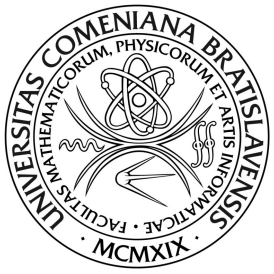
COMENIUS UNIVERSITY IN BRATISLAVA
FACULTY OF MATHEMATICS, PHYSICS AND INFORMATICS



ASYMPTOTIC PROPERTIES OF SUPPORT VECTOR MACHINES

Master's Thesis

Bc. Jana Janková

BRATISLAVA 2012

Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics
Department of Applied Mathematics and Statistics

Vrije Universiteit Amsterdam
Faculty of Sciences
Department of Mathematics

Asymptotic Properties of Support Vector Machines

Master's Thesis

Bc. Jana JANKOVÁ

Study Programme:	Economic and Financial Mathematics
Branch of Study:	1114 Applied Mathematics
Supervisor:	Prof. RNDr. Marek Fila, DrSc. Comenius University in Bratislava
Co-supervisor:	Prof. dr. Aad W. van der Vaart Vrije Universiteit Amsterdam

BRATISLAVA 2012

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky
Katedra aplikovanej matematiky a štatistiky

Vrije Universiteit Amsterdam
Faculty of Sciences
Department of Mathematics

Asymptotické vlastnosti metódy oporných bodov

Diplomová práca

Bc. Jana JANKOVÁ

Študijný Program:	Ekonomická a finančná matematika
Študijný odbor:	1114 Aplikovaná matematika
Školiteľ:	Prof. RNDr. Marek Fila, DrSc. Univerzita Komenského v Bratislave
Konzultant:	Prof. dr. Aad W. van der Vaart Vrije Universiteit Amsterdam

BRATISLAVA 2012

77351015

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Jana Janková
Študijný program: ekonomická a finančná matematika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: 9.1.9. aplikovaná matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: anglický

Názov: Asymptotic Properties of Support Vector Machines

Cieľ: Study of asymptotic properties of support vector machines

Vedúci: prof. RNDr. Marek Fila, DrSc.
Konzultant: prof. Dr. Aad W. van der Vaart
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky

Dátum zadania: 13.01.2011

Dátum schválenia: 14.01.2011

prof. RNDr. Daniel Ševčovič, CSc.
garant študijného programu

.....
študent

.....
vedúci práce,
konzultant práce

Acknowledgement

I would like to express my gratitude to my supervisor and co-supervisor, without whom this research project would not have been possible. In particular, to Prof. Aad van der Vaart for a nice topic suggestion, helpful discussions and for all the time he spent with me. Special thanks go to my mum, dad and my sister for supporting me throughout the duration of my studies.

Jana Janková
Amsterdam, 2012

Declaration on Word of Honour

I declare that all parts of this thesis have been written by myself using only the references explicitly referred to in the text and consultations with my supervisor.

.....

Jana Janková

Bratislava, 2012

JANKOVÁ, Jana. Asymptotic Properties of Support Vector Machines, Master's Thesis, Bratislava 2012.

Supervisor: Prof. RNDr. Marek Fila, DrSc.

Department of Applied Mathematics and Statistics
Faculty of Mathematics, Physics and Informatics
Comenius University in Bratislava

Co-supervisor: Prof. dr. Aad W. van der Vaart

Department of Mathematics
Faculty of Sciences
Vrije Universiteit Amsterdam

Abstract

The aim of this work is to study the support vector machine (SVM) algorithm from a statistical perspective using tools of empirical processes and concentration theory. The gist of this approach lies in casting the support vector machine as a regularized empirical minimization scheme where the regularizer is the squared norm in a Hilbert space of functions on the input space. Recent results on convergence rates of empirical contrast estimators, and specifically, the support vector machine, to the Bayes risk are presented and discussed. For support vector machines with Gaussian kernels, we follow the approach of Van der Vaart and Wellner in [11], Bousquet, Blanchard and Massart [2] and finally Steinwart and Scovel in [9]. Under Tsybakov noise assumption and a geometric noise assumption on the underlying distribution, and using properties of reproducing kernel Hilbert spaces with a Gaussian kernel analyzed in [12], we obtain an improvement of the bound on the convergence rates presented in [9]. Finally, we rederive the same rates using a different bound on the entropy number of a unit ball in a reproducing kernel Hilbert space with a Gaussian kernel.

Keywords: support vector machine, Bayes risk, empirical process, Gaussian kernels, entropy

JANKOVÁ, Jana. Asymptotické vlastnosti metódy oporných bodov, Diplomová práca, Bratislava 2012.

Školiteľ: Prof. RNDr. Marek Fila, DrSc.

Katedra aplikovanej matematiky a štatistiky
Fakulta matematiky, fyziky a informatiky
Univerzita Komenského v Bratislave

Konzultant: Prof. dr. Aad W. van der Vaart

Department of Mathematics
Faculty of Sciences
Vrije Universiteit Amsterdam

Abstrakt

V práci sa venujeme štúdiu metódy oporných bodov prostredníctvom teórie empirických procesov a teórie koncentrácie pravdepodobnostných mier. Hlavnou ideou tohto prístupu je pozorovanie, že metóda oporných bodov je špeciálnym prípadom štatistickej procedúry, známej ako minimalizácia empirického rizika s regularizáciou, kde regularizáciu predstavuje norma v Hilbertovom priestore funkcií definovaných na množine vstupov. Predstavíme a porovnáme najnovšie výsledky pre rýchlosť konvergencie odhadov založených na minmizalizácii empirického rizika k Bayesovskému riziku a zhrnieme známe výsledky pre metódu oporných bodov. Zameriame sa na metódu oporných bodov s Gaussovskými jadrami a študujeme ich podobne ako v práci Van der Vaarta a Wellnera [11], Bousqueta, Blancharda a Massarta [2] a napokon Steinwarta a Scovela [9]. Za Tsybakovho predpodkladu na šum v rozdelení a za predpokladu na geometrický šum v rozdelení, a tiež pomocou vlastností tzv. Hilbertových priestorov s Gaussovským jadrom podrobne analyzovaných v [12], odvodíme rýchlosti konvergencie k Bayesovskému riziku, ktoré vylepšujú odhady odvodené v [9]. Napokon rovnaké výsledky odvodíme pomocou iného ohraničenia na entropiu, v súlade s prácou [9].

Kľúčové slová: metóda oporných bodov, Bayesovské riziko, empirický proces, Gaussovské jadrá, entropia

Contents

List of Symbols

$\mathcal{X}$	input space
$\mathcal{Y}$	label space
$P, \mathbb{E}, \mathbb{E}_P, \mathbb{E}_P^*$	expectation with respect to the (outer) measure $\mathbb{P}(^*)$
$\mathbb{P}_X$	marginal distribution of $\mathbb{P}$ with respect to the random variable X
$\|\cdot\|_\infty, \|\cdot\|_2$	uniform norm, Euclidean norm
$L^r(P)$	space of measurable functions that are $r-$integrable w.r.t. $\mathbb{P}$
$\|\cdot\|_{P,r}$	$L^r(\mathbb{P})-$norm
$\mathbb{P}_n$	empirical measure
$\mathbb{G}_n$	empirical process
$\mathbb{H}, \mathbb{H}(\mathcal{X})$	reproducing kernel Hilbert space (RKHS) on $\mathcal{X}$
$\mathbb{H}_\sigma, \mathbb{H}_\sigma(\mathcal{X})$	Gaussian kernel RKHS on $\mathcal{X}$ with kernel width $1/\sigma$
$B_{\mathbb{H}}$	unit ball in the space $\langle \mathbb{H}, \|\cdot\|_{\mathbb{H}} \rangle$
$\lesssim$	less than up to a constant
$\sim$	$\lesssim$ and $\gtrsim$
l_θ	0-1 loss function
m_θ	hinge loss function
$O_P(1), O(1)$	stochastic order symbol and (Landau) big O notation
$L^r(\mathbb{R}^d)$	L^r-space on $\mathbb{R}^d$ with respect to the Lebesgue measure
$\mathbb{R}, \mathbb{N}, \mathbb{C}$	number sets

Introduction

Why asymptotic statistics? The use of asymptotic approximations is twofold. First, they enable us to find approximate tests and confidence regions. Second, approximations can be used theoretically to study the quality (efficiency) of statistical procedures.

Aad van der Vaart

The goal of statistical learning theory is to provide the theoretical framework to study problems of inference such as constructing models from a set of data. Under assumptions of statistical nature, this theory produces techniques that allow for studying properties of learning algorithms that are increasingly popular in a variety of applications such as speech and text recognition, image analysis and data mining.

In this work, we specify to a subclass of supervised learning procedures known as binary classification. Within this framework, data consists of label-instance pairs, where the label only takes the values -1 or 1. Given a data set, a learning algorithm aims at constructing a mapping from the space of instances to the space of labels. It seems reasonable that the mapping should aim to minimize the probability of wrong classification when predicting the label of unseen instances. Although given a training set, one could always build a function that fits the data exactly, in general it is not a wise thing to do, as in presence of some noise in the data, this would lead to very poor performance on unseen instances. In general, one aims to construct a model which fits the data well, but is, in some sense, as simple as possible. That is, one looks for some *regularities*.

