function is unique.

Proof. Let k_1 and k_2 be two kernels in \mathcal{H} with the reproducing property. By the reproducing property and the symmetry

$$k_1(x, y) = \langle k_1(\cdot, y), k_2(\cdot, x) \rangle = \langle k_2(\cdot, x), k_1(\cdot, y) \rangle = k_2(y, x) = k_2(x, y). \quad (4.20)$$

□

The converse is also true by the Moore-Arnoszjan theorem, see [1].

Theorem 4.3.1. *For each positive definite function k there uniquely exists a RKHS in which the kernel satisfies the reproducing property, i.e.*

$$\langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = f(x) \quad \forall f \in \mathcal{H}. \quad (4.21)$$

Proof. Notice that \mathcal{H} must contain all functions with form $\sum_{i=1}^n a_i k(\cdot, x_i)$ for all $a_1, \dots, a_n \in \mathbb{R}$ and $x_1, \dots, x_n \in \mathbb{X}$. Let denote the vector space

$$\widetilde{\mathcal{H}} \doteq \left\{ f : f(\cdot) = \sum_{i=1}^n a_i k(\cdot, x_i), n \in \mathbb{N}, a_1, \dots, a_n \in \mathbb{R}, x_1, \dots, x_n \in \mathbb{X} \right\}. \quad (4.22)$$

In order to ensure that the reproducing property holds the inner product should be defined as

$$\langle k(\cdot, x), k(\cdot, z) \rangle \doteq k(x, y) \quad (4.23)$$

for all $x, z \in \mathbb{X}$. Then for any $f(\cdot) = \sum_{i=1}^n a_i k(\cdot, x_i)$ and $g(\cdot) = \sum_{j=1}^m b_j k(\cdot, z_j)$ the inner product is independent of the expansion of f and g , because

$$\begin{aligned} \langle f, g \rangle &= \left\langle \sum_{i=1}^n a_i k(\cdot, x_i), \sum_{j=1}^m b_j k(\cdot, z_j) \right\rangle \\ &= \sum_{i=1}^n \sum_{j=1}^m a_i b_j k(x_i, z_j) = \sum_{i=1}^n a_i g(x_i) = \sum_{j=1}^m b_j f(z_j). \end{aligned} \quad (4.24)$$

Furthermore, substituting $k(\cdot, x)$ for g yields that the reproducing property holds. The defined inner product is clearly symmetric and linear. In order to prove its positive definiteness let $f(\cdot) = \sum_{i=1}^n a_i k(\cdot, x_i)$. Then

$$\langle f, f \rangle = \sum_{i=1}^n \sum_{j=1}^n a_i a_j k(x_i, x_j) = a^T K a \geq 0. \quad (4.25)$$

We show that $\langle f, f \rangle = 0$ if and only if $f = 0$. Let $\lambda \in \mathbb{R}$ and $x \in \mathbb{X}$, then

$$0 \leq \left\| \lambda k(\cdot, x) + \sum_{i=1}^n a_i k(\cdot, x_i) \right\|_{\mathcal{H}}^2 = \lambda^2 k(x, x) + \lambda \sum_{i=1}^n a_i k(x, x_i). \quad (4.26)$$

Since λ is arbitrary $\sum_{i=1}^n a_i k(x, x_i)$ needs to be 0.

We need to close $\widetilde{\mathcal{H}}$ to obtain a Hilbert space \mathcal{H} . It can be done on the usual manner via a density argument.

It remained to prove that the RKHS is unique. Let \mathcal{G} be another Hilbert space including \mathcal{H} which is the minimal Hilbert space induced by $\widetilde{\mathcal{H}}$. Let $h \in \mathcal{G} \ominus \mathcal{H}$. We know that \mathcal{H} is closed and linear, thus $h \perp \mathcal{H}$. Since $k(\cdot, x) \in \mathcal{H}$ by the orthogonality and the reproducing property it follows that $0 = \langle h, k(\cdot, x) \rangle = h(x)$ for all $x \in \mathbb{X}$. Therefore $h = 0$ and $\mathcal{G} = \mathcal{H}$. \square

Typical examples for reproducing kernels are the Gaussian kernel, $k(x, y) = \exp\left(\frac{-\|x-y\|^2}{2\sigma^2}\right)$ with $\sigma > 0$, the Laplacian kernel, $k(x, y) = \exp\left(\frac{-\|x-y\|}{\sigma}\right)$ with $\sigma > 0$, and the polynomial kernel,

$k(x, y) = (x^T y + c)^d$ with $c \geq 0$ and $d \in \mathbb{N}$.

Definition 4.3.3. (*universal kernel*) Let $C_b(\mathbb{X})$ denote the space of bounded continuous functions on a compact metric space \mathbb{X} with the supremum norm. A kernel is universal if the linear space $\text{span}\{k(\cdot, x) : x \in \mathbb{X}\} \subseteq \mathcal{H}$ is dense in $C_b(\mathbb{X})$: for all $f \in C_b(\mathbb{X})$ and $\varepsilon > 0$ there exists $\lambda_1, \dots, \lambda_l \in \mathbb{R}$ and $x_1, \dots, x_l \in \mathbb{X}$ such that $\sup_{u \in \mathbb{X}} \left| f(u) - \sum_{i=1}^l \lambda_i k(u, x_i) \right| < \varepsilon$.

That is, any bounded continuous function on \mathbb{X} can be uniformly approximated arbitrarily well with universal kernels, see [16]. For example the Gaussian and the Laplacian kernels are universal, see [17].

4.4 Representer Theorem

Optimizing over infinite dimensional Hilbert spaces sounds computationally demanding, still the representer theorem makes the RKHS attractive in many cases.

Theorem 4.4.1. (*representer theorem*) Given a sample $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ and a positive definite kernel $k(\cdot, \cdot)$, an associated RKHS \mathcal{H} with a norm $\|\cdot\|_{\mathcal{H}}$ induced by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, then, for any strictly monotonically increasing regularizer, $\kappa : [0, \infty) \rightarrow [0, \infty)$, and for arbitrary loss function $L : (\mathbb{X} \times \mathbb{R}^2)^n \rightarrow \mathbb{R} \cup \{\infty\}$ each minimizer of the criterion

$$\nu(f, \mathcal{D}) \doteq L((x_1, y_1, f(x_1)), \dots, (x_n, y_n, f(x_n))) + \kappa(\|f\|_{\mathcal{H}}) \quad (4.27)$$

admits the following representation $f(x) = \sum_{i=1}^n a_i k(x, x_i)$.

Observe that instead of solving an optimization problem in a high, often infinite dimensional Hilbert space, it is enough to minimize in a finite, at most n , dimensional linear space and the solution admits a finite dimensional form. That is why the representer theorem is a very powerful tool, and in many optimization problems the kernel values can be used instead of standard inner products. The proof is from the book of Scölkopf and Smola, see [18].

Proof. First notice that $\tilde{\kappa}(\|f\|_{\mathcal{H}}^2)$ can be used instead of $\kappa(\|f\|_{\mathcal{H}})$ because the quadratic function is strictly monotone increasing on $[0, \infty)$, implying that κ is strictly monotone increasing if and only if $\tilde{\kappa}$ is strictly monotone increasing.

Consider the span of the functions $k(\cdot, x_1), \dots, k(\cdot, x_n)$. It is a finite dimensional closed linear space, therefore all $f \in \mathcal{H}$ can be decomposed as

$$f = f_{\parallel} + f_{\perp}, \quad (4.28)$$

where $f_{\parallel} = \sum_{i=1}^n a_i k(\cdot, x_i)$ and $f_{\perp} \perp k(\cdot, x_i)$ for all $i \in [n]$. By the orthogonality and the reproducing property for all $j \in [n]$

$$f(x_j) = \left\langle \sum_{i=1}^n a_i k(\cdot, x_i), k(\cdot, x_j) \right\rangle + \langle f_{\perp}(\cdot), k(\cdot, x_j) \rangle = \sum_{i=1}^n a_i k(x_i, x_j) + 0, \quad (4.29)$$

that is $f(x_j)$ is independent of f_\perp from which it follows that the value $L((x_1, y_1, f(x_1)), \dots, (x_n, y_n, f(x_n)))$ is also independent of f_\perp . Furthermore, by the Pythagorean theorem

$$\kappa(\|f\|_{\mathcal{H}}) = \tilde{\kappa}\left(\left\|\sum_{i=1}^n a_i k(\cdot, x_i)\right\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2\right) \geq \tilde{\kappa}\left(\left\|\sum_{i=1}^n a_i k(\cdot, x_i)\right\|_{\mathcal{H}}^2\right) = \kappa(\|f_\parallel\|_{\mathcal{H}}), \quad (4.30)$$

therefore $f_\perp = 0$ when f is a minimizer. \square

Though the representer theorem does not say anything about the existence of a minimizer, in practice the form of the loss function often ensures it. Notice that if κ is monotone increasing but not strictly monotone the same proof yields that if a minimizer exists, then there is one with finite dimensional representation. By this theorem RR and SVM can be generalized for higher dimensional optimization problems in a RKHS, because instead of the normal inner products $\langle x_i, x_j \rangle$ in the euclidean space, we can use the inner products in the RKHS $\langle k(\cdot, x_i), k(\cdot, x_j) \rangle_{\mathcal{H}}$, which can be efficiently computed with the kernel function.

4.5 Kernel Mean Embedding

An important intuition behind the kernel functions is that it can be viewed as a feature map. Namely the function $x \rightarrow k(\cdot, x)$ maps the inputs to a higher dimensional feature space, which is the RKHS.

Similarly we can define a mapping from the sets of distributions on a space \mathbb{X} to the elements of a RKHS with the help of the kernel. This is the idea of *kernel mean embedding*, see [17].

Definition 4.5.1. (*kernel mean embedding*) Let (\mathbb{X}, Σ) be a measurable space and let $M_+(\mathbb{X})$ denote the space of all probability measures on it. The kernel mean embedding of these probability measures into a RKHS \mathcal{H} endowed with a reproducing kernel $k : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ is defined as

$$\begin{aligned} \mu &: M_+(\mathbb{X}) \rightarrow \mathcal{H}, \\ P &\rightarrow \int k(x, \cdot) dP(x), \end{aligned} \quad (4.31)$$

where the integral is a Bochner integral.

The following claim provides a sufficient condition for the existence of the kernel mean embedding, see [17].

Claim 4.5.1. If $\mathbb{E}_{X \sim P}[\sqrt{k(X, X)}] < \infty$, then $\mu_P \in \mathcal{H}$ and $\mathbb{E}_{X \sim P}[f(X)] = \langle f, \mu_P \rangle_{\mathcal{H}}$

Proof. Consider the linear functional $L_P f \doteq \mathbb{E}_{X \sim P}[f(X)]$. It is bounded, because for all $f \in \mathcal{H}$

$$|L_P f| \leq \mathbb{E}_{X \sim P} |f(X)| = \mathbb{E}_{X \sim P} [|\langle f(\cdot), k(\cdot, X) \rangle|] \leq \mathbb{E}_{X \sim P} [\sqrt{k(X, X)}] \|f\|_{\mathcal{H}}, \quad (4.32)$$

where we applied Jensen's inequality and the Cauchy-Schwarz inequality. By the Riesz representation theorem we obtain that there exists $\mu_P \in \mathcal{H}$ such that $L_P f = \langle \mu_P, f \rangle_{\mathcal{H}}$. Then for

$f(\cdot) = k(\cdot, x)$ with an arbitrary $x \in \mathbb{X}$ it follows that

$$\mu_P(x) = \langle \mu_P(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = \mathbb{E}_{X \sim P}[k(X, x)] = \int k(u, x) dP(u), \quad (4.33)$$

thus the claim is proved. \square

A kernel is called *characteristic* if the embedding, μ , is *injective* (e.g., the Gaussian kernel). In this case the embedded element captures all informations about the distribution, e.g., for all $P, Q \in M_+(\mathbb{X})$, $\|\mu_P - \mu_Q\|_{\mathcal{H}} = 0$ if and only if $P = Q$. Hence, the embedding induces a metric on $M_+(\mathbb{X})$. When \mathbb{X} is a compact metric space and k is a universal kernel on \mathbb{X} , then one can show that k is also characteristic [17].

The kernel mean embedding has good properties even if the kernel is not characteristic. For example, for polynomial kernels with degree d it holds that $\|\mu_P - \mu_Q\|_{\mathcal{H}} = 0$ if and only if the first d moments of P and Q are the same, see [17].

Furthermore, several fundamental operations can be performed in \mathcal{H} instead of dealing with the distributions themselves. For example kernel mean embedding can be performed on conditional distributions and the conditional expected value can be expressed via an inner product similarly to Claim 4.5.1. Consequently a wide range of tools can be performed in \mathcal{H} such as the sum, product and Bayes rules, see [17].

The underlying probability distribution of the sample is typically unknown, therefore the kernel mean embedding should be estimated from empirical data. An important tool to prove the validity of such approaches is the *Strong Law of Large Numbers* (SLLN) for random elements taking values in a *separable* Hilbert space \mathcal{H} , see [22, Theorem 3.2.4].

Theorem 4.5.1. *Let $\{X_n\}_{n \geq 1}$ be a sequence of independent random elements taking values in a separable Hilbert space \mathcal{H} . If*

$$\sum_{n=1}^{\infty} \frac{D^2(X_n)}{n^2} < \infty \quad (4.34)$$

where $D^2(X) \doteq \mathbb{E}[\|X - \mathbb{E}[X]\|_{\mathcal{H}}^2]$, then

$$\left\| \frac{1}{n} \sum_{k=1}^n (X_k - \mathbb{E}[X_k]) \right\|_{\mathcal{H}} \xrightarrow{a.s.} 0 \quad \text{as } n \rightarrow \infty. \quad (4.35)$$

Chapter 5

Confidence Regions

In several cases it is not enough to find a good estimator for a target function, but we also need to quantify the uncertainty of the estimation for example for the purpose of stability, safety or quality. In this chapter we are going to deal with the problem of classification and we want to find non-asymptotic stochastic guarantee tags for finding the unknown regression function. In this chapter the presented results were carried out together by me and my supervisor. In the conference paper [7] we have already published some of these results.

We present the uncertainty quantification via confidence sets. Notice that usually the probability that a point estimate f_n hits the target function exactly is zero, i.e. $\mathbb{P}(f_n = f_*) = 0$. *Confidence regions* were introduced to overcome this difficulty by constructing a set of estimator functions that contains the target function with high probability. In this thesis we present a broad framework which enables us to build *exact* confidence regions, that contain the regression function with a user-chosen probability, under mild statistical conditions. These random sets provide us bounds on the misclassification probabilities as well, that are also very important in practice. The methods, that are going to be presented, are *distribution-free* and *non-asymptotic*, that is the confidence bounds will hold for all distributions and sample sizes. Furthermore, we consider a broad model class for the regression function, e.g. the model class can be infinite dimensional, and nonparametric. Beside the finite sample advantages of the construction schemes we also provide bounds for the asymptotical behavior of our region estimates.

5.1 Resampling Framework

We have seen already (see Claim 1.1.2) that the regression function is identical to the conditional expected value function in case of the 0/1 loss, i.e.

$$f_*(x) \doteq \mathbb{E}[Y \mid X = x] = 2 \cdot \mathbb{P}(Y = +1 \mid X = x) - 1 = 2\eta(x) - 1. \quad (5.1)$$

In this chapter we assume that

(A0) the sample $\mathcal{D}_0 = \{(X_i, Y_i)\}_{i=1}^n$ is i.i.d.,

(A1) given a parameterized family of the possible regression functions that contains the true regression function, i.e. $f_* \in \mathcal{F} \doteq \{f_\theta : \mathbb{X} \rightarrow [-1, +1] \mid \theta \in \Theta\}$,

(A2) the parameterization is injective in the $L_2(P_X)$ sense, that is for all $\theta_1 \neq \theta_2 \in \Theta$

$$\|f_{\theta_1} - f_{\theta_2}\|_P^2 \doteq \int_{\mathbb{X}} (f_{\theta_1}(x) - f_{\theta_2}(x))^2 dP_X(x) \neq 0. \quad (5.2)$$

For the sake of simplicity we call Θ the parameter space, but it can be infinite dimensional. For examples the functions themselves can be the parameters. Let the true parameter, which corresponds to the true regression function, be θ_* , that is $f_{\theta_*} = f_*$.

As an example we consider the case when we observe a sample point from class “+1” with probability p or a sample point from class “−1” with probability $1-p$ and the input distributions corresponding to class “+1” and “−1” are determined by density functions φ_1 and φ_2 . Then it is easy to see that the regression function has the following form

$$f_*(x) = \mathbb{E}[Y \mid X = x] = \frac{p\varphi_1(x) - (1-p)\varphi_2(x)}{p\varphi_1(x) + (1-p)\varphi_2(x)} \mathbb{I}(p\varphi_1(x) + (1-p)\varphi_2(x) \neq 0). \quad (5.3)$$

Observe that if we have candidate densities for inputs with various labels and we know their mixing probability, then we can compute the regression function. However, we see that the regression function does not determine φ_1, φ_2 and p .

These type of regression functions were the test objects for the numerical examples that are presented in the end of this chapter. Notice that f_* does not determine the joint distribution of the sample. In fact it contains almost no information about the distribution of the inputs. Therefore our approach can be called semi-parametric.

Notice that the observed i.i.d. input-output dataset can be seen as an \mathbb{S}^n -valued random vector, where $\mathbb{S} = \mathbb{X} \times \{+1, -1\}$. One of our core ideas is that if a candidate θ is given, then we can generate (resample) alternative labels for the available inputs using the conditional distribution induced by f_θ , which is

$$\begin{aligned} \mathbb{P}_\theta(Y = +1 \mid X = x) &= \frac{f_\theta(x) + 1}{2}, \\ \mathbb{P}_\theta(Y = -1 \mid X = x) &= \frac{1 - f_\theta(x)}{2}, \end{aligned} \quad (5.4)$$

as it immediately follows from our observations in (5.1).