This work focuses on a special instance of statistical learning algorithms for binary classification known as the support vector machine (SVM). A support vector machine (in high-dimensional or infinite-dimensional space) constructs a hyperplane that separates the two groups of data with a gap as wide as possible. The original formulation was proposed by Vapnik [15] in 1963 as a linear classifier. However, this simple formulation only applies when the two data sets are linearly separable. In 1995, Vapnik and Cortes [5] proposed a soft-margin version of the algorithm that allows for mislabeled examples. Furthermore, in 1992, Boser, Guyon and Vapnik suggested a way of creating nonlinear classifiers by applying the kernel trick [3]. This essentially means that the problem is transformed into a higher dimensional space, where the original algorithm is applied. However, in the original space the corresponding classifier may be nonlinear.

The success of the SVM algorithm is mainly due to the number of experimental results

that have been obtained in very diverse domains of application, such as pattern recognition and regression. From the computational point of view required in applications, the algorithm is mostly treated as a convex optimization problem. However, this approach does not allow for investigation of its statistical behaviour which still remains only partially understood. This may be achieved by expressing the problem as the minimization of a regularized functional where the regularizer is the squared norm in a Hilbert space of functions on the input space. Our goal in this work is to adopt the latter approach to investigate the properties of the support vector machine algorithm in a statistical setting. As the main tool that develops the theoretical framework to study statistical properties of learning algorithms, we introduce some results from the theory of empirical processes.

In the first chapter, we outline and explain binary classification, introduce the support vector machine algorithm and transform it to a more suitable form which will allow us to apply results derived for empirical risk minimization estimation tasks. We introduce reproducing kernel Hilbert spaces and discuss some of their properties.

In the second chapter we give an introduction to the theory of empirical processes and the penalized empirical minimization problem and note that the support vector machine may be cast in this framework, as derived in Chapter 1. We explain the general approach which gives us the rates of convergence for specific problems arising in empirical risk minimization procedures. We reformulate the results to fit the framework of the SVM. In the end of the chapter, we discuss entropy bounds for unit balls in RKHSs, which are crucial to obtain any rates of convergence in light of the presented results.

The third chapter further presents some recent results on convergence rates for SVMs. First we present and discuss a result for the special case of Gaussian kernel RKHS as given in [9] and then a more general result derived in [2]. In the second part of this chapter, we adopt the approaches in [11], [9] and [2] for reproducing kernel Hilbert spaces with a Gaussian kernel function. We follow these to obtain convergence rates to the Bayes risk for SVMs with a Gaussian kernel. We summarize our results and discuss their implications.

Chapter 1

Notation and Preliminaries

This chapter provides an introduction to the problem of statistical classification and treats relevant theoretical framework. We first introduce binary classification in general and then specialize to a binary classification algorithm known as the support vector machine, treating it first mainly as a problem of convex optimization. Later on we underline the connection between the SVM algorithm and a statistical procedure called the *penalized empirical contrast*.

We provide an overview of *kernel functions* which lead to a generalization of the SVM algorithm and we introduce associated *reproducing kernel Hilbert spaces*. We specifically consider reproducing kernel Hilbert spaces associated with a Gaussian kernel functions. We briefly discuss the notion of consistency of a classifier.

To simplify the notation, we shall use the shorthand $Pf = \int f \mathrm{d}\mathbb{P}$ for a given distribution $\mathbb{P}$.

1.1 Classification

Statistical classification is the problem of identifying the group to which a new observation belongs, purely on the basis of a training set of data for which the group labels are known.

We assume that the input data comes from a Hilbert space $\mathbb{H}$, i.e. a real (or complex) vector space endowed with an inner product and associated norm and metric that is also a complete metric space. Consider therefore the Hilbert space $(\mathcal{X}, \langle \cdot, \cdot \rangle)$ that represents the input data space and $\mathcal{Y}$ that represents the corresponding group labels of the data. We shall only consider the case $\mathcal{Y} = \{-1, 1\}$, which corresponds to the assumption that the data belongs to one of the two groups labeled with 1 or -1 (i.e. binary classification). Consider an (unknown) probability measure $\mathbb{P}$ defined on $\mathcal{X} \times \mathcal{Y}$. We are also given a *training set* $(X_1, Y_1), \ldots, (X_n, Y_n)$ that represents an i.i.d. sample from the distribution $\mathbb{P}$. Each $Y_i \in \mathcal{Y}$ indicates to which group $X_i \in \mathcal{X}$ belongs.

Based on the training set, we would like to construct a classifier that will assign a group label to a new given observation X. Formally, a classifier is simply any measurable function $\theta : \mathcal{X} \to \mathcal{Y}$. (Let us remark that we may instead consider a classifier as a function from $\mathcal{X} \to \mathbb{R}$ and classify an instance as 1 or -1 according to the sign

of $\theta(X)$. Later on we adopt this approach). The challenge of constructing a classifier now lies in how to construct a "good" classifier and, in fact, what are the requirements a classifier should satisfy.

It seems natural to require that a classifier, say θ, in some sense minimizes the risk of misclassification, that is, minimizes the probability that the predicted label $\theta(X)$ does not equal the true label Y,

$$\mathbb{P}(Y \neq \theta(X)). \tag{1.1}$$

At this point, we will make a slight diversion from the minimization problem (1.1) with the aim of introducing some more general notation which will prove to be more convenient in the sections to follow. First note that we may rewrite

$$\mathbb{P}(Y \neq \theta(X)) = P(1_{Y \neq \theta(X)}) \tag{1.2}$$

where P denotes the expectation w.r.t. $\mathbb{P}$. In general, we may denote by Pr_θ the *risk* associated with a *loss function* $r_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ given a classifier θ (w.r.t. the distribution $\mathbb{P}$). (In fact, we could equivalently consider loss functions as functions of $\theta(x)$ and y (or their product) only, since we are only interested in the label that the classifier θ assigns to x, not in the x itself.) For instance, in (1.2) the corresponding risk function is given by $r_\theta(x, y) = 1_{y \neq \theta(x)}$. We will denote this specific loss function by l_θ.

Although we did not impose any specific requirements on a loss function, some loss functions seem to be a more reasonable choice. Intuitively, loss functions represent the price we are willing to pay for predicting $\theta(x)$ in place of y. The choice of a loss function is mainly an empirical problem, and it is often the case that the choice strongly depends upon computational issues. But as a reasonable requirement, we typically want the loss function to be convex in the variable $t := \theta(x)y$. We give a few examples of the most commonly used loss functions. Note that the 0-1 loss l_θ that appears in the classification error (1.2) is not a convex function.

- 0-1 loss $l_\theta(x, y) = 1_{y\theta(x)<0}$
- hinge loss $m_\theta(x, y) = (1 - y\theta(x))_+$
- square loss $r_\theta(x, y) = (\theta(x) - y)^2$
- logistic loss $r_\theta(x, y) = (\ln 2)^{-1} \ln(1 + e^{-\theta(x)y})$

These are illustrated in figure 1.1.

1.1.1 The Bayes classifier

Now we return back to the problem of minimizing (1.1). If we knew the distribution $\mathbb{P}$, then we could explicitly find the classifier which minimizes the risk given by (1.1) over all $\theta \in \Theta$, where Θ denotes the set of all measurable functions $\theta : \mathcal{X} \to \mathcal{Y}$. This classifier is called the *Bayes classifier* and the corresponding risk is called the *Bayes risk*. The following lemma identifies the Bayes classifier in terms of the distribution $\mathbb{P}$.

Lemma 1. *The risk function (1.1) is minimized over the set Θ of all measurable maps $\theta : \mathcal{X} \to \mathbb{R}$ by*

$$\theta_0(x) \quad = \quad 2\mathbf{1}_{\eta(x) > \frac{1}{2}} - 1 \qquad \textit{where} \qquad \eta(x) := \mathbb{P}(Y = 1 | X = x).$$

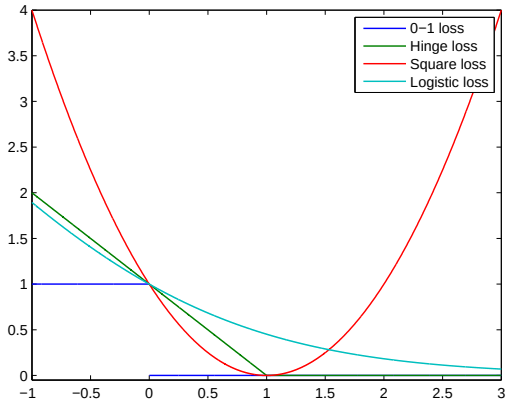


Figure 1.1: Loss functions. The variable on the x-axis is $t := y\theta(x)$ and the functions plotted are $t \mapsto V(t) = r_\theta(x, y)$. In this case, we assume an alternative definition of a classifier θ as a map from $\mathcal{X} \to \mathbb{R}$ and we classify x as 1 or -1 according to the sign of θ.

Proof. Conditioning on the input x we may write

$$\begin{aligned} Pl_\theta &= \mathbb{P}(Y = 1, \theta(X) = 1) + \mathbb{P}(Y = -1, \theta(X) = 1) \\ &= \int \eta(x)\mathbf{1}_{\theta(x)=-1} + (1 - \eta(x))\mathbf{1}_{\theta(x)=1} \mathrm{d}\mathbb{P}_X(x), \end{aligned}$$

where $\mathbb{P}_X$ denotes the marginal distribution w.r.t. X. We can minimize the integral by separately minimizing the integrand for each x. If for a given x we have $\eta(x) > 1-\eta(x)$, i.e. $\eta(x) > 1/2$, then we put $\theta_0(x) := 1$, if $\eta(x) < 1/2$ we put $\theta_0(x) := -1$. For all x such that $\eta(x) = 1/2$, the choice of the classifier is irrelevant and the minimizer is not unique on this set. Thus we see that the function $\theta_0(x) = 2\mathbf{1}_{\eta(x)>\frac{1}{2}} - 1$ is a possible choice for a minimizer. □

The function θ_0 in the preceding lemma is the Bayes classifier mentioned above. It simply classifies the input x as -1 or 1 according to the biggest of the two probabilities $\mathbb{P}(Y = 1|X = x)$ and $\mathbb{P}(Y = -1|X = x)$. In the deterministic case, i.e. $\mathbb{P}(Y = 1|X = x) \in \{0, 1\}$, we have $Y = \theta_0(X)$ almost surely and the Bayes risk is zero.

In estimation procedures based on some data, there are two restrictions we have to consider. Both arise as a consequence of the fact that the distribution $\mathbb{P}$ is unknown. Firstly, the definition of η implies that the Bayes classifier depends on the underlying distribution $\mathbb{P}$. Thus the best we can do is approximate the optimal Bayes classifier (or rather its minimal risk as outlined above) as closely as possible by a classifier $\hat{\theta}_n(.) = \hat{\theta}_n(., X_1, Y_1, \ldots, X_n, Y_n)$ based on the observations. Secondly, again since the distribution $\mathbb{P}$ is unknown, we cannot directly measure the risk (1.1) for a given estimator $\hat{\theta}_n$ based on the data. We can only measure the agreement of a candidate function with the data. For that we often consider the empirical risk

$$\mathbb{P}_n l_\theta := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\theta(X_i)\neq Y_i}$$

that is a natural estimate of the risk (1.1).

We review some common strategies that are used in learning algorithms. The basic approaches to finding an approximation to the Bayes classifier lie in

- restricting the class of functions in which minimization is considered
- modifying the criterion to be optimized.

Empirical Risk Minimization. The idea is to choose a *model* $\mathcal{G}$ of possible functions and minimize the empirical risk within that model. Thus the function returned by the algorithm, denote it $\hat{\theta}_n$, is given by

$$\hat{\theta}_n := \arg\min_{g \in \mathcal{G}} \mathbb{P}_n g.$$

This will work best if the target function θ_0 belongs to $\mathcal{G}$. However, this is often not the case.

Structural Risk Minimization. The idea is to choose an infinite sequence of models $\{\mathcal{G}_d : d = 1, 2, \dots\}$ of increasing size and minimize the empirical risk in each model with an added penalty for the size of the model. Then $\hat{\theta}_n$ is given by

$$\hat{\theta}_n := \arg\min_{g \in \mathcal{G}_d, d \in \mathbb{N}} \mathbb{P}_n g + pen(d, n).$$

The penalty $pen(d, n)$ measures the "size" of the model.

Regularization. The approach lies in choosing a large model $\mathcal{G}$ and defining a *regularizer*, typically a norm $\|g\|$. Then one has

$$\hat{\theta}_n := \arg\min_{g \in \mathcal{G}} \mathbb{P}_n g + \lambda \|g\|^2.$$

Here one has a free parameter λ, called the *smoothing parameter* which allows to choose the right trade-off between "fit" and "complexity".

We have already pointed out that it seems natural to require that the classifiers $\hat{\theta}_n$ we construct based on the training set generate risk which tends to the Bayes risk as the sample size n tends to infinity. We may thus consider the notion of *consistency* of a classifier. A *universally consistent classifier* is such, that for any probability measure the risk generated by the classifier $\hat{\theta}_n$ converges to the Bayes risk. We give a formal definition.

Definition 1. *We say that a classifier $\hat{\theta}_n(.) = \hat{\theta}_n(., X_1, Y_1, \dots, X_n, Y_n)$ based on an i.i.d. sample $(X_1, Y_1), \dots, (X_n, Y_n)$ from a distribution $\mathbb{P}$ on $\mathcal{X} \times \mathcal{Y}$ is consistent if*

$$Pl_{\hat{\theta}_n} \to Pl_{\theta_0}$$

holds in probability if $n \to \infty$. If a classifier is consistent for all distributions $\mathbb{P}$ on $\mathcal{X} \times \mathcal{Y}$, it is said to be universally consistent.

However, even if a classifier is shown to be universally consistent, this does not mean that it works well for a specific classification task. It can be shown that for no classifier there exists a uniform *rate of convergence* to the Bayes risk (a rate of convergence essentially means what sample size n we need to ensure accuracy up to some given $\varepsilon > 0$, but we will formalize the notion in the next chapter). This underlines the importance of studying the speed of convergence to the Bayes risk and will be the main focus of this work. We will return to rates of convergence in chapter 2.

1.2 Introduction to the Support Vector Machine Algorithm

Support vector machine is an algorithm for binary classification that, given an i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, where $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1, \ldots, n$ from the distribution $\mathbb{P}$, outputs a data-dependent classifier $\hat{\theta}_n : \mathcal{X} \to \{-1, 1\}$ we shall now specify.

First we shall consider the case when the data set groups $\{X_i : Y_i = 1\}$ and $\{X_i : Y_i = -1\}$ can be linearly separated in the space $\mathcal{X}$ Let us first establish some notation and a few elementary observations.

Given a "normal" vector $\beta \in \mathcal{X}$ and "displacement" $b \in \mathbb{R}$, define $H_{\beta,b}$ to be the *hyperplane*

$$H_{\beta,b} = \{x \in \mathcal{X} : \langle x, \beta \rangle + b = 0\}.$$

The distance of an arbitrary point $a \in \mathcal{X}$ to the hyperplane $H_{\beta,b}$ is given by

$$\|a - H_{\beta,b}\| = \frac{1}{\|\beta\|} |\langle a, \beta \rangle + b|,$$

where $\|x\| = \sqrt{\langle x, x \rangle}$ for every $x \in \mathcal{X}$ is the norm associated with the scalar product in the Hilbert space $\mathbb{H}$. Now we may give a definition of the support vector machine.

Definition 2. *A support vector machine is a hyperplane in $\mathcal{X}$ that separates the two sets of points $\{X_i : Y_i = -1\}$ and $\{X_i : Y_i = 1\}$ and that maximizes the minimum distance of the points to the hyperplane. Formally, a support vector machine is the hyperplane $H_{\hat{\beta},\hat{b}}$ where $(\hat{\beta}, \hat{b})$ is the solution to the optimization problem*

$$\begin{aligned} \max_{(\beta,b) \in \mathcal{X} \times \mathbb{R}} \min_i \frac{1}{\|\beta\|} |\langle X_i, \beta \rangle + b| & \\ \langle X_i, \beta \rangle + b &\leq 0 \quad \text{if } Y_i = -1 \qquad (1.3) \\ \langle X_i, \beta \rangle + b &\geq 0 \quad \text{if } Y_i = 1 \qquad (1.4) \\ & \text{for } i = 1, \ldots, n. \end{aligned}$$

The classifier corresponding to the support vector machine is as follows: classify a new given observation x as 1 if $\langle x, \hat{\beta} \rangle + \hat{b} \geq 0$ and as -1 if $\langle x, \hat{\beta} \rangle + \hat{b} < 0$.

The two halfspaces defined by the hyperplane $H_{\beta,b}$ are determined by the sign of $\langle a, \beta \rangle + b$, thus the two sets of points are separated by $H_{\beta,b}$ if and only if the numbers $Y_i(\langle X_i, \beta \rangle + b)$ for $i = 1, \ldots, n$ possess the same sign. As given in definition 2, the support vector machine is the hyperplane $H_{\beta,b}$ defined by a pair (β, b) that maximizes

$$\min_i \frac{1}{\|\beta\|} |\langle X_i, \beta \rangle + b|,$$

under these restrictions. Noting that a pair $(c\beta, cb)$ describes the same hyperplane as the pair (β, b) for any $c \in \mathbb{R}$, we replace the restrictions (1.3) and (1.4) by

$$\min_i Y_i(\langle X_i, \beta \rangle + b) = 1. \qquad (1.5)$$

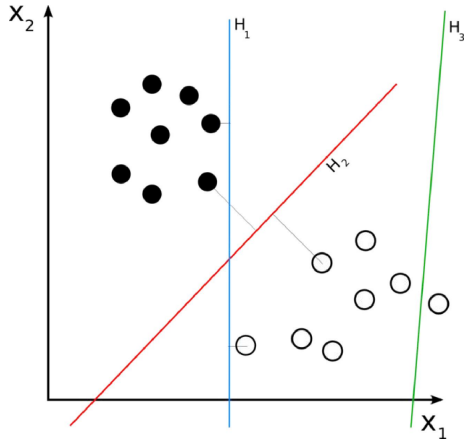


Figure 1.2: Support vector machine in $\mathcal{X} = \mathbb{R}^2$. Data point color corresponds to the class label of that particular data point (i.e. black: $Y = 1$, white: $Y = -1$). Hyperplanes H_1 and H_2 separate the two sets of points, while H_3 does not. Hyperplane H_2 gives a larger minimum distance (so-called "margin") to the data set points. [17]