Given a θ , we generate $m - 1$ alternative samples. Let these be

$$\mathcal{D}_i(\theta) \doteq ((X_1, Y_{i,1}(\theta)), \dots, (X_n, Y_{i,n}(\theta))), \quad (5.5)$$

for $i = 1, \dots, m - 1$, where for all $(i, j) \in [n] \times [m - 1]$, label $Y_{i,j}(\theta)$ is generated randomly according to the conditional distribution $\mathbb{P}_\theta(Y \mid X = X_j)$. For notational simplicity, we extend this to \mathcal{D}_0 , that is $\forall \theta : \mathcal{D}_0(\theta) \doteq \mathcal{D}_0$ and $\forall j \in [n] Y_{0,j}(\theta) \doteq Y_j$.

Naturally, for all index i , dataset $\mathcal{D}_i(\theta)$ can also be identified with a random vector in \mathbb{S}^n , and $\mathcal{D}_1(\theta), \dots, \mathcal{D}_{m-1}(\theta)$ are always *conditionally i.i.d.*, given the inputs, $\{X_j\}_{j=1}^n$. Observe that

in case $\theta \neq \theta^*$ the distribution of \mathcal{D}_0 is in general different than that of $\mathcal{D}_i(\theta)$, $\forall i \neq 0$; while \mathcal{D}_0 and $\mathcal{D}_i(\theta^*)$ have the same distribution for $i \in [m - 1]$.

Based on these observations our methods are going to operate on the following manner. For a given θ we generate $m - 1$ alternative samples. If these samples are “similar” to the original one we include the examined parameter, otherwise we exclude the parameter from the confidence region. Consequently we need to derive a method to compare the datasets.

Our algorithms do the comparison via rank statistics. The key notion here is the *ranking function*, which informally compares its first argument to the other arguments.

Definition 5.1.1. (*ranking function*) Let \mathbb{A} be a measurable space (with some σ -algebra), a (measurable) function $\psi : \mathbb{A}^m \rightarrow [m]$, where $[m] \doteq \{1, \dots, m\}$, is called a ranking function if for all $(a_1, \dots, a_m) \in \mathbb{A}^m$ it satisfies the following properties:

(P1) For all permutations μ of the set $\{2, \dots, m\}$, we have

$$\psi(a_1, a_2, \dots, a_m) = \psi(a_1, a_{\mu(2)}, \dots, a_{\mu(m)}),$$

that is the function is invariant w.r.t. reordering the last $m - 1$ terms of its arguments.

(P2) For all $i, j \in [m]$, if $a_i \neq a_j$, then we have

$$\psi(a_i, \{a_k\}_{k \neq i}) \neq \psi(a_j, \{a_k\}_{k \neq j}), \quad (5.6)$$

where the simplified notation is justified by (P1).

The value taken by the ranking function is called the *rank*. An important observation about the rank of exchangeable variables is the following.

Lemma 5.1.1. Let A_1, \dots, A_m be exchangeable, almost surely pairwise different random elements taking values in \mathbb{A} . Then, $\psi(A_1, A_2, \dots, A_m)$ has discrete uniform distribution, that is for all $k \in [m]$

$$\mathbb{P}(\psi(A_1, A_2, \dots, A_m) = k) = \frac{1}{m}. \quad (5.7)$$

Proof. Since random elements $\{A_i\}_{i=1}^m$ are exchangeable

$$\mathbb{P}(\psi(A_1, \dots, A_m) = k) = \mathbb{P}(\psi(A_{\mu(1)}, \dots, A_{\mu(m)}) = k) \quad (5.8)$$

for all $k \in [m]$ and for all permutations μ on set $[m]$. Fix the value of k . In addition, notice that because of property P1 we have

$$\{\psi(A_1, \dots, A_m) = k\} = \{\psi(A_1, A_{\sigma(2)}, \dots, A_{\sigma(m)}) = k\} \quad (5.9)$$

for all permutation σ on set $\{2, \dots, m\}$, therefore similarly to the convention in P2 we can use the notation $C_i \doteq \{\psi(A_i, \{A_j\}_{j \neq i}) = k\}$. It is easy to see that events $\{C_i\}_{i=1}^m$ are disjoint and they cover an event which occurs with probability 1, i.e. $\cup_{i=1}^m C_i \supseteq \Omega \setminus \Omega_0$ where Ω_0 happens with probability zero. The disjoint property of the events is ensured by P2. Let Ω_0 denote the zero

probability event which contains the cases when there exists i and $j \in [m]$ such that $A_i = A_j$. Then for all $\omega \in \Omega \setminus \Omega_0$ the values $\psi(A_j(\omega), \{A_k(\omega)\}_{k \neq j})$ are distinct for all indices $j \in [m]$ because of P2. Furthermore, by definition these values are in $[m]$, therefore there exists an index $i \in [m]$ such that $\psi(A_i(\omega), \{A_k(\omega)\}_{k \neq i}) = k$, i.e. $\omega \in \bigcup_{i=1}^m C_i$. This is true for all $\omega \in \Omega \setminus \Omega_0$, thus $\Omega \setminus \Omega_0 \subseteq \bigcup_{i=1}^m C_i$ holds, implying that $\bigcup_{i=1}^m C_i$ is a 1 probability event. Combining this with the disjoint property and applying that the random elements $\{A_i\}_{i=1}^m$ are exchangeables yield

$$\begin{aligned} 1 = \mathbb{P}(\Omega \setminus \Omega_0) &= \mathbb{P}\left(\bigcup_{i=1}^m C_i\right) = \sum_{i=1}^m \mathbb{P}(C_i) = \sum_{i=1}^m \mathbb{P}\left(\psi\left(A_i, \{A_j\}_{j \neq i}\right) = k\right) \\ &= \sum_{i=1}^m \mathbb{P}\left(\psi\left(A_1, \{A_j\}_{j \neq 1}\right) = k\right) = m \mathbb{P}\left(\psi\left(A_1, \dots, A_m\right) = k\right). \end{aligned} \quad (5.10)$$

Dividing both sides by m we obtain that $\mathbb{P}\left(\psi\left(A_1, \dots, A_m\right) = k\right) = \frac{1}{m}$ for all $k \in [m]$. \square

Observe that this lemma does not assume anything about the distribution of the random elements $\{A_i\}_{i=1}^n$, only exchangeability is required, which holds for the original and the alternative samples generated from the conditional distribution determined by θ^* . In order to ensure pairwise difference, which does not always hold, we can extend the samples with the different elements of a random permutation $\pi : [m] \rightarrow [m]$ by setting $\mathcal{D}_i^\pi(\theta) \doteq (\mathcal{D}_i(\theta), \pi(i))$ for all $i = 0, \dots, m-1$. With this idea we can apply the Lemma 5.1.1 on exchangeable samples in general.

5.2 Non-Asymptotic Confidence Regions

Inspired by finite sample system identification methods [5, 6, 15], the core idea of the proposed algorithms is to compare the original dataset to alternative samples which are randomly generated according to a given hypothesis. The comparison will be based on the rank of the original dataset among all the available samples, therefore the ranking function is in the heart of all proposed algorithms and the differences between the presented methods primarily originate from the various ways they rank.

First we state the general result for arbitrary ranking functions, then we present four approaches to define concrete examples for ranking. Let ψ be a ranking function on the extended samples, i.e. $\psi : (\mathbb{X} \times \mathbb{Y})^m \times [m] \rightarrow [m]$. Furthermore, let $p, q \in [m]$ be user-chosen hyperparameters such that $p \leq q$ hold. Then we define a *confidence region* as

$$\Theta_\varrho^\psi \doteq \left\{ \theta \in \Theta : p \leq \psi\left(\mathcal{D}_0^\pi, \{\mathcal{D}_k^\pi(\theta)\}_{k \neq 0}\right) \leq q \right\}, \quad (5.11)$$

where $\varrho \doteq (m, p, q)$ denotes the hyperparameters. We are going to see that the probability level of the confidence region can be controlled by varying m , p and q and we can reach any (rational) probability level. Now we prove our main abstract result in this thesis with the help of Lemma 5.1.1.

Theorem 5.2.1. Assume that A0, A1 and A2 hold. Then for all ranking function ψ and

hyper-parameter $\varrho = (m, p, q)$ with integers $1 \leq p \leq q \leq m$,

$$\mathbb{P}(\theta^* \in \Theta_{\varrho}^{\psi}) = \frac{q - p + 1}{m}. \quad (5.12)$$

Proof. It is clear that $\mathcal{D}_0(\theta^*), \mathcal{D}_1(\theta^*), \dots, \mathcal{D}_{m-1}(\theta^*)$ are identically distributed and conditionally independent with respect to the inputs $\{X_i\}_{i=1}^n$, hence they are exchangeables. We extend these samples with the different elements of a random permutation $\pi : [m] \rightarrow [m]$ generated independently uniformly from the symmetric group. Observe that the extended samples, $\mathcal{D}_0^{\pi}(\theta^*), \mathcal{D}_1^{\pi}(\theta^*), \dots, \mathcal{D}_{m-1}^{\pi}(\theta^*)$, are almost surely pairwise different and still exchangeables implying that we can apply Lemma 5.1.1. It follows that for all values $k \in [m]$ the rank $\psi(\mathcal{D}_0^{\pi}, \{\mathcal{D}_k^{\pi}(\theta)\}_{k \neq 0}) = k$ with probability $\frac{1}{m}$, thus $p \leq \psi(\mathcal{D}_0^{\pi}, \{\mathcal{D}_k^{\pi}(\theta)\}_{k \neq 0}) \leq q$ with probability $\frac{q-p+1}{m}$. \square

Observe that this theorem guarantees an exact covering probability level under very mild statistical conditions. We do not assume anything about the distribution of the sample, i.e. the result is *distribution-free*. Furthermore, the theorem holds for any finite sample size providing us non-asymptotic guarantee tags. Moreover, the probability level can be chosen in advance and any rational value is reachable.

Because of the generality some degenerate construction is allowed. For example we can define ranking functions which only depend on the tie-breaking random permutations that are appended to the datasets. We would like to avoid such cases, therefore we analyze the asymptotic behaviour of the defined algorithms.

The so-called (pointwise) universal strong consistency is considered. Intuitively a method is universally strongly consistent if for all possible distributions of the sample any bad parameters are excluded from the confidence regions when the data size goes to infinity. Formally:

Definition 5.2.1. (*universal strong consistency*) A method is universally strongly consistent if for all distributions of (X, Y) we have

$$\mathbb{P} \left(\bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} \{ \theta \in \Theta_{\varrho, n}^{\psi} \} \right) = 0, \quad (5.13)$$

for all parameters $\theta \neq \theta^*$, $\theta \in \Theta$, where $\Theta_{\varrho, n}^{\psi}$ denotes the confidence region constructed from a sample of size n .

Obviously purely randomized algorithms are not consistent. We also consider a stronger version of the consistency where the bad parameters are excluded uniformly. For this definition we consider the confidence sets in the function space, i.e. let

$$\mathcal{F}_{\varrho}^{\psi} \doteq \{ f_{\theta} \in \mathcal{F} : \theta \in \Theta_{\varrho}^{\psi} \}. \quad (5.14)$$

The $L_2(P_X)$ norm induces a metric on \mathcal{F} which is a natural choice to define uniformity. Let $B(f_*, \varepsilon)$ denote the closed ball with center f_* and radius ε in the $L_2(P_X)$ -metric.

Definition 5.2.2. (*universal strong uniform consistency*) A method is universally strongly uniformly consistent if for all distributions of (X, Y) , for all $\varepsilon > 0$

$$\mathbb{P} \left(\bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} \{ \mathcal{F}_{\varrho, n}^{\psi} \subseteq B(f_*, \varepsilon) \} \right) = 1, \quad (5.15)$$

where $\mathcal{F}_{\varrho, n}^{\psi}$ denotes the confidence region in \mathcal{F} constructed from a sample of size n .

Sometimes it can be more convenient to consider a metric in the parameter space. When the inverse of the parameterization is Lipschitz-continuous with respect to the $L_2(P_X)$ -metric, i.e. there exists $L > 0$ such that for all $\theta_1, \theta_2 \in \Theta$ we have

$$\|\theta_1 - \theta_2\| \leq L \|f_{\theta_1} - f_{\theta_2}\|_{L_2(P_X)}, \quad (5.16)$$

from (5.15) it follows that strong uniform consistency holds in the parameter space as well, that is for all $\varepsilon > 0$

$$\mathbb{P} \left(\bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} \{ \Theta_{\varrho, n}^{\psi} \subseteq B(\theta^*, \varepsilon) \} \right) = 1, \quad (5.17)$$

In the following sections we introduce four algorithms which construct exact, non-asymptotic confidence regions and have advantageous asymptotic properties such as universal strong consistency or strong uniform consistency.

5.3 Algorithm I (ERM Based)

The first algorithm is based on empirical risk minimizer estimates. We assume that

(A3) the model space is part of a linear space, i.e. there exists a square-integrable basis $\varphi_k : \mathbb{R}^d \rightarrow [-1, 1]$ for $k = 1, \dots, r$ such that

$$\mathcal{F} = \left\{ f_{\theta} \mid f_{\theta}(x) = \sum_{i=1}^r \theta_i \varphi_i(x), \theta \in \mathbb{R}^r, \sup_{x \in \mathbb{R}^d} |f_{\theta}(x)| \leq 1 \right\}, \quad (5.18)$$

i.e. $\Theta \subseteq \mathbb{R}^r$. Let the linear space be denoted by $\tilde{\mathcal{F}}$.

(A4) The matrix determined elementwise for $k = 1, \dots, p$ and $j = 1, \dots, n$ as $\Psi_{j,k} = \varphi_k(X_j)$ is skinny ($n > r$) and has full rank ($\text{rank}(\Psi) = r$) with probability 1.

The idea is that we apply the ERM method to estimate the regression function based on the extended original sample, \mathcal{D}_0^{π} , and similarly for a given model parameter θ we estimate the regression function based on the extended alternative samples, $\{\mathcal{D}_i^{\pi}(\theta)\}_{i=1}^{m-1}$. Let

$$\tilde{f}_{\theta, n}^{(0)} \in \arg \min_{f \in \tilde{\mathcal{F}}} \frac{1}{n} \sum_{j=1}^n (f(X_j) - Y_j)^2, \quad (5.19)$$

$$\tilde{f}_{\theta, n}^{(i)} \in \arg \min_{f \in \tilde{\mathcal{F}}} \frac{1}{n} \sum_{j=1}^n (f(X_j) - Y_{i,j})^2, \quad (5.20)$$

for $i = 1, \dots, m-1$. Notice that these function estimates can be determined as least squares solutions that exist because of (A4). Unfortunately, these estimates are not necessarily bounded in $[-1, 1]$ therefore we truncate them, i.e.

$$f_{\theta,n}^{(i)} \doteq T_1 \tilde{f}_{\theta,n}^{(i)} \quad (5.21)$$

for $i = 0, \dots, m-1$, where $T_L f(x) \doteq \max(L, |f(x)|) \cdot \text{sign}(f(x))$ for $L \in \mathbb{R}^+$.

We define reference variables with the help of the empirical L_2 distance as

$$Z_n^{(i)}(\theta) \doteq \frac{1}{n} \sum_{i=1}^n (f_\theta(X_i) - f_{\theta,n}^{(i)}(X_i))^2 \quad (5.22)$$

for $i = 0, \dots, m-1$. Then, we can define the *rank* of $Z_n^{(0)}(\theta)$ among $\{Z_n^{(i)}(\theta)\}_{i=0}^{m-1}$ as

$$\mathcal{R}_n(\theta) \doteq 1 + \sum_{i=1}^{m-1} \mathbb{I}\left(Z_n^{(i)}(\theta) \prec_\pi Z_n^{(0)}(\theta)\right), \quad (5.23)$$

where binary relation “ \prec_π ” is the standard “ $<$ ” with random tie-breaking, similarly as in Section 3.6. Consequently in case of Algorithm I, the ranking function is

$$\psi\left(\mathcal{D}_0^\pi, \{\mathcal{D}_k^\pi(\theta)\}_{k \neq 0}\right) = \mathcal{R}_n(\theta). \quad (5.24)$$

As we will see (cf. the proof of Theorem 5.3.3), for any fixed false parameter, $Z_n^{(0)}(\theta)$ tends to have the largest rank, therefore, we fix $p = 1$ and only exclude parameters which lead to high ranks. That is, similarly as in (5.11), the confidence set is

$$\Theta_{\varrho,n}^{(1)} \doteq \left\{ \theta \in \Theta : \mathcal{R}_n(\theta) \leq q \right\}, \quad (5.25)$$

where $\varrho \doteq (m, q)$ again denotes the user-chosen hyper-parameters with $1 \leq q \leq m$. Let the confidence set in the model space induced by these parameters be

$$\mathcal{F}_{\varrho,n}^{(1)} \doteq \left\{ f_\theta \in \mathcal{F} : \mathcal{R}_n(\theta) \leq q \right\}. \quad (5.26)$$

Before analyzing the algorithm we state two theorems from [11] that are going to be applied to prove Theorem 5.3.3.

Theorem 5.3.1. *Assume that*

$$\sigma^2 = \sup_{x \in \mathbb{R}^d} D^2(Y | X = x) < \infty, \quad \text{and} \quad \sup_{x \in \mathbb{R}^d} |f_*(x)| \leq L \quad (5.27)$$

hold for some $L \in \mathbb{R}^+$. Let \mathcal{F} be a linear vector space of functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$ which contains f_ . Let r be the dimension of \mathcal{F} . Define estimator f_n as*

$$f_n = T_L \tilde{f}_n \quad \text{where} \quad \tilde{f}_n \in \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (f(X_i) - Y_i)^2. \quad (5.28)$$

Then there exists some universal c such that

$$\mathbb{E} \int (f_n(x) - f_*(x))^2 dP_X(x) \leq c \cdot \max(\sigma^2, L^2) \frac{(\log(n) + 1) \cdot r}{n} \quad (5.29)$$

The proof can be found in [11].

Theorem 5.3.2. Let $\tilde{\mathcal{F}}$ be an r dimensional linear space of functions $f : \mathbb{R}^d \rightarrow \mathbb{R}$. For $R > 0$, $\varepsilon > 0$ and $z_1, \dots, z_n \in \mathbb{R}^d$

$$\mathcal{N}_2\left(\varepsilon, \left\{f \in \mathcal{F} : \frac{1}{n} \sum_{i=1}^n |f(z_i)|^2 \leq R^2\right\}, \{z_1, \dots, z_n\}\right) \leq \left(\frac{4R + \varepsilon}{\varepsilon}\right)^r. \quad (5.30)$$

Proof. Fix z_1, \dots, z_n . Let $\mathcal{G} \doteq \{f \in \mathcal{F} : \frac{1}{n} \sum_{i=1}^n |f(z_i)|^2 \leq R^2\}$. For f and g we denote the inner product w.r.t. the empirical measure by $\langle f, g \rangle_n$, i.e.

$$\langle f, g \rangle_n \doteq \frac{1}{n} \sum_{i=1}^n f(z_i)g(z_i). \quad (5.31)$$

We consider the norm induced by the inner product and denote it by $\|f\|_n^2 = \langle f, f \rangle_n$. Let $\{f_1, \dots, f_N\}$ be an ε -packing of \mathcal{G} w.r.t. the defined norm. Because of Lemma 2.9.3 it is sufficient to prove that

$$N \leq \left(\frac{4R + \varepsilon}{\varepsilon}\right)^r. \quad (5.32)$$

Let $\varphi_1, \dots, \varphi_r$ be a basis of functions in \mathcal{F} and $a, b \in \mathbb{R}^r$. Let Φ be defined elementwise as $\Phi_{i,j} = \langle \varphi_i, \varphi_j \rangle_n$ for $1 \leq i, j \leq r$. Then

$$\left\| \sum_{i=1}^r a_i \varphi_i - \sum_{i=1}^r b_i \varphi_i \right\|_n^2 = (a - b)^T \Phi (a - b). \quad (5.33)$$

It is easy to see that Φ is positive semidefinite because $a^T \Phi a \geq 0$ for all $a \in \mathbb{R}^r$. Therefore there exists a symmetric matrix $\Phi^{1/2}$ such that $\Phi = \Phi^{1/2} \Phi^{1/2}$. Then

$$\|(a - b)^T \Phi^{1/2}\|^2 = (a - b)^T \Phi (a - b), \quad (5.34)$$

where $\|\cdot\|$ denotes the euclidean metric on \mathbb{R}^r . Let $f_i(x) = \sum_{j=1}^r a_j^{(i)} \varphi_j(x)$. Because of the definition of \mathcal{G}

$$\|\Phi^{1/2} a^{(i)}\| = \|f_i\|_n \leq R. \quad (5.35)$$

In addition for all $i \neq j$:

$$\|\Phi^{1/2} a^{(i)} - \Phi^{1/2} a^{(j)}\| = \|f_i - f_j\|_n \geq \varepsilon, \quad (5.36)$$

i.e. the euclidean balls $B(\Phi^{1/2} a^{(i)}, \varepsilon/4)$ for $i \in [N]$ in \mathbb{R}^r are disjoint. Furthermore, all of them are inside $B(0, R + \varepsilon/4)$. Therefore the euclidean volume of the sum of the small balls is smaller than the volume of the large ball centered in the origin, i.e.

$$N c_d \left(\frac{\varepsilon}{4}\right)^r \leq c_d \left(R + \frac{\varepsilon}{4}\right)^r \quad (5.37)$$

holds, where c_d is the volume of the r dimensional unit ball. It suffices for the proof. \square

The following theorem summarizes the most important properties of Algorithm I.

Theorem 5.3.3. *Assume that A0, A1, A3 and A4 hold, then*

$$\mathbb{P}\left(\theta^* \in \Theta_{\varrho,n}^{(1)}\right) = \frac{q}{m}, \quad (5.38)$$

for all sample size n . Furthermore, if $q < m$ Algorithm I is strongly uniformly consistent.

Proof. It is easy to show that the defined ranking function has properties P1 and P2, hence the coverage probability is exact because of Theorem [5.2.1](#)

The proof of strong uniform consistency will be presented in several steps. Let $f_{*,n}^{(0)}$ denote $f_{\theta,n}^{(0)}$ which is independent of θ , since this estimate is computed from the original sample. First notice that for all $f_\theta \in \mathcal{F} \setminus B(f_*, \varepsilon)$ and for all $i \in [m-1]$

$$\begin{aligned} Z_n^{(0)}(\theta) - Z_n^{(i)}(\theta) &= \frac{1}{n} \sum_{j=1}^n (f_\theta(X_j) - f_{\theta,n}^{(0)}(X_j))^2 - \frac{1}{n} \sum_{j=1}^n (f_\theta(X_j) - f_{\theta,n}^{(i)}(X_j))^2 \geq \\ &\geq \frac{1}{n} \sum_{j=1}^n (f_\theta(X_j) - f_*(X_j))^2 + \frac{1}{n} \sum_{j=1}^n (f_*(X_j) - f_{*,n}^{(0)}(X_j))^2 + \\ &+ 2 \frac{1}{n} \sum_{j=1}^n (f_\theta(X_j) - f_*(X_j))(f_*(X_j) - f_{*,n}^{(0)}(X_j)) - \frac{1}{n} \sum_{j=1}^n (f_\theta(X_j) - f_{\theta,n}^{(i)}(X_j))^2 \geq \\ &\geq \mathbb{E}[(f_\theta(X) - f_*(X))^2] - \mathbb{E}[(f_\theta(X) - f_*(X))^2] + \frac{1}{n} \sum_{j=1}^n (f_\theta(X_j) - f_*(X_j))^2 \\ &+ \mathbb{E}[(f_*(X) - f_{*,n}^{(0)}(X))^2] - \mathbb{E}[(f_*(X) - f_{*,n}^{(0)}(X))^2] + \frac{1}{n} \sum_{j=1}^n (f_*(X_j) - f_{*,n}^{(0)}(X_j))^2 \\ &- \mathbb{E}[(f_\theta(X) - f_{\theta,n}^{(i)}(X))^2] + \mathbb{E}[(f_\theta(X) - f_{\theta,n}^{(i)}(X))^2] - \frac{1}{n} \sum_{j=1}^n (f_\theta(X_j) - f_{\theta,n}^{(i)}(X_j))^2 \\ &- \frac{4}{n} \sum_{j=1}^n |f_*(X_j) - f_{*,n}^{(0)}(X_j)|. \end{aligned} \quad (5.39)$$

Since $f_\theta \notin B(f_*, \varepsilon)$ the following holds

$$\inf_{f \notin B(f_*, \varepsilon)} \mathbb{E}[(f_\theta(X) - f_*(X))^2] \geq \varepsilon^2 > 0. \quad (5.40)$$

Furthermore, bounding each term with a proper supremum yields

$$\begin{aligned} &\mathbb{E}[(f_\theta(X) - f_*(X))^2] - \frac{1}{n} \sum_{j=1}^n (f_\theta(X_j) - f_*(X_j))^2 \\ &+ \mathbb{E}[(f_*(X) - f_{*,n}^{(0)}(X))^2] - \frac{1}{n} \sum_{j=1}^n (f_*(X_j) - f_{*,n}^{(0)}(X_j))^2 \\ &- \mathbb{E}[(f_\theta(X) - f_{\theta,n}^{(i)}(X))^2] + \frac{1}{n} \sum_{j=1}^n (f_\theta(X_j) - f_{\theta,n}^{(i)}(X_j))^2 \end{aligned}$$

$$\leq 3 \sup_{f \in \mathcal{F}, g \in \mathcal{G}} \left| \frac{1}{n} \sum_{j=1}^n (f(X_j) - g(X_j))^2 - \mathbb{E}[(f(X) - g(X))^2] \right|, \quad (5.41)$$

where $\mathcal{G} \doteq T_1 \tilde{\mathcal{F}} = \{g \mid \exists f \in \tilde{\mathcal{F}} : g = T_1 f\}$. In addition, by Theorem 5.3.1 we know that

$$\mathbb{E}[(f_\theta(X) - f_{\theta,n}^{(i)}(X))^2] \leq c \cdot \max(\sigma^2, L^2) \frac{(\log(n) + 1)r}{n} \leq \frac{c \cdot r \cdot (\log(n) + 1)}{n} = a_n, \quad (5.42)$$

because $\sigma^2 = \sup_{x \in \mathbb{R}^d} D^2(Y \mid X = x) \leq 1$ and $L \leq 1$ for all $\theta \in \Theta$. Applying the Cauchy-Schwarz inequality for the last term and subtracting its expected value yields that

$$\begin{aligned} & \frac{4}{n} \sum_{i=1}^n |f_*(X_i) - f_{*,n}^{(0)}(X_i)| \leq 4 \sqrt{\frac{1}{n^2} \sum_{i=1}^n (f_*(X_i) - f_{*,n}^{(0)}(X_i))^2 \sum_{i=1}^n 1} \\ & \leq 4 \sqrt{\frac{1}{n} \sum_{i=1}^n (f_*(X_i) - f_{*,n}^{(0)}(X_i))^2 - \mathbb{E}[(f_*(X) - f_{*,n}^{(0)}(X))^2] + \mathbb{E}[(f_*(X) - f_{*,n}^{(0)}(X))^2]} \\ & \leq 4 \sqrt{\sup_{f,g} \left| \frac{1}{n} \sum_{i=1}^n (f(X_i) - g(X_i))^2 - \mathbb{E}[(f(X) - g(X))^2] \right| + a_n}. \end{aligned} \quad (5.43)$$

We obtain the following by combining these equations

$$\begin{aligned} & \inf_{f_\theta \notin B(f_*, \varepsilon), i \in [m-1]} (Z_n^{(0)}(\theta) - Z_n^{(i)}(\theta)) > \varepsilon^2 - a_n \\ & - 3 \sup_{f \in \mathcal{F}, g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n (f(X_i) - g(X_i))^2 - \mathbb{E}[(f(X) - g(X))^2] \right| \\ & - 4 \sqrt{\sup_{f,g} \left| \frac{1}{n} \sum_{i=1}^n (f(X_i) - g(X_i))^2 - \mathbb{E}[(f(X) - g(X))^2] \right| + a_n}, \end{aligned} \quad (5.44)$$

where a_n is a sequence with zero limit. Let $\mathcal{H} \doteq \{(f - g)^2 \mid f \in \mathcal{F}, g \in T_1 \tilde{\mathcal{F}}\}$. We are going to prove that the following holds

$$\sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n h(X_i) - \mathbb{E}h(X) \right| \xrightarrow{a.s.} 0. \quad (5.45)$$

The VC theory provides us applicable conditions for this ULLN. We are going to prove that there exists a universal C such that for all $n \in \mathbb{N}$ and sample $\{X_1, \dots, X_n\}$

$$\mathbb{E} \mathcal{N}_1(\varepsilon, \mathcal{H}, \{X_1, \dots, X_n\}) \leq C. \quad (5.46)$$

Then, Theorem 2.10.1 implies (5.45).

The proof of (5.46) will be in several steps. Fix a realization of X_1, \dots, X_n . Let these be $\mathbf{x} = \{x_1, \dots, x_n\}$. First we prove that

$$\mathcal{N}_1(\varepsilon, \mathcal{H}, \mathbf{x}) \leq \mathcal{N}_2(\varepsilon/8, \mathcal{F}, \mathbf{x}) \cdot \mathcal{N}_1(\varepsilon/8, \mathcal{G}, \mathbf{x}). \quad (5.47)$$

Let $\nu(A) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(x_i \in A)$ be the empirical measure based on \mathbf{x} . Let f_1, \dots, f_N be an $\varepsilon/8$ -

cover of \mathcal{F} w.r.t. the $L_2(\nu)$ norm and g_1, \dots, g_M an $\varepsilon/8$ -cover of \mathcal{G} w.r.t. the $L_1(\nu)$ norm. It can be assumed w.l.o.g. that f_i and g_j are bounded in $[-1, 1]$ (otherwise we can truncate them) for all $i \in [N]$ and $j \in [M]$. We are going to prove that the set $\{(f_i - g_j)^2 \mid i \in [N], j \in [M]\}$ is an ε -cover of \mathcal{H} w.r.t. the $L_1(\nu)$ norm. Let $h \in \mathcal{H}$ arbitrary fixed. Then there exists $f \in \mathcal{F}$ and $g \in \mathcal{G}$ such that $h = (f - g)^2$. For this f and g there are f_i and g_j such that

$$\begin{aligned} \|f - f_i\|_{L_2(\nu)} &\leq \varepsilon/8 \quad \text{and} \\ \|g - g_j\|_{L_1(\nu)} &\leq \varepsilon/8. \end{aligned} \tag{5.48}$$

Applying (5.48), the Cauchy-Schwarz inequality and that $f, g, f_i, g_j \in [-1, 1]$ yields that

$$\begin{aligned} \int |h - (f_i - g_j)^2| d\nu &\leq \int |(f - g)^2 - (f_i - g_j)^2| d\nu \\ &\leq \int |((f - g) + (f_i - g_j)) \cdot ((f - g) - (f_i - g_j))| d\nu \\ &\leq 4 \int (|f - f_i| + |g - g_j|) d\nu \leq 4 \sqrt{\int (f - f_i)^2 d\nu} + 4 \int |g - g_j| d\nu \leq \varepsilon \end{aligned} \tag{5.49}$$

In the second step we bound $\mathcal{N}_2(\varepsilon/8, \mathcal{F}, \mathbf{x})$. Notice that for all $f \in \mathcal{F}$ we know that $\frac{1}{n} \sum_{j=1}^n f^2(x_j) \leq 1$. We apply Theorem 5.3.2 to obtain that

$$\mathcal{N}_2(\varepsilon/8, \mathcal{F}, \mathbf{x}) \leq \left(\frac{4 + \varepsilon/8}{\varepsilon/8} \right)^r = C_1, \tag{5.50}$$

where C_1 is independent of the data points.

In the third step we bound $\mathcal{N}_1(\varepsilon/8, \mathcal{G}, \mathbf{x})$. First we prove that $V_{\mathcal{G}^+} \leq V_{\tilde{\mathcal{F}}^+}$, where recall that $V_{\mathcal{G}^+}$ denotes the VC dimension of the subgraphs of functions in \mathcal{G} . Notice that if \mathcal{G}^+ shatters $C = \{(x_1, t_1), \dots, (x_l, t_l)\} \subseteq \mathbb{R}^d \times \mathbb{R}$, then $t_i \in [-1, 1]$ for all $i = 1, \dots, l$ because if $t_1 < -1$ then $t_i \leq g(x_i)$ for all $g \in \mathcal{G}$ which contradicts the fact that \mathcal{G}^+ shatters C . The reasoning is similar for $t_i > 1$. When \mathcal{G}^+ shatters C and $t_i \in [-1, 1]$ then $\hat{\mathcal{F}}^+$ also shatters C , because for all $f \in \hat{\mathcal{F}}$ and $T_1 f \in \mathcal{G}$ we have

$$f^+ \cap C = T_1 f^+ \cap C. \tag{5.51}$$

Therefore $V_{\mathcal{G}^+} \leq V_{\tilde{\mathcal{F}}^+}$ holds. Furthermore, notice that

$$\begin{aligned} \tilde{\mathcal{F}}^+ &= \{ \{(x, t) \in \mathbb{R}^d \times \mathbb{R} \mid t \leq f(x)\} \mid f \in \hat{\mathcal{F}} \} \\ &\subseteq \mathcal{F}^{++} \doteq \{ \{(x, t) \in \mathbb{R}^d \times \mathbb{R} \mid \alpha t + f(x) \geq 0\} : \alpha \in \mathbb{R}, f \in \hat{\mathcal{F}} \}. \end{aligned} \tag{5.52}$$

If $\tilde{\mathcal{F}}$ is a vector space with dimension r , then

$$\{ \alpha t + f(x) \mid \alpha \in \mathbb{R}, f \in \tilde{\mathcal{F}} \} = \left\{ \alpha t + \sum_{i=1}^r \beta_i \varphi_i(x) \mid \alpha \in \mathbb{R}, \beta \in \mathbb{R}^r \right\} \tag{5.53}$$

is a vector space of $\mathbb{R}^d \times \mathbb{R} \rightarrow \mathbb{R}$ type functions with dimension $r + 1$. For all $(t, x) \in \mathbb{R} \times \mathbb{R}^d$

consider the point $(t, \varphi_1(x), \dots, \varphi_r(x))$ in \mathbb{R}^{r+1} , then

$$\text{sign} \left(\alpha t + \sum_{i=1}^r \beta_i \varphi(x) \right) = 1 \Leftrightarrow \alpha t + \sum_{i=1}^r \beta_i \varphi(x) \geq 0 \quad (5.54)$$

therefore $V_{\mathcal{F}^{++}}$ is at most the VC dimension of the linear homogenous hyperplanes in \mathbb{R}^{r+1} which equals to $r + 1$ by the proof of Claim 4.2.1. In conclusion

$$V_{\mathcal{G}^+} \leq V_{\tilde{\mathcal{F}}^+} \leq V_{\mathcal{F}^{++}} \leq r + 1. \quad (5.55)$$

To ensure the conditions of Lemma 2.9.3, we translate the functions in \mathcal{G} by 1 then divide the functions by 2, i.e. let $\tilde{\mathcal{G}} \doteq \{1/2(g + 1) \mid g \in \mathcal{G}\}$. Clearly the VC dimensions are not influenced by this adjustment, that is $V_{\mathcal{G}^+} = V_{\tilde{\mathcal{G}}^+}$, and for the covering numbers the following holds: $\mathcal{N}_1(\varepsilon, \mathcal{G}, \|\cdot\|_{L_1(\nu)}) = \mathcal{N}_1(\varepsilon/2, \tilde{\mathcal{G}}, \|\cdot\|_{L_p(\nu)})$. Then applying Lemma 2.9.3, Theorem 2.9.4 and (5.58) yields that for all $\varepsilon > 0$ we have

$$\begin{aligned} \mathcal{N}_1(\varepsilon/8, \mathcal{G}, \mathbf{x}) &= \mathcal{N}_1(\varepsilon/16, \tilde{\mathcal{G}}, \mathbf{x}) \leq \mathcal{M}(\varepsilon/16, \tilde{\mathcal{G}}, \|\cdot\|_{L_1(P_X)}) \\ &\leq e(V_{\tilde{\mathcal{G}}^+} + 1) \left(\frac{32e}{\varepsilon} \right)^{V_{\tilde{\mathcal{G}}^+}} \leq e(r + 2) \left(\frac{32e}{\varepsilon} \right)^{r+1} = C_2, \end{aligned} \quad (5.56)$$

where C_2 is independent of \mathbf{x} . Combining it together with (5.50) yields that for all X_1, \dots, X_n

$$\mathbb{E} \mathcal{N}_1(\varepsilon, \mathcal{H}, \{X_1, \dots, X_n\}) \leq C_1 \cdot C_2, \quad (5.57)$$

thus (5.46) holds.