The criterion function $\frac{1}{\|\beta\|}|\langle X_i, \beta\rangle + b|$ then reduces to $\frac{1}{\|\beta\|}$; equivalently, we minimize $\frac{1}{2}\|\beta\|^2$ (we include a factor $\frac{1}{2}$ for mathematical convenience). The restriction can next be relaxed by replacing the equality in (1.5) by $\geq$, as a value (β, b) giving a minimum $m = \min_i Y_i(\langle X_i, \beta\rangle + b)$ strictly larger than 1 gives a larger value to the criterion than the value $(\beta/m, b/m)$. Hence we may recast the support vector machine to be the hyperplane $H_{\beta,b}$ found by minimizing the criterion

$$f(\beta, b) := \frac{1}{2}\|\beta\|^2$$

over $(\beta, b) \in \mathcal{X} \times \mathbb{R}$ under the constraints

$$g_i(\beta, b) := 1 - Y_i(\langle X_i, \beta\rangle + b) \leq 0 \quad \text{for } i = 1, \ldots, n.$$

The Lagrangian corresponding to this optimization problem takes the form

$$\mathcal{L}(\beta, b, \lambda) := f(\beta, b) + \lambda^T g(\beta, b) = \frac{1}{2}\|\beta\|^2 + \sum_{i=1}^{n} \lambda_i \left(1 - y_i(\langle x_i, \beta\rangle + b)\right).$$

This *quadratic programming problem* may be expressed as

$$\min_{(\beta,b)\in\mathcal{X}\times\mathbb{R}} \max_{\lambda\geq 0} \mathcal{L}(\beta, b, \lambda), \tag{1.6}$$

i.e. we are looking for a saddle point of the Lagrangian.

We may note that the term $b\sum_{i=1}^n \lambda_i y_i$ is linear in b. Thus if $\sum_{i=1}^n \lambda_i y_i \neq 0$, then the term minimizes to $-\infty$. But such values of λ (i.e. $\lambda : \sum_{i=1}^n \lambda_i y_i \neq 0$,) will not yield a maximum over λ in the maximization step of the optimization problem (1.6), thus the optimal λ satisfies

$$\sum_{i=1}^{n} \lambda_i y_i = 0.$$

Hence the variable b does not enter the criterion which is then a function of β only

$$f(\beta) = \frac{1}{2}\|\beta\|^2 - \langle \beta, \sum_{i=1}^{n} \lambda_i y_i x_i \rangle + \sum_{i=1}^{n} \lambda_i. \tag{1.7}$$

Function f attains its minimum for β (as a function of λ) given by

$$\beta_\lambda = \sum_{i=1}^{n} \lambda_i y_i x_i.$$

Plugging this value into the criterion function (1.7) yields

$$\sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j \langle x_i, x_j \rangle.$$

Maximizing this function over $\lambda \geq 0$ gives an optimal λ^*. Thus the optimal $\beta^* = \beta_{\lambda^*}$. By the Kuhn-Tucker theorem, the saddle point $(\beta^*, b^*, \lambda^*)$ satisfies, for every $i = 1, \ldots, n$

$$\lambda_i^* \left(1 - y_i(\langle x_i, \beta \rangle + b^*)\right) = 0.$$

The Lagrange multiplier λ_i^* can be non-zero if and only if $\langle x_i, \beta \rangle + b = \pm 1$, i.e. the point x_i is a point nearest to the hyperplane. Such points x_i are called *support points* of the support vector machine. The optimal β^* is a linear combination of such x_i, given by $\beta_{\lambda^*} = \sum_{i=1}^{n} \lambda_i^* y_i x_i$. The support points are also illustrated in figure 1.3.

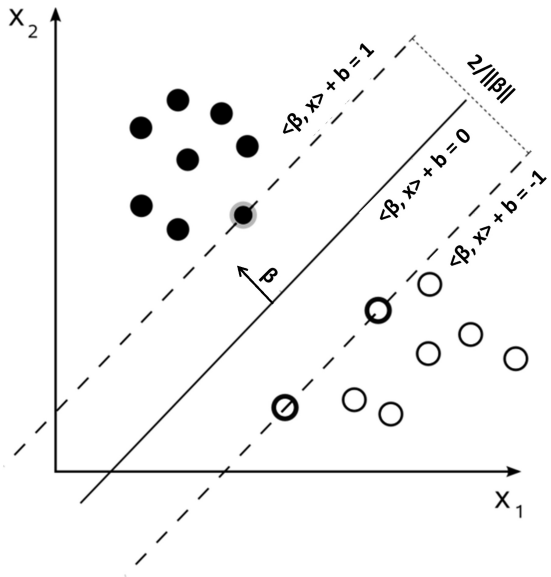


Figure 1.3: Support vector machine in $\mathcal{X} = \mathbb{R}^2$. The gap between the data sets corresponding to the maximum margin hyperplane is $2/\|\beta\|$. The support vectors are labeled with a bold circle. They lie on the hyperplanes $\langle \beta, x \rangle + b = \pm 1$. [17]

1.2.1 Soft Margin SVM

A support vector machine exists if and only if the sets $\{X_i : Y_i = 1\}$ and $\{X_i : Y_i = -1\}$ are linearly separable. Moreover, even in the separable case it is fruitful to include a tuning parameter. The *soft margin support vector machine* allows for the training points to be misclassified. It is defined as the hyperplane $H_{\beta,b}$ given by (β, b) obtained from minimizing, for a given nonnegative tuning parameter λ,

$$\lambda^2\|\beta\|^2 + \frac{1}{n}\sum_{i=1}^{n}\xi_i$$

over $(\beta, b, \xi) \in \mathcal{X} \times \mathbb{R} \times \mathbb{R}^n$ under the constraints

$$\xi_i \geq 0, \quad 1 - \xi_i \leq Y_i(\langle X_i, \beta\rangle + b) \quad i = 1, \ldots, n.$$

A value $\xi_i > 1$ allows the point X_i to fall on the "wrong" side of the hyperplane, and a value of $\xi_i \in (0, 1)$ is said to allow the point X_i to fall "within the margin". This optimization problem always has a solution. Misclassification and points within the margin are discouraged by inclusion of the penalty $\frac{1}{n}\sum_i \xi_i$ in the criterion. Lower values of the tuning constant λ increase the influence of this penalty.

The Lagrangian for this problem takes the form

$$\mathcal{L}(\beta, b, \lambda, \mu) := \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\lambda_i\left(1 - \xi_i - y_i(\langle x_i, \beta\rangle + b)\right) = \mu^T\xi.$$

We may proceed similarly as in the separable case and solve the optimization problem by expressing it in the following form

$$\min_{(\beta,b,\xi)\in\mathbb{H}\times\mathbb{R}\times\mathbb{R}^n}\ \max_{(\lambda,\mu)\in\mathbb{R}^n_+\times\mathbb{R}^n_+} \mathcal{L}(\beta, b, \lambda, \mu).$$

This procedure will yield a saddle point $(\beta^*, b^*, \xi^*, \lambda^*, \mu^*)$ of the Lagrangian.

But we are interested in expressing the problem in a slightly different form that is often studied in statistics. First observe that the constraints can also be written as $\xi_i \geq (1 - Y_i(\langle X_i, \beta\rangle + b))_+$ for $i = 1, \ldots, n$. The constrained minimization of $\lambda^2\|\beta\|^2 + \frac{1}{n}\sum_{i=1}^n \xi_i$ over $(\beta, b, \xi) \in \mathcal{X} \times \mathbb{R} \times \mathbb{R}^n$ is therefore equivalent to minimizing over $(\beta, b) \in \mathcal{X} \times \mathbb{R}$

$$\frac{1}{n}\sum_{i=1}^{n}\left(1 - Y_i(\langle X_i, \beta\rangle + b)\right)_+ + \lambda\|\beta\|^2. \tag{1.8}$$

This is obvious from the fact that $\lambda^2\|\beta\|^2 + \frac{1}{n}\sum_i \xi_i$ can be minimized in two steps: first with respect to ξ and next with respect to (β, b); the minimum over ξ is assumed at the boundary of the constraint. The formulation (1.8) will be important later on, let us now only remark that it is a special case of a general formulation known as *penalized empirical contrast procedure*. This observation will prove to be useful to analyze the rates of convergence of the SVM. But first we generalize the support vector machine by the "kernel trick."

1.2.2 The Kernel Trick

We may note that the procedure that gave us the optimal solution to the soft margin SVM in the previous section depended only on the inner products of the data points $x_i, i = 1, \ldots, n$. Thus even for very complex inputs, the computations only involve a matrix of dimensions determined by the size of the training set n. This allows us to use even very high-dimensional, or even infinite-dimensional, Hilbert spaces. This enables us to make use of the following idea of transforming the data into a higher dimensional space.