By (5.46) and by $a_n \rightarrow 0$ it follows that

$$\liminf_{n \rightarrow \infty} \inf_{f_\theta \notin B(f_*, \varepsilon), i \in [m-1]} \left(Z_n^{(0)}(\theta) - Z_n^{(i)}(\theta) \right) \geq \varepsilon^2 > 0 \quad (5.58)$$

with probability 1.

Let Ω_1 be the 1 probability event where (5.58) holds. For all $\omega \in \Omega_1$ and for $\varepsilon^2/2$ there exists $n_0(\omega)$ such that for all $n \geq n_0(\omega)$, for all $i = 1, \dots, m-1$ and for all $f_\theta \notin B(f_*, \varepsilon)$ it holds that $Z_n^{(0)} - Z_n^{(i)}(\theta) > \varepsilon^2 - \varepsilon^2/2$, that is $\mathcal{R}_n(\theta) = m$ for all $f_\theta \notin B(f_*, \varepsilon)$. Hence $B(f_*, \varepsilon) \supseteq \mathcal{F}_{\varrho, n}^{(1)}(\omega)$ for all $n > n_0(\omega)$ implying that $\omega \in \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} \left\{ \mathcal{F}_{\varrho, n}^{(1)} \subseteq B(f_*, \varepsilon) \right\}$, therefore

$$\mathbb{P} \left(\bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} \left\{ \mathcal{F}_{\varrho, n}^{(1)} \subseteq B(f_*, \varepsilon) \right\} \right) = 1. \quad (5.59)$$

The theorem is proved. \square

5.4 Numerical Experiments I

We carried out numerical experiments on synthetic datasets to illustrate this algorithm. The linear model class was spanned by $\varphi_1(x) = \exp(-(x + 1)^2/2)$ and $\varphi_2(x) = \exp(-(x - 1)^2/2)$,

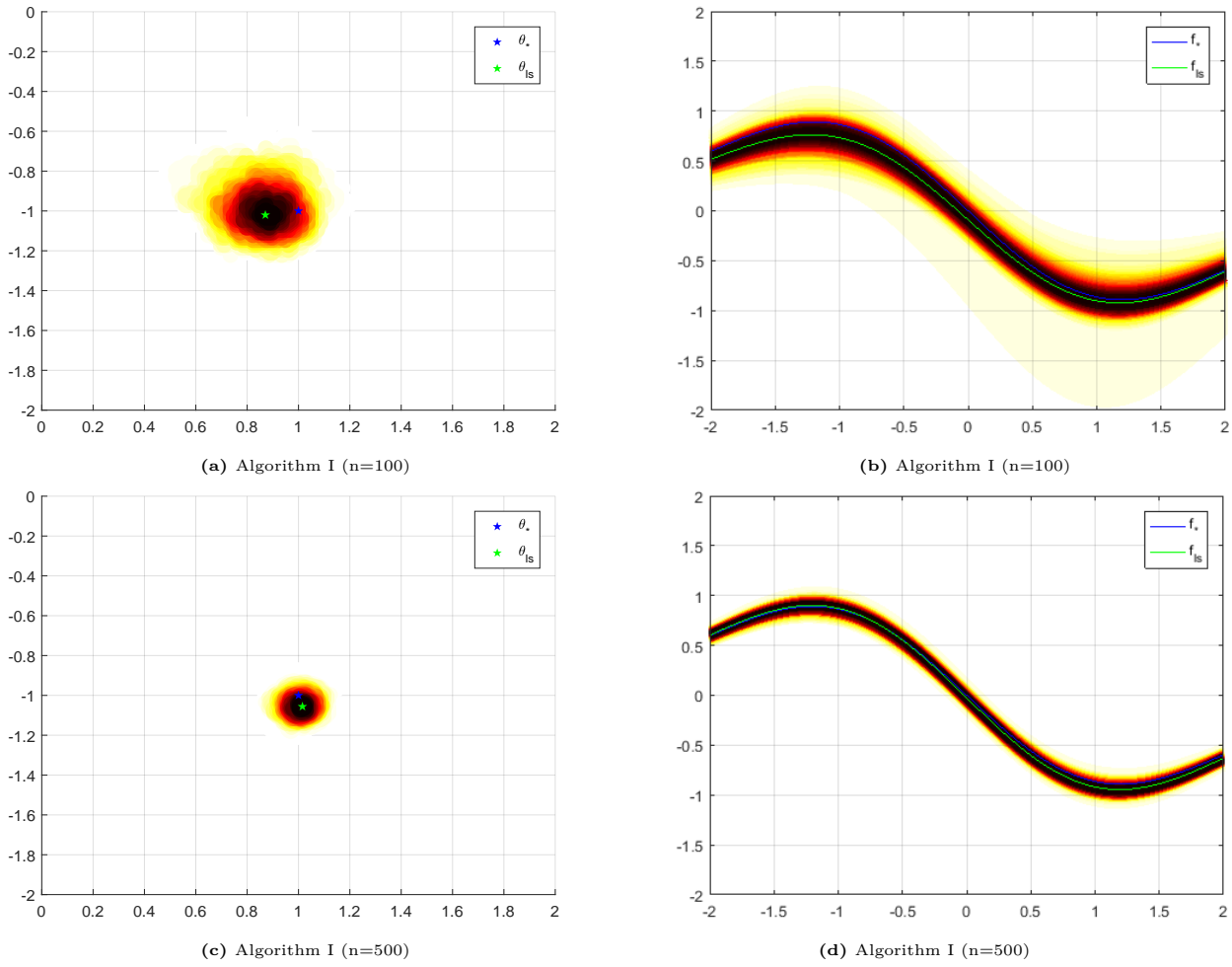


Figure 5.1: The ranks of the reference variables are indicated with the color for the parameters in figures (a) and (c), and for the models in figures (b) and (d), thus darker points and models indicate reference variables with lower ranks. The true parameter was $\theta_* = (\alpha_*, \beta_*)$, where $\alpha_* = 1$ (x -axis) and $\beta_* = -1$ (y -axis).

i.e. the model class was part of the linear space

$$\bar{\mathcal{F}} = \{f(x) = \alpha\varphi_1(x) + \beta\varphi_2(x) \mid \alpha, \beta \in \mathbb{R}\}. \quad (5.60)$$

The true regression function was $f_*(x) = \exp(-(x+1)^2/2) - \exp(-(x-1)^2/2)$, so we wanted to estimate the true parameters $\alpha_* = 1$ and $\beta_* = -1$. In this simple example the marginal distribution was uniform on $[-2, 2]$. The results of these tests can be seen in figure 5.1. The ranks of the reference variables are indicated with the color. The confidence regions are evaluated in the parameter space, see (a) and (c), and in the corresponding model space, see (b) and (d). The sample size was $n = 100$ for figures (a) and (b) and $n = 500$ for figures (c) and (d). We can see that the regions shrink around the true model as the sample size increases. We used $m = 40$ samples including both the original and the alternative ones. In conclusion when the parameterization is linear the constructed confidence regions provide us non-asymptotic guarantees for finding the true regression function, that can be used in practice.

5.5 Algorithm II (Local Averaging Based)

The second algorithm is based on local averaging kernel estimates. We want to generalize Algorithm I for more complex model classes, because the linear form is often too restrictive for a regression function. Notice that when \mathcal{F} is not parameterized linearly then application of the ERM principle can become computationally demanding, therefore instead of the least squares solution we are going to use a local averaging estimate. For this section we assume that

(B1) $\mathbb{X} \subseteq \mathbb{R}^d$, \mathbb{X} is compact,

(B2) the support of P_X is the whole \mathbb{X} space, i.e. $\text{supp } P_X = \mathbb{X}$,

(B3) P_X is absolutely continuous for the Lebesgue measure,

(B4) $K : \mathbb{X} \rightarrow \mathbb{R}$ is a regular kernel.

The kernel estimates for all $i = 0, \dots, m-1$ are defined as in (3.8) with bandwidth h_n

$$f_{\theta,n}^{(i)}(x) = \frac{\sum_{j=1}^n Y_{i,j} K\left(\frac{x-X_j}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)} \mathbb{I}\left(\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right) \neq 0\right). \quad (5.61)$$

The reference variables are defined as the L_2 errors of these estimates, that is for all $i = 0, \dots, m-1$

$$Z_n^{(i)}(\theta) \doteq \|f_\theta - f_{\theta,n}^{(i)}\|_2. \quad (5.62)$$

These are well-defined, because f_θ and $f_{\theta,n}^{(i)}$ are both measurable and bounded. The ranking function and the confidence region are defined similarly as in Algorithm I, see (5.25) and (5.24), that is

$$\Theta_{\varrho,n}^{(2)} \doteq \left\{ \theta \in \Theta : \mathcal{R}_n(\theta) \leq q \right\}. \quad (5.63)$$

Theorem 5.5.1. *Assume that A0, A1, A2, B1, B2, B3 and B4 hold. Then*

$$\mathbb{P}\left(\theta^* \in \Theta_{\varrho,n}^{(2)}\right) = \frac{q}{m}, \quad (5.64)$$

for all sample size n . In addition, if $h_n \rightarrow 0$ and $nh_n^d \rightarrow \infty$ as $n \rightarrow \infty$, and $q < m$, then Algorithm II is strongly consistent (5.13).

Proof. From Theorem 5.2.1 the exact coverage probability follows.

For strong consistency we use that kernel estimates are strongly consistent under the conditions of Theorem 5.5.1, see Theorem 3.5.2.

Fix $\theta \neq \theta_*$. Because of the injectivity of the parameterization there exists a measurable set $C \subseteq \mathbb{R}^d$ with nonzero measure, such that $f_\theta(x) \neq f_*(x)$ for all $x \in C$, hence $\kappa \doteq \|f_* - f_\theta\|_2 > 0$. Since kernel estimates are strongly consistent

$$\|f_{\theta,n}^{(0)} - f_*\|_P \xrightarrow{a.s.} 0 \quad \text{as } n \rightarrow \infty, \quad (5.65)$$

$$\|f_{\theta,n}^{(i)} - f_\theta\|_P \xrightarrow{a.s.} 0 \quad \text{as } n \rightarrow \infty \quad (5.66)$$

for all $i = 0, \dots, m-1$. Conditions B2 and B3 imply that almost sure convergence occurs when we use the $\|\cdot\|_2$ instead of the $\|\cdot\|_P$ norm. Then

$$Z_n^{(i)}(\theta) \xrightarrow{a.s.} 0 \quad \text{as } n \rightarrow \infty \quad (5.67)$$

for all $i = 1, \dots, m-1$. Furthermore, applying the triangle inequality and the reversed triangle inequality yields

$$Z_n^{(0)}(\theta) = \|f_\theta - f_{\theta,n}^{(0)}\|_2 \leq \|f_\theta - f_*\|_2 + \|f_* - f_{\theta,n}^{(0)}\|_2 \xrightarrow{a.s.} \kappa \quad (5.68)$$

$$Z_n^{(0)}(\theta) = \|f_\theta - f_{\theta,n}^{(0)}\|_2 \geq \left| \|f_\theta - f_*\|_2 - \|f_* - f_{\theta,n}^{(0)}\|_2 \right| \xrightarrow{a.s.} \kappa \quad (5.69)$$

as $n \rightarrow \infty$, therefore $Z_n^{(0)}(\theta) \xrightarrow{a.s.} \kappa$.

Let Ω_1 be the 1 probability event where the convergences of $Z_n^{(i)}(\theta)$ for $i = 0, \dots, m-1$ occur. Let $\varepsilon = \kappa/3$. For all $\omega \in \Omega_1$ there exists $n_0(\omega)$ such that for all $n \geq n_0(\omega)$ for all $i = 1, \dots, m-1$: $Z_n^{(i)}(\theta) < \kappa/3$ and $Z_n^{(0)}(\theta) > 2/3\kappa$, thus $\theta \notin \Theta_{\varepsilon,n}^{(2)}(\omega)$ for all $n > n_0$ implying that $\omega \notin \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} \{\theta \in \Theta_{\varepsilon,n}^{(2)}\}$, therefore

$$\mathbb{P} \left(\bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} \{\theta \in \Theta_{\varepsilon,n}^{(2)}\} \right) = 0. \quad (5.70)$$

The theorem is proved. \square

It is clear that $\{f_{\theta,n}^{(i)}\}_{i=0}^{m-1}$ can be computed from the sample and they are piece-wise constant. The distance $\|f_{\theta,n}^{(i)} - f_{\theta,n}^{(j)}\|_2$ can also be calculated from the available data. Nevertheless, one may use the Monte Carlo approximation

$$\|f_{\theta,n}^{(i)} - f_{\theta,n}^{(j)}\|_2 \approx \sqrt{\frac{1}{\ell_n} \sum_{k=1}^{\ell_n} \left(f_{\theta,n}^{(i)}(\bar{X}_k) - f_{\theta,n}^{(j)}(\bar{X}_k) \right)^2}, \quad (5.71)$$

where ℓ_n is a constant and $\{\bar{X}_k\}$ are i.i.d. random variables having uniform distribution on \mathbb{X} . Note that we know from the *strong law of large numbers* (SLLN) that the square root of the sum in (5.71) almost surely converges to $\|f_{\theta,n}^{(i)} - f_{\theta,n}^{(j)}\|_2$, as $\ell_n \rightarrow \infty$. It is relatively easy to see that using the approximation in (5.71), instead of (5.62), does not affect the *exact* coverage probability of the algorithm. Moreover, if $\ell_n \rightarrow \infty$ as $n \rightarrow \infty$, then one can also show the strong consistency of the Monte Carlo approximated variant. Hence, the theoretical properties of Theorem 5.3.3 remain valid even under (5.71), but the sizes of regions are of course affected by the approximation.

The regular kernel for the local averaging estimator, which is in the core of Algorithm II can be chosen arbitrarily. The Gaussian kernel is applied in the examples in the end of this chapter. Notice that kNN estimators also can be used, that are in fact can be seen as kernel estimators using a variable bandwidth rectangular window. Therefore a natural approach is to apply the

kNN rule for local averaging, thus we can redefine functions $\{f_{\theta,n}^{(i)}\}$ for $i = 0, \dots, m-1$ as

$$f_{\theta,n}^{(i)}(x) \doteq \frac{1}{k_n} \sum_{j=1}^n Y_{i,j}(\theta) \mathbb{I}(X_j \in N(x, n_k)), \quad (5.72)$$

where $N(x, n_k)$ denotes the k_n closest neighbors of x from $\{X_j\}_{j=1}^n$ in a given metric, and $k_n \leq n$ is a constant (window size), which can depend on n . This approach also leads to alternative confidence region constructions and typically also builds confidence regions with *exact* coverage probabilities. Moreover, as kNN estimates are strongly consistent when $k_n \rightarrow \infty$ and $k_n/n \rightarrow 0$ as $n \rightarrow \infty$, see Theorem 3.6.1, the modified Algorithm II inherits these properties, the resulting confidence sets are also *strongly consistent*. The corresponding coverage and consistency theorems could be proved analogously to Theorem 5.5.1.

In general arbitrary estimation technique can be applied in (5.61). These local averaging estimates are usually computationally light and in several cases have universal guarantees for consistency, therefore when our model class is too complex to find an ERM estimate or we do not have prior knowledge about the structure of the regression function the use of these nonparametric estimates is preferable. On the other hand when the structure of \mathcal{F} is known parametric techniques such as least squares estimates can perform better.

5.6 Algorithm III (Embedding Based)

The core idea of Algorithm III is to embed the distribution of the original sample and that of the alternative ones in a RKHS using a characteristic kernel. If the underlying distributions are different, then the original dataset results in a different element than the one the alternative datasets are being mapped to, which can be detected statistically.

Let $\mathbb{S} \doteq \mathbb{X} \times \{+1, -1\}$ be the sample space. Assume that

(C1) \mathcal{H} is a *separable* RKHS containing $\mathbb{S} \rightarrow \mathbb{R}$ type functions,

(C2) the reproducing kernel function, $k(\cdot, \cdot)$, corresponding to \mathcal{H} is *characteristic* and bounded.

If $\mathbb{X} = \mathbb{R}^d$, then $\mathbb{S} = \mathbb{R}^d \times \{+1, -1\}$, and we can use, for example, the Gaussian or the Laplacian kernel, both which are characteristic [17].

Let us introduce the following kernel mean embeddings

$$h_*(\cdot) \doteq \mathbb{E}[k(\cdot, S_*)], \quad (5.73)$$

$$h_\theta(\cdot) \doteq \mathbb{E}[k(\cdot, S_\theta)], \quad (5.74)$$

where S_* and S_θ are random elements from \mathbb{S} . Variable S_* has the “true” distribution of the observations, while S_θ has a distribution where the output, Y , is generated according to the conditional probability (5.4), parameterized by θ , while the marginal distribution of the input, P_X , remains the same.

Since the kernel is bounded, $\mathbb{E}[\sqrt{k(S_\theta, S_\theta)}] < \infty$ for all θ , which ensures that h_θ exists and belongs to \mathcal{H} for all $\theta \in \Theta$.

We know that $h_\theta = h_*$ if and only if $\theta = \theta^*$ because the kernel is characteristic. Now, let us introduce the following empirical versions of the embedded distributions,

$$h_{\theta,n}^{(i)}(\cdot) \doteq \frac{1}{n} \sum_{j=1}^n k(\cdot, s_{i,j}(\theta)), \quad (5.75)$$

for $i = 0, \dots, m-1$, where $s_{i,j}(\theta) \doteq (X_j, Y_{i,j}(\theta))$; and recall that for $i = 0$, we have $Y_{i,j}(\theta) = Y_j$. In other words, $s_{i,j}(\theta)$ has the same distribution as S_θ for $i \neq 0$ and its distribution is the same as that of S_* for $i = 0$.

Let B_k be a constant that satisfies $|k(s_1, s_2)| \leq B_k$ for all $s_1, s_2 \in \mathbb{S}$. Then, obviously $|h_\theta(s)| \leq B_k$ for all $s \in \mathbb{S}$ as well. Applying the reproducing property yields the bound

$$\begin{aligned} D^2(k(\cdot, S)) &= \mathbb{E}[\|k(\cdot, S) - h(\cdot)\|_{\mathcal{H}}^2] \leq \mathbb{E}[\|k(\cdot, S)\|_{\mathcal{H}}^2] + \mathbb{E}[\|h(\cdot)\|_{\mathcal{H}}^2] \\ &\quad + 2\mathbb{E}[\langle k(\cdot, S), h(\cdot) \rangle_{\mathcal{H}}] \leq \mathbb{E}[\|k(\cdot, S)\|_{\mathcal{H}}^2] + \|h(\cdot)\|_{\mathcal{H}}^2 + 2\mathbb{E}[|h(S)|] \\ &\leq \mathbb{E}[\langle k(\cdot, S), k(\cdot, S) \rangle_{\mathcal{H}}] + \|h(\cdot)\|_{\mathcal{H}}^2 + 2B_k \\ &= \mathbb{E}[k(S, S)] + \|h(\cdot)\|_{\mathcal{H}}^2 + 2B_k \leq 3B_k + \|h(\cdot)\|_{\mathcal{H}}^2 < \infty, \end{aligned} \quad (5.76)$$

where S is either S_* or S_θ , and $h \doteq \mathbb{E}[k(\cdot, S)]$.

Then, we know from the SLLN for Hilbert space valued elements (Theorem 4.5.1) that $\|h_{\theta,n}^{(i)} - h_\theta\|_{\mathcal{H}} \rightarrow 0$ (a.s.), as $n \rightarrow \infty$, for $i \neq 0$, additionally, $\|h_{\theta,n}^{(0)} - h_*\|_{\mathcal{H}} \rightarrow 0$ (a.s.), as $n \rightarrow \infty$.

Now, we can define the reference variables $\{Z_n^{(i)}(\theta)\}_{i=0}^{m-1}$ similarly to (5.62) as

$$Z_n^{(i)}(\theta) \doteq \sum_{j=0}^{m-1} \|h_{\theta,n}^{(i)} - h_{\theta,n}^{(j)}\|_{\mathcal{H}}^2, \quad (5.77)$$

which is the total cumulative squared distance of $h_{\theta,n}^{(i)}$ from all other embedded estimates, and construct the confidence set as in (5.25).

Theorem 5.6.1. *Assume that A0, A1, A2, C1 and C2 hold, then the following is true for the confidence regions constructed by Algorithm III:*

$$\mathbb{P}(\theta^* \in \Theta_{\varrho,n}^{(3)}) = \frac{q}{m} \quad (5.78)$$

for all sample size n and hyperparameter $\varrho = (q, m)$. Moreover when $q < m$ and $3 \leq m$ hold, Algorithm III is universally strongly consistent.

Proof. The exact confidence level again follows from Theorem 5.2.1 by noting that the ranking function satisfies P1 and P2.

The proof of consistency follows the ideas of the proof of Theorem 5.5.1. Namely, let us fix a false parameter $\theta \in \Theta$ with $\theta \neq \theta^*$. Since the parameterization is injective, we know that \mathcal{D}_0 and $\{\mathcal{D}_i(\theta)\}_{i \neq 0}$ have different distributions. As the kernel is characteristic, we know that

the RKHS embedded distributions $h_*(\cdot)$ and $h_\theta(\cdot)$ are different. We then apply the SLLN for Hilbert space valued elements, Theorem 4.5.1, and use the construction of the $\{Z_n^{(i)}\}$ variables to get the limits

$$Z_n^{(i)}(\theta) \rightarrow \kappa \quad \text{as } n \rightarrow \infty, \quad (5.79)$$

$$Z_n^{(0)}(\theta) \rightarrow (m-1)\kappa \quad \text{as } n \rightarrow \infty, \quad (5.80)$$

for $i \neq 0$, almost surely, where $\kappa \doteq \|h_* - h_\theta\|_{\mathcal{H}} > 0$. Thus, $Z_n^{(0)}(\theta)$ again tends to take rank m (a.s.), as $n \rightarrow \infty$, which leads to the (a.s.) asymptotic exclusion of the false parameter $\theta \neq \theta^*$ (for more details, see the proof of Theorem 5.5.1). \square

The squared distance of the empirical versions of the embeddings $\|h_{\theta,n}^{(i)} - h_{\theta,n}^{(j)}\|_{\mathcal{H}}^2$ can be computed by applying the reproducing property of the kernel and the Gram matrix of sample $s_{i,1}(\theta), \dots, s_{i,n}(\theta), s_{j,1}(\theta), \dots, s_{j,n}(\theta)$.

Algorithm III has a nice theoretical interpretation as comparing embedded distributions in a RKHS. However, as the Gram matrices, which are required to compute variables $\{Z_n^{(i)}(\theta)\}$, depend on θ , this method has a large computational burden, hence the importance of Algorithm III is mainly theoretical. Nevertheless, motivated by its ideas, in the next section we suggest a computationally much lighter algorithm.

5.7 Algorithm IV (Discrepancy Based)

Algorithm IV follows the intuitions behind Algorithm III, but ensures that we can work with the same Gram matrix for all θ . Moreover, it has a simpler construction for $\{Z_n^{(i)}(\theta)\}$, which also makes it computationally more appealing.

For Algorithm IV we assume that

(D1) (\mathbb{X}, d) is a *compact, polish* metric space, i.e. complete and separable,

(D2) all $f \in \mathcal{F}$ are *continuous*,

(D3) \mathcal{H} is a *separable* RKHS containing $\mathbb{X} \rightarrow \mathbb{R}$ type functions,

(D4) \mathcal{H} is endowed with a measurable, *bounded* and *universal* kernel.

Let us introduce the notation $\varepsilon_{i,j}(\theta) \doteq Y_{i,j}(\theta) - f_\theta(X_j)$, for $i = 0, \dots, m-1$ and $j = 1, \dots, n$. Note that if $i \neq 0$, $\varepsilon_{i,j}(\theta)$ has zero mean for all $j \in [n]$, as $f_\theta(X_j) = \mathbb{E}_\theta[Y_{i,j}(\theta) | X_j]$.

The fundamental objects of Algorithm IV are the following variables

$$Z_n^{(i)}(\theta) \doteq \left\| \frac{1}{n} \sum_{j=1}^n \varepsilon_{i,j}(\theta) k(\cdot, X_j) \right\|_{\mathcal{H}}^2 \quad (5.81)$$

for $i = 0, \dots, m-1$. Observe that $Z_n^{(i)}(\theta)$ can be easily computed using the Gram matrix $K_{i,j} \doteq k(X_i, X_j)$, as

$$Z_n^{(i)}(\theta) = \frac{1}{n^2} \varepsilon_i^T(\theta) K \varepsilon_i(\theta), \quad (5.82)$$

applying the vector notation $\varepsilon_i(\theta) \doteq (\varepsilon_{i,1}(\theta), \dots, \varepsilon_{i,n}(\theta))^T$.

From this point, we follow the construction of Algorithms I, II and III. Namely, we define the ranking function as in (5.24), and the confidence region as in (5.63), but we apply our new reference variables, (5.81), for the definition of the ranking function.

Theorem 5.7.1. *Assume that A0, A1, A2, D1, D2 and D3 hold. The confidence regions of Algorithm III have*

$$\mathbb{P}(\theta^* \in \Theta_{\theta,n}^{(4)}) = q/m \quad (5.83)$$

for any sample size n ; and for $q < m$ they are strongly consistent.

Proof. The exact confidence follows from Theorem 5.2.1

For the proof of strong consistency, let us fix $\theta \neq \theta^*$ and an $i \neq 0$. To simplify the notations, introduce $e_j \doteq \varepsilon_{i,j}(\theta)$ and $\bar{Y}_j \doteq Y_{i,j}(\theta)$. We first show that $e_j k(\cdot, X_j)$ has zero mean

$$\begin{aligned} \mathbb{E}[e_j k(x, X_j)] &= \mathbb{E}[\mathbb{E}[e_j k(x, X_j) | X_j]] = \mathbb{E}[\mathbb{E}[(\bar{Y}_j - f_\theta(X_j))k(x, X_j) | X_j]] \\ &= \mathbb{E}[\mathbb{E}[\bar{Y}_j | X_j]k(x, X_j) - f_\theta(X_j)k(x, X_j)] = \mathbb{E}[(f_\theta(X_j) - f_\theta(X_j))k(x, X_j)] = 0. \end{aligned} \quad (5.84)$$

About the variance of $e_j k(\cdot, X_j)$, observe that

$$D^2(e_j k(\cdot, X_j)) = \mathbb{E}[\|e_j k(\cdot, X_j)\|_{\mathcal{H}}^2] \leq 4B_k, \quad (5.85)$$

where $|k(x_1, x_2)| \leq B_k$, for any x_1 and $x_2 \in \mathbb{X}$ since the kernel is bounded; also note that $\|k(\cdot, x)\|_{\mathcal{H}}^2 = k(x, x)$, for any $x \in \mathbb{X}$, because of the reproducing property of the kernel.

Therefore, we can apply the Hilbert space valued SLLN to conclude that $Z_n^{(i)}(\theta) \rightarrow 0$ (a.s.), as $n \rightarrow \infty$, for all $i \neq 0$.

Now, let $e_j^* \doteq \varepsilon_{0,j}(\theta) = Y_j - f_\theta(X_j)$. We will prove that the mean of $e_j^* k(\cdot, X_j)$ is not the zero function. We can again show that

$$\mathbb{E}[e_j^* k(\cdot, X_j)] = \mathbb{E}[(f_*(X_j) - f_\theta(X_j))k(\cdot, X_j)], \quad (5.86)$$

using similar steps as in (5.84), except in the last one, where in our case we have $\mathbb{E}[Y_j | X_j] = f_*(X_j)$. We will argue that the term $\mathbb{E}[(f_*(X_j) - f_\theta(X_j))k(\cdot, X_j)]$ cannot be zero.

Let us introduce $f_0 \doteq f_* - f_\theta$, and assume by contradiction that $\mathbb{E}[f_0(X_j)k(\cdot, X_j)]$ is the zero function. Then, for all $x \in \mathbb{X}$, $\langle f_0, k(x, \cdot) \rangle_P \doteq \mathbb{E}[f_0(X_j)k(x, X_j)] = 0$ (note that a RKHS is a space of functions and not that of equivalence classes of functions). Since the kernel is universal, \mathbb{X} is compact, and f_0 is continuous, we know that for all $\varepsilon > 0$, there exists an

$$\hat{f}(\cdot) = \sum_{k=1}^N \alpha_k k(\cdot, \bar{x}_k), \quad (5.87)$$

with some points $\{\bar{x}_k\}_{k=1}^N$ and coefficients $\{\alpha_k\}_{k=1}^N$, such that $\|\hat{f} - f_0\|_\infty < \varepsilon$. Then, clearly

$$\int_{\mathbb{X}} (\hat{f} - f_0)^2 dP_X(x) \leq \int_{\mathbb{X}} \|\hat{f} - f_0\|_\infty^2 dP_X(x) < \int_{\mathbb{X}} \varepsilon^2 dP_X(x) = \varepsilon^2, \quad (5.88)$$

since P_X is a probability measure on \mathbb{X} . Hence, for all $\varepsilon > 0$,

$$\|\hat{f} - f_0\|_P^2 = \|\hat{f}\|_P^2 + \|f_0\|_P^2 - 2\langle f_0, \hat{f} \rangle_P < \varepsilon^2. \quad (5.89)$$

For all $x \in \mathbb{X}$, we know that $\langle f_0, k(x, \cdot) \rangle_P = 0$, thus

$$\begin{aligned} \langle f_0, \hat{f} \rangle_P &= \int_{\mathbb{X}} \sum_{k=1}^N \alpha_k k(x, \bar{x}_k) f_0(x) dP_X(x) \\ &= \sum_{k=1}^N \alpha_k \int_{\mathbb{X}} k(x, \bar{x}_k) f_0(x) dP_X(x) = \sum_{k=1}^N \alpha_k \langle f_0, k(\bar{x}_k, \cdot) \rangle_P = 0. \end{aligned} \quad (5.90)$$

Then, combining (5.89) and (5.90) yields that for all $\varepsilon > 0$ the following holds

$$\|f_0\|_P^2 \leq \|\hat{f}\|_P^2 + \|f_0\|_P^2 < \varepsilon^2, \quad (5.91)$$

which implies that $\|f_0\|_P^2 = 0$. On the other hand, we know from (5.2) that this norm cannot be zero if $\theta \neq \theta^*$. Therefore, we have reached a contradiction, hence $\mathbb{E}[(f_*(X_j) - f_\theta(X_j))k(\cdot, X_j)]$ cannot be the zero element of the RKHS.

We can use a similar argument to (5.76) to show that $D^2(e_j^* k(\cdot, X_j))$ is bounded, also using that $\{e_j^*\}$ are bounded. Then, applying the Hilbert space variant of SLLN, Theorem 4.5.1,

$$\frac{1}{n} \sum_{j=1}^n e_j^* k(\cdot, X_j) \xrightarrow{a.s.} h_0 \neq 0, \quad \text{as } n \rightarrow \infty. \quad (5.92)$$

Therefore, summarizing our results, we have

$$Z_n^{(0)}(\theta) \rightarrow \|h_0\|_{\mathcal{H}}^2 \quad \text{as } n \rightarrow \infty, \quad (5.93)$$

$$Z_n^{(i)}(\theta) \rightarrow 0 \quad \text{as } n \rightarrow \infty, \quad (5.94)$$

for $i \neq 0$, almost surely, where $\|h_0\|_{\mathcal{H}}^2 > 0$. Thus, $Z_n^{(0)}(\theta)$ again tends to take rank m (a.s.), as $n \rightarrow \infty$, which leads to the (a.s.) asymptotic exclusion of the parameter $\theta \neq \theta^*$. \square

5.8 Numerical Experiments II

To illustrate Algorithm II-IV we carried out similar numerical experiments as for Algorithm I. Since the regression functions are not required to be linearly parameterized in these cases we considered a more general setup. In these presented tests the joint probability distribution of the data was assumed to be the mixture of two Laplace distributions with different location parameters, μ_1, μ_2 , but with the same scale parameter λ . It was assumed that with probability p we observe the “+1” class, and with probability $1 - p$ we see an element of the “−1” class. Thus selecting p, μ_1, μ_2 and λ induces a regression function, see (5.3).

During the experiments the confidence regions were built for parameters p and λ , while the location parameters were fixed, $\mu_1 = 1$ and $\mu_2 = -1$, to allow two dimensional figures. Figure 5.2 demonstrates the obtained ranks $\{\mathcal{R}_n(\theta)\}$ for various $\theta = (p, \lambda)$ using Algorithm II with a Gaussian kernel (a), Algorithm II with a kNN approach (b), Algorithm III with a Gaussian

kernel (c) and Algorithm IV with a Gaussian kernel. The corresponding confidence regions in the model spaces can be seen on figures (e), (f), (g) and (h). Darker colors indicate smaller ranks, that is, the darker the color is, the more likely the parameter and the corresponding model are included in a confidence region. The true parameters were $p = 1/2$ (x -axis) and $\lambda = 1$ (y -axis). The sample size was $n = 500$ and $m = 40$ (original and alternative) samples were generated. For the Gaussian kernel we chose parameter $\sigma = 1/2$. The regions were evaluated on a fine grid in the parameter space and the corresponding models were indicated in the model space. It can be seen that the constructed confidence regions are comparable in size and shape.

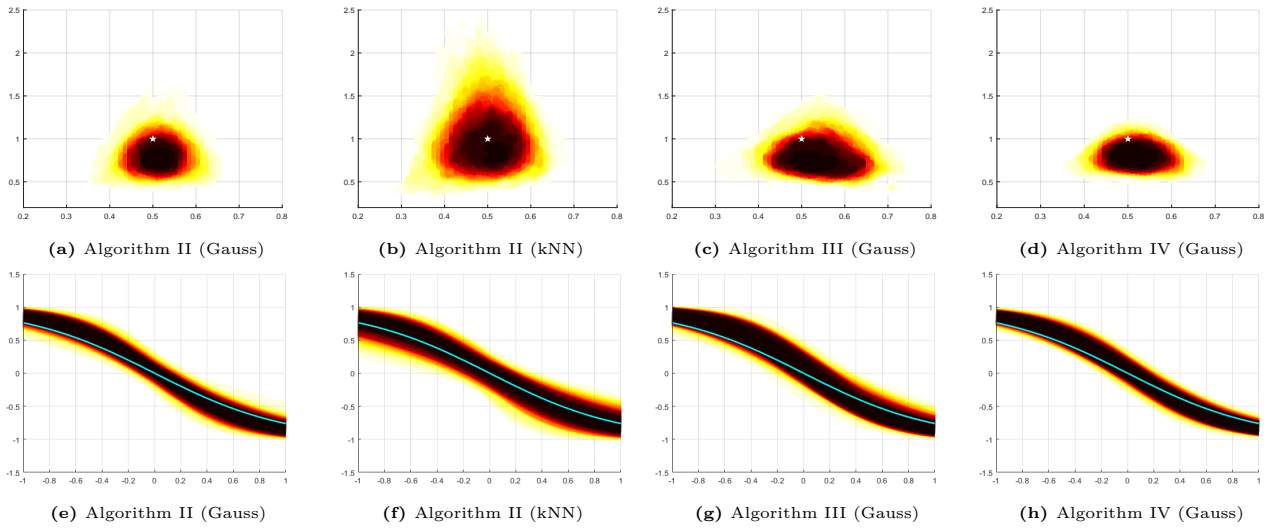


Figure 5.2: The ranks of the reference variables are indicated with the color for the parameters in figures (a), (b), (c) and (d) and for the corresponding models in figures (e), (f), (g) and (h), that is darker points and models indicate reference variables with lower ranks. The true parameter was $\theta_* = (p_*, \lambda_*)$, where $p_* = \frac{1}{2}$ (x -axis) and $\lambda_* = 1$ (y -axis).