Suppose that an input x in an arbitrary set $\mathcal{X}$ can be mapped into a Hilbert space H by a map $\phi : \mathcal{X} \to H$. The support vector machine can then be applied to the data $(\phi(X_i), Y_i)$ and finds a hyperplane in H parametrized by a pair $(\beta^*, b^*) \in H \times \mathbb{R}$. A new instance x is then classified according to the sign of $\langle \phi(x), \beta^* \rangle + b^*$. The *kernel trick* is that this procedure can be implemented without explicitly defining the *feature map* ϕ, but solely in terms of the inner products

$$K(u, v) := \langle \phi(u), \phi(v) \rangle_H \quad u, v \in \mathcal{X}. \tag{1.9}$$

The *kernel function* $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ in 1.9 is clearly symmetric and positive definite (in the sense that the matrix $(K(x_i, x_j))_{i,j=1,\ldots,n}$ is nonnegative definite for every finite set $x_1, \ldots, x_n \in \mathcal{X}$). It can be shown that symmetricity and positive definiteness also characterize a kernel. That is, any symmetric positive definite function can be represented in the form $K(u, v) = \langle \phi(u), \phi(v) \rangle_H$ for some feature map ϕ.

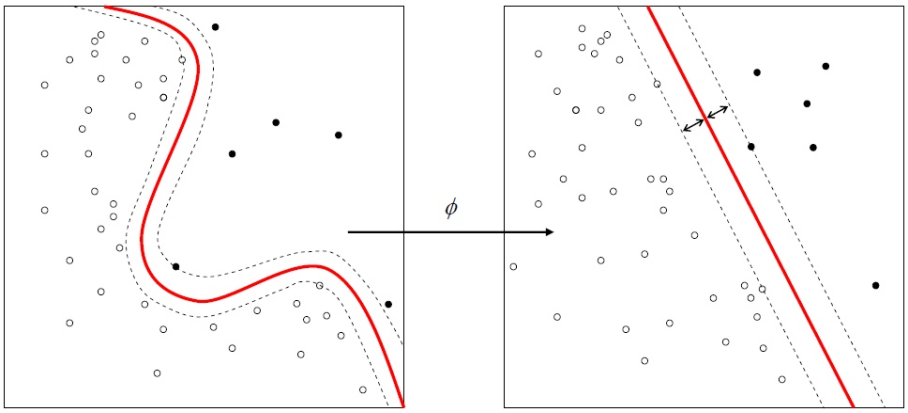


Figure 1.4: A feature map ϕ from the input space $\mathcal{X}$ to the feature space H. [17]

It is also interesting to note that kernels are in fact the same objects as the covariance functions. It is well known that every symmetric positive definite function arises as a covariance function of some centered (i.e. mean-zero) stochastic process $(G_x : x \in \mathcal{X})$

$$K(x_1, x_2) = \mathbb{E} G_{x_1} G_{x_2}.$$

Conversely, every covariance function is symmetric and positive definite.

Example 1. The covariance function of Brownian motion $(W_t)_{t \geq 0}$, given by

$$k(s, t) = \min(s, t) \quad s, t \in \mathbb{R}$$

is an example of a kernel function.

The above considerations lead us to the notion of reproducing kernel Hilbert spaces, which we now define (see e.g. [14]).

Definition 3. *A Hilbert space* $(\mathbb{H}, \langle ., . \rangle_{\mathbb{H}})$ *of functions* $h : \mathcal{X} \to \mathbb{R}$ *on an arbitrary set* $\mathcal{X}$ *is called a reproducing kernel Hilbert space if there exists a symmetric function* $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *such that the function* $y \mapsto K(x, y)$ *is an element of* $\mathbb{H}$ *for every* $x \in \mathcal{X}$ *and, for every* $h \in \mathbb{H}$ *and* $x \in \mathcal{X}$ *we have*

$$h(x) = \langle K(x, .), h \rangle_{\mathbb{H}}. \tag{1.10}$$

The following lemma (see e.g. [14]) shows that the feature space H we introduced above can be without loss of generality taken to be a *reproducing kernel Hilbert space* (RKHS) $\mathbb{H}$ and describes the support vector machine in terms of $\mathbb{H}$.

Lemma 2. *If* $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is a symmetric positive definite function on the product of an arbitrary set* $\mathcal{X}$ *with itself, then there exists a unique Hilbert space* $(\mathbb{H}, \langle \cdot, \cdot \rangle_{\mathbb{H}})$ *of functions* $h : \mathcal{X} \to \mathbb{R}$ *such that:*

(i) the map $K(x, \cdot) : \mathcal{X} \to \mathbb{R}$ *is contained in* $\mathbb{H}$ *for every* $x \in \mathcal{X}$ *and the span of these maps is dense in* $\mathbb{H}$*;*

(ii) $h(x) = \langle K(x, \cdot), h \rangle_{\mathbb{H}}$ *for every* $h \in \mathbb{H}$ *and* $x \in \mathcal{X}$*;*

(iii) $K(x, y) = \langle K(x, \cdot), K(y, \cdot) \rangle_{\mathbb{H}}$ *for every* $x, y \in \mathcal{X}$*;*

(iv) the map $\bar{\phi} : x \mapsto K(x, \cdot)$ *is* $1 - 1$ *from* $\mathcal{X}$ *to* $\mathbb{H}$ *and* $\langle \bar{\phi}(x), \bar{\phi}(y) \rangle_{\mathbb{H}} = K(x, y)$. *Furthermore, if* $K(x, y) = \langle \phi(x), \phi(y) \rangle_H$ *for some map* $\phi : \mathcal{X} \to H$ *into a Hilbert space* H*, then* $\beta^* \in H$ *minimizes*

$$\beta \mapsto \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i(\langle X_i, \beta \rangle + b))_+ + \lambda \|\beta\|_H^2$$

over H *if and only if* $f_{\beta^*} \in \mathbb{H}$ *given by* $f_{\beta^*}(x) = \langle \phi(x), \beta^* \rangle_H$ *minimizes over* $\mathbb{H}$ *the map*

$$f \mapsto \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i(f(X_i) + b))_+ + \lambda \|f\|_{\mathbb{H}}^2. \tag{1.11}$$

If we denote the functions $f + b$ by θ, add the constants to $\mathbb{H}$ and define the seminorm $\|\theta\|_{\mathbb{H}}$ to be invariant under shifts, then we can identify a "kernelized" support vector machine with the map $\theta : \mathcal{X} \to \mathbb{R}$ that minimizes

$$\theta \mapsto \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \theta(X_i))_+ + \lambda^2 \|\theta\|_{\mathbb{H}}^2. \tag{1.12}$$

This observation will play a crucial role later on.

Examples of Kernels

We now give a few examples of the most commonly used kernel functions in the case $\mathcal{X} = \mathbb{R}^p$.

- polynomial (homogeneous) $k(x_1, x_2) = x_1^T x_2$

- polynomial (inhomogeneous) $k(x_1, x_2) = (1 + x_1^T x_2)^k$
- Gaussian $k_\sigma(x_1, x_2) = \exp(-\sigma\|x_1 - x_2\|^2)$
- hyperbolic tangent $k(x_1, x_2) = \tanh(cx_1^T x_2 + c_2)$

Gaussian Kernel RKHS

In this work, we mainly focus on RKHS associated with Gaussian kernels, which are the most widely used kernels in practice [9]. We give a precise definition of the Gaussian kernel function.

Definition 4. *A Gaussian kernel function is of the form*

$$k_\sigma(x_1, x_2) := \exp\left(-\sigma^2\|x_1 - x_2\|_2^2\right)$$

where $x_1, x_2 \in \mathcal{X}$, $\sigma > 0$ is a free parameter whose inverse $1/\sigma$ is called the width of the k_σ. We shall denote the corresponding RKHS by $\mathbb{H}_\sigma(\mathcal{X})$ or simply $\mathbb{H}_\sigma$.

The Gaussian reproducing kernel Hilbert space is the Hilbert space attached to a kernel function k_σ in view of lemma 2. For a thorough analysis of the Gaussian RKHSs see e.g. [12]. We shall give a characterization of Gaussian kernel RKHS as derived in [12], but first let us fix some notation. By $L^r(\mathbb{P})$ we shall denote the space of measurable functions that are $r-$integrable with respect to the measure $\mathbb{P}$ and by $\|\cdot\|_{\mathbb{P},r}$ we denote the norm corresponding to the space $L^r(\mathbb{P})$.
The reproducing kernel Hilbert space with a Gaussian kernel is shown in [12] to be the set of real parts of the functions $h_\psi : \mathcal{X} \to \mathbb{C}$ given by

$$h_\psi : x \mapsto \int e^{i\lambda^T x}\psi(\lambda)\mathrm{d}\mu_\sigma(\lambda),$$

where ψ runs through the complex Hilbert space $L^2(\mu_\sigma)$ with RKHS norm equal to $\|\psi\|_{\mu_\sigma,2}$ and μ_σ is the measure with density relative to the Lebesgue measure given by

$$f : \lambda \mapsto \frac{\sigma^d}{2^d\pi^{d/2}}e^{\frac{-\sigma^2\|\lambda\|^2}{4}}.$$

Chapter 2

Empirical Processes And Rates of Convergence

This chapter provides brief introduction to the theory of empirical processes, which is the main tool that will allow us to study statistical properties of support vector machines. We give heuristic arguments for estimation procedures related to empirical risk minimization and analyze the type of errors that arise within these procedures. The error analysis leads us to study measures of complexity of sets of functions such as metric entropy and we relate the rate of entropy increase to the empirical process. We then cast the support vector machine as a penalized empirical minimization problem. A general result derived in [11] that provides rates of convergence of estimators in specific classes of empirical risk minimization procedures will be presented, with the aim to apply it to the support vector machine algorithm.