Note that in this special example it is possible to construct individual confidence regions for parameters p and λ based on standard results. One can use, for example, Hoeffding’s inequality, see Theorem [A.3.1](#), to derive confidence bounds for probability p , and λ can be estimated based on the remark that the variance of the inputs of the observations, for both classes, is $2\lambda^2$. Nevertheless, such approaches require the specific interpretations of the parameters, on how they influence the observations. Furthermore, even in this very special case it is not obvious how to construct a joint confidence region for the pair (p, λ) . Simply intersecting the two confidence tubes (i.e., if we extend the confidence intervals for p and λ to \mathbb{R}^2 , then they define two infinite “stripes”, a vertical and a horizontal one) produces a rectangle set with a lower confidence than those of the original sets, and leads to conservative confidence regions.

On the other hand, the Algorithm II-IV do not presuppose any interpretation of the tested parameters, apart from the fact that they determine a regression function. They do not need a fully parameterized joint distribution, indeed, the regression function is compatible with infinitely many joint distributions having widely different (marginal) input distributions. In addition, if $\theta \in \mathbb{R}^d$, then the algorithms automatically construct joint and non-conservative, exact confidence sets. Hence, another advantage of the presented framework, apart from its strong theoretical guarantees, is its flexibility.

Conclusion

In this thesis a general resampling technique was introduced for uncertainty quantification for the problem of classification and the theoretical background of the applied tools was presented. The new results provide us strong, non-conservative stochastic guarantees for finding the estimated model, which are very important in several scientific and industrial applications.

The first part of the thesis contains a general introduction to statistical learning theory, where a wide range of the most important concepts and algorithms are presented focusing especially on those materials that are applied in our new results.

The main findings of this thesis are in Chapter 5, where a general framework is introduced for inferring non-asymptotic stochastic guarantees in the form of confidence regions for the problem of classification. The main idea was to test candidate regression function models by generating alternative samples based on them, and then computing the performance of an estimator on all samples. One of our main observations was that if the candidate model is wrong, then our algorithms behave differently on the alternatively generated samples than on the original one, which can be detected statistically by ranking. Four algorithms were introduced, all of which constructs exact confidence region for the regression function under mild statistical assumptions for any user-chosen rational probability level. The first method uses empirical risk minimizer estimates, as it compares least squares solutions to the examined parameters. The second algorithm is based on nonparametric local averaging kernel estimates. The third and fourth constructions are built on the theory of reproducing kernel Hilbert spaces and kernel mean embeddings. It is proved that the presented algorithms have good asymptotic behaviors, as Algorithm I is strongly uniformly consistent and Algorithm II-IV are all universally strongly consistent under general conditions. Furthermore, the proposed framework is semi-parametric, because the regression function does not determine the joint probability distribution of the data, as it does not affect the distribution of the inputs, and that is why only the labels are resampled. Note that the construction scheme is not restricted to specific model classes of regression functions. Any such model can be tested as long as it determines a proper conditional distribution with respect to the inputs.

In conclusion, the presented non-asymptotic, distribution-free methods can provide auspicious alternatives to those uncertainty evaluating procedures that are based on strong distributional assumptions or on asymptotic results.

Appendix A

Tail and Concentration Inequalities

In learning theory non-asymptotic guarantees are always preferable to asymptotic results, because in real-world problems datasets always have a finite size. For this reason it is our interest to obtain high probability statements for fixed sample size n and dimension d . We used such bounds in the thesis, nevertheless they are also important on their own. In several cases, we are interested in finding tail bounds for a random variable or to quantify how close a random variable is to its mean in absolute value. In Appendix [A](#) we present the most important elementary techniques for deriving high probability bounds for both concentration and tail deviation based on the book of Wainwright [\[25\]](#) and the book of Györfi et al. [\[11\]](#).

We are on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, where Ω is the set of all possible outcomes, \mathcal{A} is the σ -algebra of the events and \mathbb{P} is a probability measure.

A.1 Deriving the Chernoff Bound

Our starting point is Markov's inequality.

Claim A.1.1. (*Markov's inequality*) *Let X be a nonnegative variable. If $\mathbb{E}X < \infty$ then for all $t > 0$ we have*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}X}{t}. \quad (\text{A.1})$$

When X has a finite second moment, then applying Markov's inequality to variable $|X - \mathbb{E}X|^2$ yields Chebyshev's inequality

$$\mathbb{P}(|X - \mathbb{E}X| > t) \leq \frac{D^2(X)}{t^2}, \quad (\text{A.2})$$

where $D^2(X)$ is the variance of X . It can be showed that both Markov's inequality and Chebyshev's inequality are sharp. Generalizing this idea we can say that for a variable X with $\mathbb{E}X^k < \infty$ we have

$$\mathbb{P}(|X - \mathbb{E}X| > t) \leq \frac{\mathbb{E}(|X - \mathbb{E}X|^k)}{t^k}. \quad (\text{A.3})$$

Similarly we can derive bounds for different functions than polynomials. Assume that X has a generating function which is finite at least in some neighborhood of zero, i.e. there is a $\delta > 0$ such that $\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] < \infty$ for all $\lambda \in [0, \delta)$. Let $\mu = \mathbb{E}X$. Then for all $\lambda \in [0, \delta)$ we obtain that

$$\mathbb{P}(X - \mu \geq t) = \mathbb{P}(e^{\lambda(X-\mu)} \geq e^{\lambda t}) \leq \frac{\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}]}{e^{\lambda t}}. \quad (\text{A.4})$$

Hence, the tail bound depends on the growth of the moment generating function. We can minimize the quantity in the right hand side to obtain the Chernoff bound as

$$\mathbb{P}(X - \mu \geq t) \leq \inf_{\lambda \in [0, \delta)} \frac{\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}]}{e^{\lambda t}}. \quad (\text{A.5})$$

It can be showed that a proper choice of k for the previous method always gives at least as good result as the Chernoff bound. Still Chernoff bound is most widely applied because of the well-known techniques for manipulating moment generating functions.

A.2 Sub-Gaussian Variables

As an example we derive the Chernoff bound for X when it has a Gaussian distribution with expected value μ and variance σ^2 . The moment generating function of a Gaussian variable is well-known, that is for all λ

$$\mathbb{E}[e^{\lambda(X-\mu)}] = e^{\sigma^2 \lambda^2 / 2}. \quad (\text{A.6})$$

For the Chernoff bound we need to minimize the following quantity

$$\frac{\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}]}{e^{\lambda t}} = \frac{e^{\sigma^2 \lambda^2 / 2}}{e^{\lambda t}} = e^{\sigma^2 \lambda^2 / 2 - \lambda t}. \quad (\text{A.7})$$

Since the exponential function is strictly monotone increasing we need to minimize function $f(\lambda) = \sigma^2 \lambda^2 / 2 - \lambda t$ in variable λ . It is easy to see that a minimum is achieved when $\lambda_* = \frac{t}{\sigma^2}$ and $f(\lambda_*) = -\frac{t^2}{2\sigma^2}$. Therefore, the Chernoff bound for a Gaussian X with expected value μ and variance σ^2 is

$$\mathbb{P}(X - \mu \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right). \quad (\text{A.8})$$

The two sided version can be proved easily using that this inequality holds for $-X$ with expected value $-\mu$ implying that the following holds for all $t \geq 0$

$$\mathbb{P}(|X - \mu| \geq t) \leq 2 \exp\left(-\frac{t^2}{2\sigma^2}\right). \quad (\text{A.9})$$

Motivated by this example the notion of sub-Gaussianity can be introduced.

Definition A.2.1. (*sub-Gaussian variable*) Let X be a variable such that $\mu = \mathbb{E}X$. We say

that it is sub-Gaussian with a positive parameter σ if the following holds

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\sigma^2 \lambda^2 / 2} \quad \forall \lambda \in \mathbb{R}. \quad (\text{A.10})$$

Notice that when X is a Gaussian variable with parameters (μ, σ^2) then it is also sub-Gaussian with parameter σ . Besides, there are a large number of sub-Gaussian variables.

Claim A.2.1. *Let X be a bounded variable taking value in $[a, b]$. It is sub-Gaussian with parameter $\sigma = (b - a)$.*

The presented proof is from [25].

Proof. Let X be a variable bounded between numbers a and b . Let X' be an identically distributed variable independent of X .

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] = \mathbb{E}[e^{\lambda(X-\mathbb{E}X')}] \leq \mathbb{E}_{X, X'}[e^{\lambda(X-X')}], \quad (\text{A.11})$$

because of Jensen's inequality. We use a symmetrization principle. Let ξ be a random sign (Rademacher variable) such that $\mathbb{P}(\xi = +1) = \mathbb{P}(\xi = -1) = 1/2$. Notice that $\xi(X - X')$ has the same distribution as $X - X'$, since X and X' are independent identically distributed (i.i.d.). Therefore the following holds

$$\mathbb{E}_{X, X'}[e^{\lambda(X-X')}] = \mathbb{E}_{X, X', \xi}[e^{\lambda \xi(X-X')}] = \mathbb{E}[\mathbb{E}[e^{\lambda \xi(X-X')} | X, X']]. \quad (\text{A.12})$$

First we compute the conditional expected value

$$\mathbb{E}[e^{\lambda \xi(X-X')} | X, X'] = \mathbb{E}[e^{\lambda \xi(x-x')}]_{x=X, x'=X'}. \quad (\text{A.13})$$

For fixed $\gamma = \lambda(x - x')$ we find that

$$\mathbb{E}[e^{\gamma \xi}] = \frac{1}{2}e^{-\gamma} + \frac{1}{2}e^{\gamma} = \frac{1}{2} \sum_{k=0}^{\infty} \frac{(-\gamma)^k}{k!} + \frac{1}{2} \sum_{k=0}^{\infty} \frac{\gamma^k}{k!} \quad (\text{A.14})$$

$$= 1 + \frac{1}{2} \sum_{k=1}^{\infty} \frac{(-\gamma)^k + \gamma^k}{k!} = 1 + \frac{1}{2} \sum_{k=1}^{\infty} \frac{\gamma^{2k}}{(2k)!} \leq \sum_{k=0}^{\infty} \frac{(\gamma^2)^k}{2^k k!} = e^{\gamma^2/2} \quad (\text{A.15})$$

Substituting it back to equation (A.13) yields

$$\mathbb{E}[e^{\lambda \xi(X-X')} | X, X'] = \exp\left(\lambda^2(X-X')^2/2\right). \quad (\text{A.16})$$

We can finish the proof by applying that $(X - X')^2 \leq (b - a)^2$, because both X and X' are in $[a, b]$ almost surely.

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] \leq \mathbb{E}\left[\exp\left(\lambda^2(X-X')^2/2\right)\right] \leq e^{\lambda^2(b-a)^2/2} \quad (\text{A.17})$$

Hence, X is sub-Gaussian with parameter $(b - a)$. \square

It can be showed with a more thorough reasoning that $\sigma = \frac{(b-a)}{2}$ is also a good sub-Gaussian parameter for bounded variables, see Exercise 2.4 in [25].

Furthermore, when X is sub-Gaussian then so is $-X$ with the same parameter. Therefore exactly as in the Gaussian case, one can show that for any sub-Gaussian variable X (A.8) and (A.9) hold.

A.3 Hoeffding's Inequality

Now we derive Hoeffding's inequality which is our main tool to prove several results related to PAC learnability, uniform convergence and consistency.

Theorem A.3.1. (*Hoeffding's inequality*) Let X_1, \dots, X_n be independent sub-Gaussian variables each with $\mathbb{E}X_i = \mu_i$ and parameter σ_i for all $i = 1, \dots, n$, then for all $t \geq 0$

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mu_i) > t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right). \quad (\text{A.18})$$

Proof. First, we are going to prove that the sum of sub-Gaussian variables is sub-Gaussian. Using the independence we can see that

$$\mathbb{E}\left[\exp\left(\lambda\left(\sum_{i=1}^n (X_i - \mu_i)\right)\right)\right] = \prod_{i=1}^n \mathbb{E}\left(\exp(\lambda(X_i - \mu_i))\right) \quad (\text{A.19})$$

$$\leq \prod_{i=1}^n \exp\left(\frac{\lambda^2 \sigma_i^2}{2}\right) \leq \exp\left(\frac{\lambda^2 \sum_{i=1}^n \sigma_i^2}{2}\right), \quad (\text{A.20})$$

i.e. $\sum_{i=1}^n X_i$ is sub-Gaussian with parameter $\sqrt{\sum_{i=1}^n \sigma_i^2}$. We argued that for sub-Gaussian variables the Chernoff bound can be used, see (A.8), from which we obtain the theorem. \square

The two sided version of this theorem can be stated similarly to (A.9).

Since bounded variables are sub-Gaussian with parameter $\sigma = \frac{(b-a)}{2}$, we can state the theorem for them as a corollary. The literature often refers to this version as the Hoeffding inequality.

Corollary A.3.1.1. (*Hoeffding's inequality for bounded variables*) Let X_1, \dots, X_n be independent variables taking values from $[a, b]$, with expected values $\mathbb{E}X_i = \mu_i$ for $i = 1, \dots, n$. Then for all $t \geq 0$ we have

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mu_i) > t\right) \leq \exp\left(-\frac{2t^2}{n(b-a)^2}\right). \quad (\text{A.21})$$

Repeating the arguments for the negated variables, $-X_1, \dots, -X_n$, yields similar bounds which can be used to formulate the theorem for the absolute deviation as

$$\mathbb{P}\left(\left|\sum_{i=1}^n (X_i - \mu_i)\right| > t\right) \leq 2 \exp\left(-\frac{2t^2}{n(b-a)^2}\right). \quad (\text{A.22})$$

An important example is when X_1, \dots, X_n are i.i.d. bounded variables and we are interested in bounding the deviation of the empirical mean from the expectation. Then applying the

theorem for variables $\frac{1}{n}X_1, \dots, \frac{1}{n}X_n$ yields

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X_1\right| > t\right) \leq 2 \exp\left(-\frac{2nt^2}{(b-a)^2}\right), \quad (\text{A.23})$$

because $\frac{1}{n}X_i \in [a/n, b/n]$ almost surely when $X_i \in [a, b]$ for $i = 1, \dots, n$.

A.4 Generalization to Martingale Differences

We can apply these tools on martingale differences. In Chapter 3 such results were used to derive universal bounds for local averaging kernel estimates.

Definition A.4.1. (*martingale*) Let $\{X_k\}_{k=0}^\infty$ be an adopted sequence of variables to the filtration $\{\mathcal{F}_k\}_{k=0}^\infty$. We call the $\{(X_k, \mathcal{F}_k)\}_{k=0}^\infty$ sequence a martingale if for all natural $k \geq 1$ we have $\mathbb{E}|X_k| < \infty$ and with probability 1

$$\mathbb{E}(X_{k+1} | \mathcal{F}_k) = X_k. \quad (\text{A.24})$$

In addition, we call $D_k \doteq X_k - X_{k-1}$ the k^{th} martingale difference and the sequence $\{(D_k, \mathcal{F}_k)\}_{k=1}^\infty$ martingale difference sequence. Notice that each D_k has zero mean.

Theorem A.4.1. (*Azuma-Hoeffding's inequality*) Let $\{(D_k, \mathcal{F}_k)\}_{k=1}^\infty$ be a martingale difference sequence for which $D_k \in [a_k, b_k]$ with probability 1 for all $k \geq 1$. Then for all $t \geq 0$

$$\mathbb{P}\left(\left|\sum_{i=1}^n D_i\right| \geq t\right) \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \quad (\text{A.25})$$

Proof. We show that $\sum_{i=1}^n D_i$ is sub-Gaussian with parameter $\sqrt{\frac{\sum_{i=1}^n (b_i - a_i)^2}{4}}$.

$$\begin{aligned} \mathbb{E}\left(e^{\lambda \sum_{i=1}^n D_i}\right) &= \mathbb{E}\left(e^{\lambda \sum_{i=1}^{n-1} D_i} \mathbb{E}(e^{\lambda D_n} | \mathcal{F}_1, \dots, \mathcal{F}_{n-1})\right) \\ &\leq \mathbb{E}\left(e^{\lambda \sum_{i=1}^{n-1} D_i}\right) e^{\lambda^2 (b_n - a_n)^2 / 8} \leq \exp\left(\frac{\lambda^2 \sum_{i=1}^n (b_i - a_i)^2}{2 \cdot 4}\right) \end{aligned} \quad (\text{A.26})$$

In the first inequality we used that $\mathbb{E}(e^{\lambda D_n} | \mathcal{F}_1, \dots, \mathcal{F}_{n-1}) \leq e^{\lambda^2 (b_n - a_n)^2 / 8}$. It holds because $D_n | \mathcal{F}_{n-1}$ is in $[a_n, b_n]$ almost surely since $D_n \in [a_n, b_n]$ with probability 1, therefore the stronger version of Claim A.2.1 can be applied. Then the two sided Hoeffding's inequality yields the theorem. \square

A neat application of this theorem results McDiarmid's inequality.