For two sequences $\{f_n\}_n, \{g_n\}_n$ we use the notation $f_n \lesssim g_n$ meaning that there exists a universal constant $C > 0$ such that $f_n \leq C g_n$ over a pre-specified range of values n (e.g. all n sufficiently large). We use the notation $\gtrsim$ with a similar meaning and $\sim$ when both $\lesssim$ and $\gtrsim$ hold.

In the chapters that follow, we shall often consider a supremum of uncountably many random variables. Note that this is not necessarily a random variable. Some results in the following are only true if certain measurability conditions are satisfied. We shall ignore the technicalities of measurability (see [11] for full details), although sometimes we write $\mathbb{E}^*$ or $\mathbb{P}^*$ for expectations and probabilities, where the asterisk refers to "outer" measure.

2.1 Relationship to Empirical Processes

The motivation for studying empirical measures is that in reality, it is often impossible to know the true underlying probability measure $\mathbb{P}$. We collect observations $X_1, X_2, \ldots, X_n$ and we can estimate $\mathbb{P}$, or its corresponding distribution function F by means of *empirical measure* or *empirical distribution function*, respectively.

Assume that we have an i.i.d. sequence of random variables $X_1, \ldots, X_n$ in a measurable space $(\Omega, \mathcal{S})$. Let $\mathbb{P}_n$ denote the *empirical measure* induced by the sequence

$X_1, \ldots, X_n$, i.e. for any $A \in \mathcal{S}$ define

$$\mathbb{P}_n(A) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_A(X_i),$$

where $\mathbf{1}_A$ is the indicator function of the set A. The empirical measure $\mathbb{P}_n$ maps measurable functions $f : \Omega \to \mathbb{R}$ from a given set $\mathcal{F}$ to their *empirical mean* by a map from $\mathcal{F} \to \mathbb{R}$ given by

$$f \mapsto \mathbb{P}_n f = \int_\Omega f \mathrm{d}\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

Let $\mathbb{P}$ denote the true distribution of the X_is. Then for a fixed measurable function f, $\mathbb{P}_n f$ is a random variable with mean Pf and variance $\frac{1}{n} P\left(f - Pf\right)^2$.

The centered and scaled version of the map $f \mapsto \mathbb{P}_n f$ is called the $\mathcal{F}$-indexed *empirical process* $\mathbb{G}_n$ given by

$$f \mapsto \mathbb{G}_n f := \sqrt{n}(\mathbb{P}_n - P)f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(f(X_i) - Pf\right).$$

Example 2. *For $\Omega := \mathbb{R}$ and $\mathcal{F}$ the set of all intervals of type $\mathbf{1}_{(-\infty,x]}$, $x \in \mathbb{R}$, the empirical measure indexed by $\mathcal{F}$ can be identified with the empirical distribution function $\mathbb{F}_n$ given by*

$$x \mapsto \mathbb{F}_n(x) := \mathbb{P}_n \mathbf{1}_{(-\infty,x]} = \sum_{i=1}^{n} \mathbf{1}_{X_i \leq x}.$$

Remark 1. For the empirical measure $\mathbb{P}_n$, we shall use the shorthand $\mathbb{P}_n f = \int f \mathrm{d}\mathbb{P}_n$ (to prevent possible confusion between $P_n f$ and Pf if we used the notation $P_n f = \int f \mathrm{d}\mathbb{P}_n$ and $Pf = \int f \mathrm{d}\mathbb{P}$).

The relationship of learning algorithms and empirical processes has already been suggested in section 1.1.1. Recall that we are mainly interested in estimating the risk Pm_θ (assume now a given loss function m_θ) generated by candidate classifiers $\theta \in \Theta$ and we would like to choose a classifier $\hat{\theta}_n$ such that this risk is minimal. However, the true risk cannot be directly observed since it depends on the underlying unknown distribution $\mathbb{P}$. One way to make a statement about this quantity is to say how it relates to the *empirical risk* $\mathbb{P}_n m_\theta$ of a classifier θ given by

$$\mathbb{P}_n m_\theta := \frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i, Y_i)$$

for a loss function m given a data set $(X_1, Y_1), .., (X_n, Y_n)$. In general, learning algorithms use the empirical risk $\mathbb{P}_n m_\theta$ as a criterion function to minimize (maximize) instead of the true risk.

The criterion function $\mathbb{P}_n m_\theta$ converges to the asymptotic criterion function Pm_θ as $n \to \infty$. We may thus expect that the minimizers $\hat{\theta}_n$ of $\mathbb{P}_n m_\theta$ converge to the minimizer θ_0 of $\theta \mapsto Pm_\theta$. If we introduce some metric d, we may be interested in the rate

of convergence of $d(\hat{\theta}_n, \theta_0)$ as $n \to \infty$.

Let us now finally formalize what we actually mean by the *rate of convergence* of a random variable. The definition is analogous to the well known "big-O" notation that is used to describe limiting behaviour of a deterministic function. If we were only interested in the limiting behaviour (as $n \to \infty$) of the expectation of a sequence of random variables, say $\{Z_n\}_n$, we write that $\mathbb{E}|Z_n| = O(\delta_n)$ for some deterministic function δ_n of n meaning that there exists a constant M such that

$$\mathbb{E}|Z_n| \leq M\delta_n$$

for all n sufficiently large. Similarly, for the random variable Z_n, we write $Z_n = O_P^*(\delta_n)$ if for all $\varepsilon > 0$ there exists a constant $M \in \mathbb{R}$ such that

$$\mathbb{P}^* \left(\frac{|Z_n|}{\delta_n} > M \right) < \varepsilon$$

for all n sufficiently large. E.g. $O_P(1)$ means that a sequence is bounded in probability.

For the sake of clarity, we shall set the metric d equal to the quantity of our interest, i.e.

$$d^2(\hat{\theta}_n, \theta_0) := Pm_{\hat{\theta}_n} - Pm_{\theta_0}$$

that will measure the distance between $\hat{\theta}_n$ and θ_0. Let us only remark that the results that will follow remain true with any chosen metric, or even for an arbitrary map from $\Theta \to [0, \infty)$ (see [11]), under an additional assumption that the asymptotic criterion decreases quadratically when θ moves away from θ_0, i.e.

$$P(m_\theta - m_{\theta_0}) \geq d^2(\theta, \theta_0) \tag{2.1}$$

Before we introduce results that give actual rates of convergence to the Bayes risk for specific estimation tasks, we will provide a brief insight into the sources of errors that are inherent to empirical risk minimization problems. We are mainly interested in how much the risk generated by a data-dependent classifier $\hat{\theta}_n$ differs from the Bayes risk, i.e. the so called *excess risk*

$$Pm_{\hat{\theta}_n} - Pm_{\theta_0}.$$

An estimation algorithm in general chooses its output $\hat{\theta}_n$ from a class of functions $\mathcal{F} \subset \Theta$, which is not necessarily equal to the whole space Θ. Note that this implies that θ_0 does not necessarily lie in $\mathcal{F}$. If we denote $\theta^* = \arg\inf_{\theta \in \mathcal{F}} Pm_\theta$, i.e. θ^* is the best function that can be chosen from the model $\mathcal{F}$(with respect to risk minimization), then we may decompose the excess risk as follows

$$Pm_{\hat{\theta}_n} - Pm_{\theta_0} = (Pm_{\theta^*} - Pm_{\theta_0}) + (Pm_{\hat{\theta}_n} - Pm_{\theta^*}). \tag{2.2}$$

The first term, $Pm_{\theta^*} - Pm_{\theta_0}$, is called the *approximation error* and it measures how well can functions or classifiers in $\mathcal{F}$ approximate the target Bayes risk (it would be zero if $\theta_0 \in \mathcal{F}$). Note that the approximation error does not depend on the data, it is solely a property of $\mathcal{F}$. The second term is called the *estimation error* and it measures how close is $\hat{\theta}_n$ to the best possible choice $\theta^* \in \mathcal{F}$ in terms of the theoretical risk P.

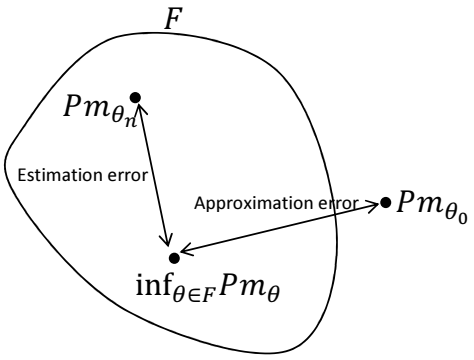


Figure 2.1: Schematic diagram of approximation and estimation errors.

Note that there is some trade-off between the approximation and the estimation error. When the class of functions $\mathcal{F}$ is large, $\inf_{\theta \in \mathcal{F}} Pm_\theta$ may be close to Pm_{θ_0}, but the estimation error may be large. On the other hand, if the class $\mathcal{F}$ is small, there could be a gap between $\inf_{\theta \in \mathcal{F}} Pm_\theta$ and Pm_{θ_0} so the approximation error could be substantial.

Results that give us some information about the rate of convergence to the Bayes risk, i.e. the rate of convergence of the random variable $P(m_{\hat{\theta}_n} - m_\theta)$ to zero, typically assume that we are able to obtain some bounds of the supremum of the empirical process $\mathbb{G}_n(m_\theta - m_{\theta_0})$ over a given class of functions θ. How to obtain such bounds will be discussed in the next section. Let us now only remark that these bounds are strongly related to measures of complexity of spaces of functions, such as metric entropy.

2.2 Entropy Bounds

While the approximation error is a property of the class $\mathcal{F}$ only, the two types of error that depend on data $Pm_{\hat{\theta}_n} - Pm_{\theta^*}$, i.e. the estimation error, and $\mathbb{P}_n m_{\hat{\theta}_n} - Pm_{\hat{\theta}_n}$ which represents the error due to the estimation of the risk from the data, can be reduced by increasing the sample size. As both errors involve random quantities ($\mathbb{P}_n m_{\hat{\theta}_n}$ and $Pm_{\hat{\theta}_n}$), statistical learning theory mainly aims at their probabilistic bounds. That is, how to bound the tail probabilities of the differences in risk. The following inequalities suggest that the two types of error can be dealt with as one by examining $\sup_{\theta\in\mathcal{F}} |\mathbb{P}_n m_\theta - Pm_\theta|$.

$$\begin{aligned} |\mathbb{P}_n m_{\hat{\theta}_n} - Pm_{\hat{\theta}_n}| &\leq \sup_{\theta\in\mathcal{F}} |\mathbb{P}_n m_\theta - Pm_\theta| \\ Pm_{\hat{\theta}_n} - \inf_{\theta\in\mathcal{F}} Pm_\theta &\leq 2\sup_{\theta\in\mathcal{F}} |\mathbb{P}_n m_\theta - Pm_\theta|. \end{aligned}$$

While the first inequality trivially holds, the second requires a bit of manipulation and we omit the proof.

In general, bounds on expressions of the type

$$\mathbb{E}\sup_{f\in\mathcal{F}} |\mathbb{G}_n f| \tag{2.3}$$

for a given class of functions $\mathcal{F}$ can often be obtained by studying specific measures of complexity of a class of functions. One such measure is a *covering number*, which essentially tells us how many "balls" of a given radius (in a (semi)metric space) are required to cover a set.

Definition 5. *Let (T, d) be an arbitrary semimetric space. Then the covering number $N(\varepsilon, T, d)$ is the minimal number of balls of radius ε needed to cover T. The corresponding entropy number is the logarithm of the covering number.*

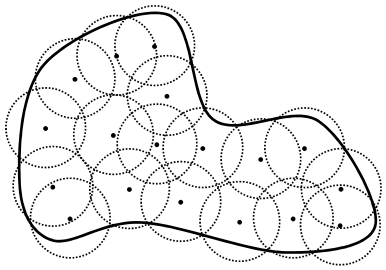


Figure 2.2: Illustration of a possible covering of a set.

First we may note that covering number is a decreasing function of ε and typically tends to ∞ for $\varepsilon \downarrow 0$. If the covering number is finite for every $\varepsilon > 0$, we say that the semimetric space T is *totally bounded*. The upper bounds on (2.3) depend on the rate at which the entropy numbers for T taken to be a class of functions $\mathcal{F}$ grow as $\varepsilon \downarrow 0$.

For the case of Gaussian kernel RKHS $\mathbb{H}_\sigma$ and the uniform norm, the entropy number $\log N(\varepsilon, B_{\mathbb{H}_\sigma}, \|.\|_\infty)$, where $B_{\mathbb{H}} = \{\theta \in \mathbb{H} : \|\theta\|_{\mathbb{H}} \leq 1\}$ denotes the unit ball in $\mathbb{H}$ may be bounded (with respect to the uniform norm) as derived in [12].

Lemma 3. *Assume that $\mathcal{X} \subset \mathbb{R}^d$. Let $\mathbb{H}_\sigma(\mathcal{X})$ be a reproducing kernel Hilbert space corresponding to the Gaussian kernel k_σ with width $1/\sigma$. Then for every $\varepsilon < \frac{1}{2}$,*

$$\log N(\varepsilon, B_{\mathbb{H}_\sigma}, \|.\|_\infty) \lesssim \sigma^d \left(\log \frac{1}{\varepsilon}\right)^{1+d}.$$

For most classes of functions $\mathcal{F}$ we might be interested in, the covering number grows to infinity as $\varepsilon \to 0$. We are mostly interested in the speed of this growth, which can be measured in terms of the *entropy integral*

$$J(\delta, \mathcal{F}, L^2) = \int_0^\delta \sup_Q \sqrt{1 + \log N(\varepsilon, \mathcal{F}, L^2(Q))} d\varepsilon$$

where the supremum is taken over all discrete probability measures Q.

An alternative way of measuring the size of a class of functions may be in terms of *bracketing numbers.*

Definition 6. *Given two functions l and u, the bracket $[l, u]$ is the set of all functions f with $l \leq f \leq u$. An $\varepsilon-$bracket in $L^r(\mathbb{P})$ is a bracket $[l, u]$ with $P|u - l|^r < \varepsilon^r$. The bracketing number $N_{[]}(\varepsilon, \mathcal{F}, L^r(\mathbb{P}))$ is the minimum number of $\varepsilon-$brackets needed cover $\mathcal{F}$. The entropy with bracketing is the logarithm of the bracketing number. (The bracketing functions l and u must have finite $L^r(\mathbb{P})$-norms, but need not belong to $\mathcal{F}$.)*

Bracketing numbers very much resemble covering numbers and they are decreasing functions of ε as well. The advantage of bracketing numbers over covering numbers is that we gain pointwise control over the function f: $l(x) \leq f(x) \leq u(x)$ for every x. The $L^r(\mathbb{P})$ balls give only control over the integral, not bounds on the function itself.

One may note that a bracket $[l, u]$ of size ε can be covered with a ball with center $\frac{l+u}{2}$ of radius $\varepsilon/2$, since $|f - \frac{u+l}{2}| \leq |u - \frac{u+l}{2}| = |l - \frac{u+l}{2}|$ implies $\|f - \frac{l+u}{2}\|_{P,r} \leq \|u - \frac{l+u}{2}\|_{P,r} < \varepsilon/2$. We thus obtain the following relationship between covering and bracketing numbers

$$N(\varepsilon/2, \mathcal{F}, L^r(\mathbb{P})) \leq N_{[]}(\varepsilon, \mathcal{F}, L^r(\mathbb{P})).$$

Note that this also implies $N(\varepsilon, \mathcal{F}, L^r(\mathbb{P})) \leq N_{[]}(\varepsilon, \mathcal{F}, L^r(\mathbb{P}))$ by the decreasingness of a covering number in ε. Hence bracketing numbers are in general bigger than covering numbers.

To measure the speed of growth of bracketing numbers for $\varepsilon \downarrow 0$ we define the *bracketing integral*

$$J_{[]}(\delta, \mathcal{F}, L^2(P)) = \int_0^\delta \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}, L^2(P))} d\varepsilon.$$

The integrand is a decreasing function of ε, since the bracketing number decreases as a function of ε. Thus the convergence of the integral depends only on the size of the bracketing numbers for $\varepsilon \downarrow 0$. We may note that since the integral $\int_0^1 \varepsilon^{-r} d\varepsilon$ converges for $r < 1$ and diverges for $r \geq 1$, then (roughly speaking) bracketing integral is finite if the growth of bracketing entropies is slower than ε^{-2}.

Theorem 15 in [11] gives us a bound on the supremum of the empirical process over a set $\mathcal{F}$ of uniformly bounded measurable functions and thus relates the complexity measures of spaces of functions to empirical processes.

Theorem 1. *Let $\mathcal{F}$ be a set of measurable functions $f : \mathcal{X} \to \mathbb{R}$ such that $|f(x)| \leq F(x) \leq 1$ for all $x \in \mathcal{X}$, and such that for some $\delta \in (0,1)$ and for every $f \in \mathcal{F}$ it holds $Pf^2 < \delta^2 PF^2$. Then*

$$\mathbb{E}_P^* \sup_{f \in \mathcal{F}} |\mathbb{G}_n f| \lesssim J(\delta, \mathcal{F}, L^2(P)) \left(1 + \frac{J(\delta, \mathcal{F}, L^2(P))}{\delta^2 \sqrt{n} \|F\|_{P,2}} \|F\|_{P,2}\right).$$

A similar result may be obtained for bracketing numbers (Theorem 23 in [11]).

Theorem 2. *Let $\mathcal{F}$ be a set of measurable functions $f : \mathcal{X} \to \mathbb{R}$ such that for some $\delta \in \mathbb{R}$ and for every f it holds $Pf^2 < \delta^2$ and $\|f\|_\infty \leq M$ for a constant M. Then*

$$\mathbb{E}_P^* \sup_{f \in \mathcal{F}} |\mathbb{G}_n f| \lesssim J_{[]}(\delta, \mathcal{F}, L^2) \left(1 + \frac{J_{[]}(\delta, \mathcal{F}, L^2)}{\delta^2 \sqrt{n}} M\right).$$

We will see in the next section how one applies entropy measures in estimation problems. Specifically for the SVM, the bound on (2.3) may be obtained by finding bounds on entropy numbers of the associated RKHS that a SVM uses. Then we are able to obtain corresponding entropy integrals and thus bound the supremum of the related empirical process, which is essential for the techniques that guarantee rates of convergence in empirical risk minimization procedures.