Theorem A.4.2. (*McDiarmid's inequality*) Let X_1, \dots, X_n be independent variables from a set \mathbb{A} and $f : \mathbb{A}^n \rightarrow \mathbb{R}$ be a function with the following property

$$\sup_{x_1, \dots, x_n, y \in \mathbb{A}} |f(x_1, \dots, x_n) - f(x_1, \dots, x_{i-1}, y, x_{i+1}, \dots, x_n)| \leq c_i \quad i \in [n], \quad (\text{A.27})$$

where $[n] \doteq \{1, \dots, n\}$. Then for all $t \geq 0$

$$\mathbb{P}\left(|f(X_1, \dots, X_n) - \mathbb{E}f(X_1, \dots, X_n)| \geq \varepsilon\right) \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right). \quad (\text{A.28})$$

Proof. Let $X_0 = 0$ and $X \doteq (X_1, \dots, X_n)$. It is easy to show that from (A.27) it follows that f is bounded, because let a_1, \dots, a_n be fixed each in \mathbb{A} . Then for all $x_1, \dots, x_n \in \mathbb{A}$ we have $|f(x_1, \dots, x_n)| \leq |f(a_1, \dots, a_n)| + \sum_{i=1}^n c_i$. Therefore $\mathbb{E}|f(X)| < \infty$ and we can define the martingale differences as

$$D_k = \mathbb{E}(f(X) | X_1, \dots, X_k) - \mathbb{E}(f(X) | X_1, \dots, X_{k-1}). \quad (\text{A.29})$$

We are going to prove that D_k is in an interval with at most width c_k . Let

$$A_k \doteq \inf_{x \in \mathbb{A}} \mathbb{E}(f(X) | X_1, \dots, X_{k-1}, X_k = x) - \mathbb{E}(f(X) | X_1, \dots, X_{k-1}), \quad (\text{A.30})$$

$$B_k \doteq \sup_{x \in \mathbb{A}} \mathbb{E}(f(X) | X_1, \dots, X_{k-1}, X_k = x) - \mathbb{E}(f(X) | X_1, \dots, X_{k-1}), \quad (\text{A.31})$$

then it is clear that $A_k \leq D_k \leq B_k$ almost surely. Therefore it is enough to show that $B_k - A_k \leq c_k$ with probability one to apply Theorem A.4.1. Let \mathbb{E}_{k+1} denote the expectation with respect to variables X_{k+1}, \dots, X_n . Because of the independence

$$\mathbb{E}(f(X) | X_1 = x_1, \dots, X_k = x_k) = \mathbb{E}_{k+1}(f(x_1, \dots, x_k, X_{k+1}, \dots, X_n)). \quad (\text{A.32})$$

Using the bounds on the differences yields almost surely that

$$\begin{aligned} B_k - A_k &\leq \sup_{x \in \mathbb{A}} \mathbb{E}_{k+1}(f(X_1, \dots, X_{k-1}, x, X_{k+1}, \dots, X_n)) \\ &\quad - \inf_{y \in \mathbb{A}} \mathbb{E}_{k+1}(f(X_1, \dots, X_{k-1}, y, X_{k+1}, \dots, X_n)) \\ &\leq \sup_{x, y \in \mathbb{A}} \left| \mathbb{E}_{k+1}[(f(X_1, \dots, X_{k-1}, x, X_{k+1}, \dots, X_n)) - (f(X_1, \dots, X_{k-1}, y, X_{k+1}, \dots, X_n))] \right|, \end{aligned} \quad (\text{A.33})$$

which is at most c_k . Then by Theorem A.4.1 we obtain the inequality. \square

There are other generalizations of these concepts. For sub-exponential variables Bernstein's inequality provides similar bounds. For further results see Chapter 2 in [25].

Appendix B

Proofs

B.1 A Bayes Optimal Classifier

Claim 1.1.2. *The following function, g_* , with domain \mathbb{X}*

$$g_*(x) = \text{sign}(2\eta(x) - 1) = \text{sign}\left(\mathbb{E}[Y | X = x]\right) \quad (\text{B.1})$$

is Bayes optimal in case of the 0/1 loss.

Proof. We are going to show that for an arbitrary classifier $g : \mathbb{X} \rightarrow \{+1, -1\}$ the following holds

$$\mathbb{P}(g_*(X) \neq Y) \leq \mathbb{P}(g(X) \neq Y). \quad (\text{B.2})$$

First, consider the conditional probability $\mathbb{P}(g(X) \neq Y | X = x)$:

$$\begin{aligned} \mathbb{P}(g(X) \neq Y | X = x) &= 1 - \mathbb{P}(g(X) = Y | X = x) \\ &= 1 - \left(\mathbb{P}(g(X) = 1, Y = 1 | X = x) + \mathbb{P}(g(X) = -1, Y = -1 | X = x)\right) \\ &= 1 - \left(\mathbb{I}(g(x) = 1)\mathbb{P}(Y = 1 | X = x) + \mathbb{I}(g(x) = -1)\mathbb{P}(Y = -1 | X = x)\right) \\ &= 1 - \left(\mathbb{I}(g(x) = 1)\eta(x) + \mathbb{I}(g(x) = -1)(1 - \eta(x))\right). \end{aligned} \quad (\text{B.3})$$

Furthermore, for all $x \in \mathbb{X}$ we have that

$$\begin{aligned} &\mathbb{P}(g(X) \neq Y | X = x) - \mathbb{P}(g_*(X) \neq Y | X = x) \\ &= \eta(x)\left(\mathbb{I}(g_*(x) = 1) - \mathbb{I}(g(x) = 1)\right) + (1 - \eta(x))\left(\mathbb{I}(g_*(x) = -1) - \mathbb{I}(g(x) = -1)\right) \\ &= (2\eta(x) - 1)\left(\mathbb{I}(g_*(x) = 1) - \mathbb{I}(g(x) = 1)\right) \geq 0, \end{aligned} \quad (\text{B.4})$$

where we used that $(2\eta(x) - 1)$ is nonnegative if and only if the $(\mathbb{I}(g_*(x) = 1) - \mathbb{I}(g(x) = 1))$ quantity is nonnegative. We have seen so far that for all conditions $X = x$:

$$\mathbb{P}(g_*(X) \neq Y | X = x) \leq \mathbb{P}(g(X) \neq Y | X = x). \quad (\text{B.5})$$

Integrating out both sides with respect to P_X and applying the monotonicity of the integral

yield that $\mathbb{P}(g_*(X) \neq Y) \leq \mathbb{P}(g(X) \neq Y)$.

The second half of the claim can be proved easily starting from the definition of η

$$\begin{aligned} 2\eta(x) - 1 &= 2\mathbb{P}(Y = 1|X = x) - 1 = \mathbb{P}(Y = 1|X = x) - (1 - \mathbb{P}(Y = 1|X = x)) \\ &= \mathbb{P}(Y = 1|X = x) - \mathbb{P}(Y = -1|X = x) = \mathbb{E}[Y|X = x]. \end{aligned} \quad (\text{B.6})$$

The proof of the claim is finished. \square

B.2 A Uniform Exponential Bound

Theorem 2.9.2. *Let \mathcal{F} be a set of functions $f : \mathbb{X} \rightarrow [0, B]$ and $\mathbf{X} = \{X_i\}_{i=1}^n$ an i.i.d. sample taking values from \mathbb{X} . For any $n \in \mathbb{N}$, and any $\varepsilon > 0$,*

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}f(X_i) \right| > \varepsilon\right) \leq 8 \mathbb{E}\mathcal{N}_1(\varepsilon/8, \mathcal{F}, \mathbf{X}) e^{-n\varepsilon^2/(128 B^2)}. \quad (\text{B.7})$$

Proof. The proof is very similar to the proof of Theorem 2.7.2. In fact let $\mathbf{X}' = \{X'_1, \dots, X'_n\}$ be an i.i.d. sample distributed as \mathbf{X} , but independent of \mathbf{X} . Furthermore, let \tilde{f} be a function such that

$$\left| \frac{1}{n} \sum_{i=1}^n \tilde{f}(X_i) - \mathbb{E}\tilde{f}(X_1) \right| > \varepsilon \quad (\text{B.8})$$

holds, dependent of \mathbf{X} . If such function does not exist, then let \tilde{f} be arbitrary (fixed). Assume that $n \geq \frac{2B^2}{\varepsilon^2}$, otherwise the bound is trivial.

In the first step we show that similarly to (2.40) the following holds

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}f(X_1) \right| > \varepsilon\right) \leq 2 \mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \frac{1}{n} \sum_{i=1}^n f(X'_i) \right| > \varepsilon/2\right). \quad (\text{B.9})$$

Applying Chebyshev's inequality yields

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n \tilde{f}(X_i) - \mathbb{E}(\tilde{f}(X'_1) | \mathbf{X}) \right| > \frac{\varepsilon}{2} \middle| \mathbf{X}\right) \leq \frac{D^2(\tilde{f}(X'_1) | \mathbf{X})}{n \varepsilon^2/4} \leq \frac{B^2/4}{\frac{n \varepsilon^2}{4}} \leq \frac{1}{2}, \quad (\text{B.10})$$

where in the second inequality we used that

$$D^2(\tilde{f}(X'_1) | \mathbf{X}) = D^2(\tilde{f}(X'_1) - B/2 | \mathbf{X}) \leq \mathbb{E}\left(|\tilde{f}(X'_1) - B/2|^2 \middle| \mathbf{X}\right) \leq \frac{B^2}{4}, \quad (\text{B.11})$$

because $\tilde{f}(X'_1) \in [0, B]$. Similarly to (2.41) we have

$$\begin{aligned} \mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \frac{1}{n} \sum_{i=1}^n f(X'_i) \right| > \varepsilon/2\right) &\geq \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n \tilde{f}(X_i) - \frac{1}{n} \sum_{i=1}^n \tilde{f}(X'_i) \right| > \varepsilon/2\right) \\ &\geq \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n \tilde{f}(X_i) - \mathbb{E}(\tilde{f}(X'_1) | \mathbf{X}) \right| > \varepsilon, \left| \frac{1}{n} \sum_{i=1}^n \tilde{f}(X'_i) - \mathbb{E}(\tilde{f}(X'_1) | \mathbf{X}) \right| \leq \varepsilon/2\right) \end{aligned}$$

$$\begin{aligned}
 &= \mathbb{E} \left[\mathbb{I} \left(\left| \frac{1}{n} \sum_{i=1}^n \tilde{f}(X_i) - \mathbb{E}(\tilde{f}(X'_1) | \mathbf{X}) \right| > \varepsilon \right) \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n \tilde{f}(X'_i) - \mathbb{E}(\tilde{f}(X'_1) | \mathbf{X}) \right| \leq \varepsilon/2 | \mathbf{X} \right) \right] \\
 &\leq \frac{1}{2} \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n \tilde{f}(X_i) - \mathbb{E}(\tilde{f}(X'_1) | \mathbf{X}) \right| > \varepsilon \right) = \frac{1}{2} \mathbb{P} \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}f(X'_1) \right| > \varepsilon \right)
 \end{aligned} \tag{B.12}$$

In the second step we introduce random signs exactly as before (2.7). Again, notice that $\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i(f(X_i) - f(X'_i)) \right|$ has the same distribution as $\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n (f(X_i) - f(X'_i)) \right|$, therefore

$$\begin{aligned}
 \mathbb{P} \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n (f(X_i) - f(X'_i)) \right| > \frac{\varepsilon}{2} \right) &= \mathbb{P} \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i(f(X_i) - f(X'_i)) \right| > \frac{\varepsilon}{2} \right) \\
 &\leq \mathbb{P} \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i(f(X_i)) \right| > \frac{\varepsilon}{4} \right) + \mathbb{P} \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i(f(X'_i)) \right| > \frac{\varepsilon}{4} \right) \\
 &= 2 \mathbb{P} \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i(f(X_i)) \right| > \frac{\varepsilon}{4} \right).
 \end{aligned} \tag{B.13}$$

In the third step we bound the probability $\mathbb{P} \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i(f(X_i)) \right| > \frac{\varepsilon}{4} \right)$ by conditioning on \mathbf{X} , which is equivalent to fixing \mathbf{x} . Let $\mathcal{F}_{\varepsilon/8}$ be an $\varepsilon/8$ -cover of \mathcal{F} with minimal size with respect to the $L_1(P_n)$ norm. For all $f \in \mathcal{F}$ there is an $\bar{f} \in \mathcal{F}_{\varepsilon/8}$ such that $\|f - \bar{f}\|_{L_1(P_n)} < \varepsilon/8$, therefore

$$\left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i) \right| \geq \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \bar{f}(x_i) \right| + \frac{1}{n} \sum_{i=1}^n |\sigma_i| \cdot |f(x_i) - \bar{f}(x_i)| < \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \bar{f}(x_i) \right| + \frac{\varepsilon}{8}. \tag{B.14}$$

Furthermore,

$$\begin{aligned}
 \mathbb{P} \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i(f(x_i)) \right| > \frac{\varepsilon}{4} \right) &\leq \mathbb{P} \left(\sup_{f \in \mathcal{F}_{\varepsilon/8}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i(\bar{f}(x_i)) \right| + \varepsilon/8 > \frac{\varepsilon}{4} \right) \\
 &\leq |\mathcal{F}_{\varepsilon/8}| \max_{f \in \mathcal{F}_{\varepsilon/8}} \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n \sigma_i(\bar{f}(x_i)) \right| > \frac{\varepsilon}{8} \right) \\
 &\leq \mathcal{N}_1(\varepsilon/8, \mathcal{F}, \mathbf{x}) \max_{f \in \mathcal{F}_{\varepsilon/8}} \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n \sigma_i(f(x_i)) \right| > \frac{\varepsilon}{8} \right).
 \end{aligned} \tag{B.15}$$

The last step is a simple application of Hoeffding's inequality for the bounded random variables $\sigma_i f(x_i) \in [-B, B]$ for $i \in [n]$. It yields

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i) \right| > \frac{\varepsilon}{8} \right) \leq 2 \exp \left(- \frac{2n(\varepsilon/8)^2}{4B^2} \right) = 2 \exp \left(- \frac{n\varepsilon^2}{128B^2} \right) \tag{B.16}$$

Averaging out on the condition yields the theorem. \square

B.3 Stone's Theorem

Theorem 3.2.1. Let $\{(X_i, Y_i)\}_{i=1}^n$ be the given i.i.d. sample, $f_n(x) = \sum_{i=1}^n w_{n,i}(x) Y_i$ be a local averaging estimate, see (3.2), and X be identically distributed as X_1 independent of the given sample. Assume that for all possible distributions of X :

i There exists a constant c such that for every (measurable) nonnegative function f satisfying $\mathbb{E}f(X) < \infty$ and any $n \in \mathbb{N}$ the following holds

$$\mathbb{E}\left(\sum_{i=1}^n |w_{n,i}(X)| f(X_i)\right) \leq c \mathbb{E}f(X). \quad (\text{B.17})$$

ii There is $D \geq 1$ such that for all $n \in \mathbb{N}$

$$\mathbb{P}\left(\sum_{i=1}^n |w_{n,i}(X)| \leq D\right) = 1. \quad (\text{B.18})$$

iii For all $a > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{E}\left(\sum_{i=1}^n |w_{n,i}(X)| \mathbb{I}(\|X_i - X\| > a)\right) = 0. \quad (\text{B.19})$$

iv

$$\sum_{i=1}^n w_{n,i}(X) \xrightarrow{p} 0 \quad (\text{B.20})$$

v

$$\lim_{n \rightarrow \infty} \mathbb{E}\left(\sum_{i=1}^n w_{n,i}^2(X)\right) = 0 \quad (\text{B.21})$$

then the corresponding local averaging estimate, f_n , is universally consistent.

Proof. Using that $(a + b + c)^2 \leq 3(a^2 + b^2 + c^2)$ we have

$$\begin{aligned} \mathbb{E}\left[(f_n(X) - f_*(X))^2\right] &\leq 3\mathbb{E}\left[\left(\sum_{i=1}^n w_{n,i}(X)Y_i - \sum_{i=1}^n w_{n,i}(X)f_*(X_i)\right)^2\right] \\ &+ 3\mathbb{E}\left[\left(\sum_{i=1}^n w_{n,i}(X)f_*(X_i) - \sum_{i=1}^n w_{n,i}(X)f_*(X)\right)^2\right] + 3\mathbb{E}\left[\left(\sum_{i=1}^n w_{n,i}(X)f_*(X) - f_*(X)\right)^2\right]. \end{aligned} \quad (\text{B.22})$$

We are going to deal with these three terms separately. First, we bound the second term by the Cauchy–Schwartz inequality and condition ii

$$\begin{aligned} &\mathbb{E}\left[\left(\sum_{i=1}^n w_{n,i}(X)f_*(X_i) - \sum_{i=1}^n w_{n,i}(X)f_*(X)\right)^2\right] \\ &\leq \mathbb{E}\left[\left(\sum_{i=1}^n \sqrt{|w_{n,i}(X)|} \sqrt{|w_{n,i}(X)|} |f_*(X_i) - f_*(X)|\right)^2\right] \\ &\leq \mathbb{E}\left[\left(\sum_{i=1}^n |w_{n,i}(X)|\right) \left(\sum_{i=1}^n |w_{n,i}(X)| |f_*(X_i) - f_*(X)|^2\right)\right] \end{aligned}$$

$$\leq D \mathbb{E} \left[\sum_{i=1}^n |w_{n,i}(X)| |f_*(X_i) - f_*(X)|^2 \right]. \quad (\text{B.23})$$

Because of the denseness result that we mentioned above, for an arbitrary $\varepsilon > 0$ there exists a bounded uniformly continuous \bar{f} such that

$$\mathbb{E} \left[(f_*(X) - \bar{f}(X))^2 \right] < \varepsilon. \quad (\text{B.24})$$

Applying the same algebraic inequality as before yields

$$\begin{aligned} \mathbb{E} \left[\sum_{i=1}^n |w_{n,i}(X)| |f_*(X_i) - f_*(X)|^2 \right] &\leq 3\mathbb{E} \left[\sum_{i=1}^n |w_{n,i}(X)| |f_*(X_i) - \bar{f}(X_i)|^2 \right] \\ &+ 3\mathbb{E} \left[\sum_{i=1}^n |w_{n,i}(X)| |\bar{f}(X_i) - \bar{f}(X)|^2 \right] + 3\mathbb{E} \left[\sum_{i=1}^n |w_{n,i}(X)| |\bar{f}(X) - f_*(X)|^2 \right] \\ &\leq 3A_1 + 3A_2 + 3A_3. \end{aligned} \quad (\text{B.25})$$

We show that $A_2 \rightarrow 0$. Since $(a - b)^2 \leq 2(a^2 + b^2)$ for all $\delta > 0$

$$\begin{aligned} A_2 &\leq \mathbb{E} \left(\sum_{i=1}^n |w_{n,i}(X)| |\bar{f}(X_i) - \bar{f}(X)|^2 \mathbb{I}(\|X_i - X\| > \delta) \right) \\ &+ \mathbb{E} \left(\sum_{i=1}^n |w_{n,i}(X)| |\bar{f}(X_i) - \bar{f}(X)|^2 \mathbb{I}(\|X_i - X\| \leq \delta) \right) \\ &\leq \mathbb{E} \left(\sum_{i=1}^n |w_{n,i}(X)| (2|\bar{f}(X_i)| + 2|\bar{f}(X)|) \mathbb{I}(\|X_i - X\| > \delta) \right) \\ &+ \mathbb{E} \left(\sum_{i=1}^n |w_{n,i}(X)| |\bar{f}(X_i) - \bar{f}(X)|^2 \mathbb{I}(\|X_i - X\| \leq \delta) \right) \\ &\leq 4 \sup_{x \in \mathbb{X}} |\bar{f}(x)| \mathbb{E} \left(\sum_{i=1}^n |w_{n,i}(X)| \mathbb{I}(\|X_i - X\| > \delta) \right) + D \sup_{u,v, \|u-v\| \leq \delta} |\bar{f}(u) - \bar{f}(v)|. \end{aligned} \quad (\text{B.26})$$

Because of condition *iii* and the uniform continuity of \bar{f}

$$\limsup_{n \rightarrow \infty} A_2 \leq D\varepsilon, \quad (\text{B.27})$$

for all $\varepsilon > 0$, hence $A_2 \rightarrow 0$. From condition *ii* it follows that

$$A_3 \leq D\varepsilon. \quad (\text{B.28})$$

By condition *i*

$$A_1 \leq c \mathbb{E} \left[(f_*(X) - \bar{f}(X))^2 \right] \leq c\varepsilon. \quad (\text{B.29})$$

Hence we showed that the second term, $\mathbb{E} \left(\sum_{i=1}^n |w_{n,i}(X)| |f_*(X_i) - f_*(X)|^2 \right)$, tends to zero.

For the first term let $\sigma^2(x) \doteq \mathbb{E} \left((f(X) - Y)^2 | X = x \right)$. From $\mathbb{E}Y^2 < \infty$ it follows that

$\mathbb{E}\sigma^2(X) < \infty$. Clearly

$$\begin{aligned} & \mathbb{E} \left[\left(\sum_{i=1}^n w_{n,i}(X) Y_i - \sum_{i=1}^n w_{n,i}(X) f_*(X_i) \right)^2 \right] \\ &= \mathbb{E} \left(\sum_{i=1}^n \sum_{j=1}^n w_{n,i}(X) w_{n,j}(X) (Y_i - f_*(X_i)) (Y_j - f_*(X_j)) \right). \end{aligned} \quad (\text{B.30})$$

For $i \neq j$ we have

$$\begin{aligned} & \mathbb{E} \left[w_{n,i}(X) w_{n,j}(X) (Y_i - f_*(X_i)) (Y_j - f_*(X_j)) \right] \\ &= \mathbb{E} \left[\mathbb{E} \left(w_{n,i}(X) (Y_i - f_*(X_i)) w_{n,j}(X) (Y_j - f_*(X_j)) \mid X_1, \dots, X_n, Y_i \right) \right] \\ &= \mathbb{E} \left[w_{n,i}(X) (Y_i - f_*(X_i)) w_{n,j}(X) \mathbb{E} \left((Y_j - f_*(X_j)) \mid X_1, \dots, X_n, Y_i \right) \right] \\ &= \mathbb{E} \left[w_{n,i}(X) (Y_i - f_*(X_i)) w_{n,j}(X) (f_*(X_j) - f_*(X_j)) \right] = 0. \end{aligned} \quad (\text{B.31})$$

From which it follows that

$$\begin{aligned} & \mathbb{E} \left(\sum_{i=1}^n \sum_{j=1}^n w_{n,i}(X) w_{n,j}(X) (Y_i - f_*(X_i)) (Y_j - f_*(X_j)) \right) \\ &= \mathbb{E} \left(\sum_{i=1}^n w_{n,i}^2(X) (Y_i - f_*(X_i))^2 \right) = \mathbb{E} \left(\sum_{i=1}^n w_{n,i}^2(X) \sigma^2(X_i) \right). \end{aligned} \quad (\text{B.32})$$

If σ^2 is bounded then $\mathbb{E} \left(\sum_{i=1}^n w_{n,i}^2(X) \sigma^2(X_i) \right) \rightarrow 0$ by condition v . For general σ^2 we can apply the denseness result since we saw that $\sigma^2 \in L_1(P_X)$. Therefore for all $\varepsilon > 0$ there exists a bounded function, $\bar{\sigma}^2(x) \leq L$ for all $x \in \mathbb{X}$, such that

$$\mathbb{E} (|\sigma^2(X) - \bar{\sigma}^2(X)|) < \varepsilon. \quad (\text{B.33})$$

We can bound the first term by the triangle inequality and condition ii

$$\begin{aligned} & \mathbb{E} \left[\left(\sum_{i=1}^n w_{n,i}(X) (Y_i - f_*(X_i)) \right)^2 \right] \leq \mathbb{E} \left(\sum_{i=1}^n w_{n,i}^2(X) \sigma^2(X_i) \right) \\ &\leq \mathbb{E} \left(\sum_{i=1}^n w_{n,i}^2(X) \bar{\sigma}^2(X_i) \right) + \mathbb{E} \left(\sum_{i=1}^n w_{n,i}^2(X) |\sigma^2(X_i) - \bar{\sigma}^2(X_i)| \right) \\ &\leq L \mathbb{E} \left(\sum_{i=1}^n w_{n,i}^2(X) \right) + D \mathbb{E} \left(\sum_{i=1}^n |w_{n,i}(X)| |\sigma^2(X_i) - \bar{\sigma}^2(X_i)| \right). \end{aligned} \quad (\text{B.34})$$

Then we obtain the following by using condition ii and v

$$\limsup_{n \rightarrow \infty} \mathbb{E} \left[\left(\sum_{i=1}^n w_{n,i}(X) (Y_i - f_*(X_i)) \right)^2 \right] \leq cD \mathbb{E} (|\sigma^2(X) - \bar{\sigma}^2(X)|) \leq cD\varepsilon. \quad (\text{B.35})$$

For the third term we apply the dominated convergence theorem with integrable dominant

$D \cdot f_*(X)$. Then by condition *iv* it follows that

$$\mathbb{E} \left[\left(\sum_{i=1}^n w_{n,i}(X) f_*(X) - f_*(X) \right)^2 \right] = \mathbb{E} \left[\left(\sum_{i=1}^n (w_{n,i}(X) - 1) f_*(X) \right)^2 \right] \rightarrow 0. \quad (\text{B.36})$$

So far we proved that $\mathbb{E}[(f_n(X) - f_*(X))^2]$ goes to zero. It is sufficient for convergence in probability, since by Markov's inequality

$$\mathbb{P} \left(\left(\sum_{i=1}^n w_{n,i}(X) f_*(X) - f_*(X) \right)^2 > \varepsilon \right) \leq \frac{\mathbb{E} \left[\left(\sum_{i=1}^n w_{n,i}(X) f_*(X) - f_*(X) \right)^2 \right]}{\varepsilon}. \quad (\text{B.37})$$

□

B.4 Banach–Steinhaus Theorem for Integral Operators

Theorem 3.5.1. *Let K_n be $\mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ type functions for $n \in \mathbb{N}$ and μ be a probability measure on \mathbb{R}^d . Assume the followings:*

i There exists $c > 0$ such that for all $n \in \mathbb{N}$ the following holds

$$\int |K_n(x, z)| \, d\mu(x) \leq c \quad (\text{B.38})$$

for μ -almost every z .

ii There exists $D \geq 1$ such that for all $x \in \mathbb{R}^d$ and for all $n \in \mathbb{N}$

$$\int |K_n(x, z)| \, d\mu(z) \leq D. \quad (\text{B.39})$$

iii For all $a > 0$

$$\lim_{n \rightarrow \infty} \int \int |K_n(x, z)| \, \mathbb{I}(\|x - z\| > a) \, d\mu(z) \, d\mu(x) = 0. \quad (\text{B.40})$$

iv

$$\lim_{n \rightarrow \infty} \text{ess sup}_x \left| \int K_n(x, z) \, d\mu(z) - 1 \right| = 0. \quad (\text{B.41})$$

Then for all $f \in L_1(\mu)$

$$\lim_{n \rightarrow \infty} \int \left| f(x) - \int K_n(x, z) f(z) \, d\mu(z) \right| \, d\mu(x) = 0. \quad (\text{B.42})$$

Proof. We use the denseness result we mentioned before Stone's theorem. For all $\varepsilon > 0$ there exists a uniformly continuous, bounded function, \tilde{f} , with compact support for which

$$\int |f(x) - \tilde{f}(x)| \, d\mu(x) < \varepsilon \quad (\text{B.43})$$

holds. Then the quantity in (B.42) can be bounded as

$$\begin{aligned} & \int \left| f(x) - \int K_n(x, z) f(z) \, d\mu(z) \right| \, d\mu(x) \leq \int \left| f(x) - \tilde{f}(x) \right| \, d\mu(x) \\ & + \int \left| \tilde{f}(x) \left(1 - \int K_n(x, z) \, d\mu(z) \right) \right| \, d\mu(x) + \int \left| \int K_n(x, z) (\tilde{f}(x) - \tilde{f}(z)) \, d\mu(z) \right| \, d\mu(x) \quad (\text{B.44}) \\ & + \int \left| \int K_n(x, z) (\tilde{f}(z) - f(z)) \, d\mu(z) \right| \, d\mu(x) = I_1 + I_2 + I_3 + I_4. \end{aligned}$$

We proceed by bounding these four terms separately. Because of the choice of \tilde{f} the first term $I_1 < \varepsilon$. From condition *iv*

$$I_2 \leq \text{ess sup}_u \left| 1 - \int K_n(u, z) \, d\mu(z) \right| \int \left| \tilde{f}(x) \right| \, d\mu(x) \rightarrow 0. \quad (\text{B.45})$$

Since \tilde{f} is uniformly continuous for $\varepsilon > 0$ let $\delta > 0$ be such that from $\|x - z\| < \delta$ it follows that $|\tilde{f}(x) - \tilde{f}(z)| < \varepsilon$. Let $B = B(x, \delta/2)$ and \bar{B} denote its complement. Then

$$\begin{aligned} I_3 & \leq \int \int_B |K_n(x, z)| |\tilde{f}(x) - \tilde{f}(z)| \, d\mu(z) \, d\mu(x) \\ & + \int \int_{\bar{B}} |K_n(x, z)| |\tilde{f}(x) - \tilde{f}(z)| \, d\mu(z) \, d\mu(x) \quad (\text{B.46}) \\ & \leq \varepsilon \int \int_B |K_n(x, z)| \, d\mu(z) \, d\mu(x) + 2 \sup_x |\tilde{f}(x)| \int \int_{\bar{B}} |K_n(x, z)| \, d\mu(z) \, d\mu(x), \end{aligned}$$

where by condition *ii* the first term is limited by εD and by condition *iii* the second term tends to zero. For I_4 the following holds by Fubini's theorem and condition *i*

$$\begin{aligned} I_4 & \leq \int \left| \int K_n(x, z) \, d\mu(x) \right| |\tilde{f}(z) - f(z)| \, d\mu(z) \\ & \leq c \int |\tilde{f}(z) - f(z)| \, d\mu(z) \leq c\varepsilon, \end{aligned} \quad (\text{B.47})$$

which finishes the proof. \square

B.5 Covering Lemma

Lemma 3.5.3. *Let K be a regular kernel. Then there exists a finite constant $\varrho = \varrho(K)$ such that for all $u \in \mathbb{R}^d$, $h > 0$ and probability measure μ*

$$\int \frac{K_h(x - u)}{\int K_h(x - z) \, d\mu(z)} \, d\mu(x) \leq \varrho. \quad (\text{B.48})$$

In addition, for all $\delta > 0$

$$\lim_{n \rightarrow \infty} \sup_u \int \frac{K_h(x - u) \mathbb{I}(\|x - u\| > \delta)}{\int K_h(x - z) \, d\mu(z)} \, d\mu(x) = 0. \quad (\text{B.49})$$

Proof. Cover the whole \mathbb{R}^d with countably many balls with radius $r/2$ on the following way. Let the centers of these balls be $x_k = (k_1 r/2, \dots, k_d r/2)$ for all $k = (k_1, \dots, k_n) \in \mathbb{Z}^d$. Then it

is easy to see that $\cup_{k \in \mathbb{Z}^d} B(x_k, r/2) \supseteq \mathbb{R}^d$. For practical reasons we reindex the centers with the natural numbers. It can be done since \mathbb{Z}^d is countable. Notice that we used infinitely many balls to cover \mathbb{R}^d , but all $x \in \mathbb{R}^d$ is covered at most $2^d + 1$ times. Because of the regularity of K

$$\begin{aligned} \sum_{i=1}^{\infty} \sup_{z \in B(x_i, r/2)} K(z) &= \sum_{i=1}^{\infty} \frac{1}{\int_{B(0, r/2)} 1 dx} \int_{B(x_i, r/2)} \sup_{z \in B(x_i, r/2)} K(z) dx \\ &\leq \frac{1}{\int_{B(0, r/2)} 1 dx} \int \sum_{i=1}^{\infty} \sup_{z \in B(x, r)} K(z) \mathbb{I}(x \in B(x_i, r/2)) dx \\ &\leq \frac{2^d + 1}{\int_{B(0, r/2)} 1 dx} \int \sup_{z \in B(x, r)} K(z) dx \leq C, \end{aligned} \quad (\text{B.50})$$

where we used that $B(x_i, r/2) \subseteq B(x, r)$ when $x \in B(x_i, r/2)$.

The covering property yields that

$$K_h(x - u) \leq \sum_{i=1}^{\infty} \sup_{x \in B(u + hx_i, rh/2)} K_h(x - u). \quad (\text{B.51})$$

In addition, for all $x \in B(u + hx_i, rh/2)$ we have

$$\int K_h(x - z) d\mu(z) \geq b\mu(B(x, rh)) \geq b\mu(B(u + hx_i, rh/2)). \quad (\text{B.52})$$

By these we can proceed as

$$\begin{aligned} \int \frac{K_h(x - u)}{\int K_h(x - z) d\mu(z)} d\mu(x) &\leq \sum_{i=1}^{\infty} \int_{B(u + hx_i, rh/2)} \frac{K_h(x - u)}{\int K_h(x - z) d\mu(z)} d\mu(x) \\ &\leq \sum_{i=1}^{\infty} \int_{B(u + hx_i, rh/2)} \frac{\sup_{z \in B(hx_i, rh/2)} K_h(z)}{b\mu(B(u + hx_i, rh/2))} d\mu(x) \\ &\leq \sum_{i=1}^{\infty} \frac{\mu(B(u + hx_i, rh/2)) \sup_{z \in B(hx_i, rh/2)} K_h(z)}{b\mu(B(u + hx_i, rh/2))} \leq \frac{1}{b} \sum_{i=1}^{\infty} \sup_{z \in B(x_i, r/2)} K(z) \leq \frac{C}{b}, \end{aligned} \quad (\text{B.53})$$

where C only depends on the dimension and on the kernel function. If we substitute $K_h(z) \mathbb{I}(\|z\| > \delta)$ in the place of $K_h(z)$ we obtain that

$$\sup_u \int \frac{K_h(x - u) \mathbb{I}(\|x - u\| > \delta)}{\int K_h(x - z) d\mu(z)} d\mu(x) \leq \frac{1}{b} \sum_{i=1}^{\infty} \sup_{z \in B(x_i, r/2)} K(z) \mathbb{I}(\|z\| > \delta/h), \quad (\text{B.54})$$

which goes to zero as $h \rightarrow 0$ by the dominant convergence theorem. \square

Bibliography

- [1] N. Aronszajn. Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.
- [2] S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [3] C. J. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.
- [4] F. P. Cantelli. Sulla determinazione empirica delle leggi di probabilita. *Giorn. Ist. Ital. Attuari*, 4(421-424), 1933.
- [5] A. Carè, B. Cs. Csáji, M. Campi, and E. Weyer. Finite-Sample System Identification: An Overview and a New Correlation Method. *IEEE Control Systems Letters*, 2(1):61 – 66, 2018.
- [6] B. Cs. Csáji, M. C. Campi, and E. Weyer. Sign-Perturbed Sums: A New System Identification Approach for Constructing Exact Non-Asymptotic Confidence Regions in Linear Regression Models. *IEEE Transactions on Signal Processing*, 63(1):169–181, 2015.
- [7] B. Cs. Csáji and A. Tamás. Semi-Parametric Uncertainty Bounds for Binary Classification. In *58th IEEE Conference on Decision and Control (CDC), Nice, France*, 2019.
- [8] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer Science & Business Media, 2013.
- [9] E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*, volume 40. Cambridge University Press, 2015.
- [10] V. Glivenko. Sulla determinazione empirica delle leggi di probabilita. *Gion. Ist. Ital. Attuari.*, 4:92–99, 1933.
- [11] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.
- [12] D. Haussler. Sphere Packing Numbers for Subsets of the Boolean n-Cube with Bounded Vapnik–Chervonenkis Dimension. *J. Comb. Theory, Ser. A*, 69(2):217–232, 1995.

- [13] D. Hush and C. Scovel. On the VC Dimension of Bounded Margin Classifiers. *Machine Learning*, 45(1):33–44, 2001.
- [14] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning*, volume 112. Springer, 2013.
- [15] S. Kolumbán. *System Identification in Highly Non-Informative Environment*. PhD thesis, Budapest University of Technology and Economics, Hungary, and Vrije Univesiteit Brussels, Belgium, 2016.
- [16] C. A. Micchelli, Y. Xu, and H. Zhang. Universal Kernels. *Journal of Machine Leing Research*, 7(Dec):2651–2667, 2006.
- [17] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends in Machine Learning*, pages 1–141, 2017.
- [18] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.
- [19] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014.
- [20] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, Stability and Uniform Convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010.
- [21] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
- [22] R. L. Taylor. *Stochastic Convergence of Weighted Sums of Random Elements in Linear Spaces*, volume 672. Springer, 1978.
- [23] V. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.
- [24] V. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer Science & Business Media, 2006.
- [25] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.