2.3 Rates of Convergence

Consider a general empirical risk minimization procedure that gives output

$$\hat{\theta}_n := \arg\min_{\theta\in\Theta} \mathbb{P}_n m_\theta.$$

We finally give a general result derived in [11] that gives rates of convergence to the Bayes risk provided that we are able to bound the supremum of the empirical process $\mathbb{G}_n$ indexed by the class of functions $\mathcal{F}_\delta := \{m_\theta - m_{\theta_0} : \theta \in \Theta, P(m_\theta - m_{\theta_0}) \leq \delta^2\}$ by some function $\phi_n(\delta)$.

Theorem 3. *Suppose that for every $\delta > 0$*

$$\mathbb{E}^* \sup_{\theta\in\Theta: Pm_\theta - Pm_{\theta_0} \leq \delta^2} |\mathbb{G}_n(m_\theta - m_{\theta_0})| \lesssim \phi_n(\delta)$$

for a function $\phi_n : [0,\infty) \to \mathbb{R}$ such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$. Then the minimizer $\hat{\theta}_n$ of $\theta \mapsto \mathbb{P}_n m_\theta$ satisfies

$$Pm_{\hat{\theta}_n} - Pm_{\theta_0} = O_P^*(\delta_n)$$

for some δ_n such that $\phi_n(\delta_n) \leq \sqrt{n}\delta_n^2$.

Proof. Fix $n \in \mathbb{N}$. The set Θ may be written as a union over $j \in \mathbb{Z}$ of disjoint sets of the form

$$S_{n,j} := \{\theta \in \Theta : \delta_n 2^{j-1} < (P(m_\theta - m_{\theta_0}))^{1/2} \leq 2^j \delta_n\}.$$

If the minimizer $\hat{\theta}_n$ of $\theta \mapsto \mathbb{P}_n m_\theta$ belongs to $S_{n,j}$ then by the definition of $S_{n,j}$ it holds $Pm_{\hat{\theta}_n} > Pm_{\theta_0} + 2^{2j-2}\delta_n^2$. By the definition of $\hat{\theta}_n$ it holds that $\mathbb{P}_n m_{\hat{\theta}_n} \leq \mathbb{P}_n m_{\theta_0}$. Combining the two inequalities gives that $\hat{\theta}_n \in S_{n,j}$ implies

$$\mathbb{P}_n m_{\hat{\theta}_n} - Pm_{\hat{\theta}_n} \leq \mathbb{P}_n m_{\theta_0} - Pm_{\theta_0} - 2^{2j-2}\delta_n^2,$$

or, rearranging,

$$\mathbb{G}_n(m_{\theta_0} - m_{\hat{\theta}_n}) \geq \sqrt{n}2^{2j-2}\delta_n^2.$$

We conclude that

$$\mathbb{P}(\hat{\theta}_n \in S_{n,j}) \leq \mathbb{P}\left(\sup_{\theta\in S_{n,j}} \mathbb{G}_n(m_{\theta_0} - m_\theta) \geq \sqrt{n}2^{2j-2}\delta_n^2\right) \leq \frac{\phi_n(2^j\delta_n)}{\sqrt{n}2^{2j-2}\delta_n^2},$$

where the last inequality follows by Markov's inequality and by the definition of ϕ_n. Since $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ we obtain $\phi_n(2^j\delta_n) \leq 2^{j\alpha}\phi_n(\delta_n)$. Then for any $M \in \mathbb{N}$ we get

$$\mathbb{P}((P(m_{\hat{\theta}_n} - m_{\theta_0}))^{\frac{1}{2}} > 2^M\delta_n) \leq \sum_{j=M}^{\infty} \frac{2^{j\alpha}\phi_n(\delta_n)}{\sqrt{n}2^{2j-2}\delta_n^2} \lesssim \sum_{j=M}^{\infty} 2^{j(\alpha-2)},$$

where the last inequality follows by the definition of δ_n. Note that the remainder of the series in the last expression converges (to zero) for $M \to \infty$ since $\alpha < 2$. □

Example 3 (**Euclidean parameter space**)**.** Suppose that $\Theta \subset \mathbb{R}^d$ and that for every θ, θ' and $x \in \mathcal{X}$

$$|m_\theta(x) - m_{\theta'}(x)| \leq M\|\theta - \theta'\|_2,$$

for a constant $M \in \mathbb{R}$. Then the class of functions $\mathcal{M}_\delta := \{m_\theta - m_{\theta_0} : \|\theta - \theta_0\|_2 < \delta\}$ is upper bounded by δM and hence the bracketing numbers satisfy

$$N_{[]}(\varepsilon M, \mathcal{M}_\delta, L^2(P)) \leq N(\varepsilon, \{\theta : \|\theta - \theta_0\|_2 < \delta\}, \|\cdot\|_\infty) \lesssim \left(\frac{\delta}{\varepsilon}\right)^d.$$

The last inequality is the consequence of the fact that (1) the $\varepsilon-$covering number of a ball of radius δ equals the $\frac{\varepsilon}{\delta}-$covering number of a unit ball, (2) the $\varepsilon-$covering number of a unit ball in $\mathbb{R}^d$ equals $\frac{1}{\varepsilon^d}$ up to a constant.

A bound on $\sup_{f \in \mathcal{M}_\delta} |\mathbb{G}_n f|$ may then be obtained in view of section 2.2 on entropy bounds as follows

$$\sup_{f \in \mathcal{M}_\delta} |\mathbb{G}_n f| \lesssim \int_0^{\delta M} \sqrt{\log\left(\frac{\delta}{\varepsilon}\right)^d}\, d\varepsilon \lesssim \delta.$$

Thus $\phi_n(\delta)$ in Theorem 3 can be taken equal to δ and the (worst) solution to the inequality $\phi_n(\delta_n) \leq \sqrt{n}\delta_n^2$ is given by $\delta_n = \frac{1}{\sqrt{n}}$. Thus the rate of convergence is $\frac{1}{\sqrt{n}}$ provided that $P(m_\theta - m_{\theta_0}) \geq c\|\theta - \theta_0\|_2^2$ for every $\theta \in \Theta$ in view of (2.1). □

In some cases, minimization of a criterion over the full space Θ may not be a good idea. For instance, consider fitting a function $\theta : [0,1] \to \mathbb{R}$ to a set of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ by least squares, i.e. we minimize

$$\theta \mapsto \frac{1}{n}\sum_{i=1}^n (Y_i - \theta(X_i))^2.$$

If the space Θ consists of all measurable functions $\theta : \mathcal{X} \to \mathbb{R}$, then the minimum is zero; the maximizer being any function $\theta : \mathcal{X} \to \mathbb{R}$ that interpolates the data points exactly, i.e. $\theta(X_i) = Y_i$ for $i = 1, \ldots, n$. This is usually not a very good estimator, which overfits the data. To prevent overfitting, we need to impose quantitative restrictions on the class of functions we minimize over, e.g. in regression one often considers the space of linear functions.

Alternative to restricting minimization (maximization), overfitting can be solved by add-ing a penalty function to the criterion function. Assume now that we want to minimize the criterion function $\mathbb{P}_n m_\theta$ and we add a *penalty function* $J : \Theta \to [0, \infty)$. We get a so-called *penalized minimum contrast estimator* that minimizes the criterion function

$$\theta \mapsto \mathbb{P}_n m_\theta + \lambda_n^2 J^2(\theta)$$

over a given class of functions, where m is a given loss function and J is a penalty function.

The purpose of the penalty is to decrease the criterion value when θ is far from θ_0. Thus we discourage those θs which give a high value $\mathbb{P}_n m_\theta$ but are not near the true parameter. This could happen, for instance, by overfitting. The trade-off between the

empirical criterion $\mathbb{P}_n m_\theta$ and the size of the penalty $J(\theta)$ is moderated by the *smoothing parameter* λ_n. Smoothing parameter decreases as n increases and asymptotically the penalty becomes inactive.

Recall that in section 1.2 we rewrote the support vector machine algorithm as the minimizer of the criterion

$$\theta \mapsto \frac{1}{n}\sum_{i=1}^{n}(1 - Y_i\theta(X_i))_+ + \lambda^2\|\theta\|_{\mathbb{H}}^2 \tag{2.4}$$

over a reproducing kernel Hilbert space $\mathbb{H}$. This is the formulation that exactly fits the penalized empirical minimization problem, with loss function equal to the hinge loss and penalty equal to the norm $\|.\|_{\mathbb{H}}$, i.e.

$$\begin{aligned} m_\theta(x,y) &:= (1 - y\theta(x))_+ \\ J(\theta) &:= \|\theta\|_{\mathbb{H}}. \end{aligned}$$

From now on, we will work with hinge loss m as our loss function, and the following result by Van der Vaart and Wellner [11] is already stated to match the needs of SVM formulation.

Theorem 4. *Let $\lambda_n > 0$. Suppose that for every $n \in \mathbb{N}$ and $\delta > 0$*

$$\mathbb{E}^* \sup\left\{\sqrt{n}|\mathbb{G}_n m_\theta - \mathbb{G}_n m_{\theta_0}| \ : \ \theta \in \mathbb{H}, Pm_\theta - Pm_{\theta_0} \le \delta^2, \|\theta\|_{\mathbb{H}} < \delta/\lambda_n\right\} \lesssim \phi_n(\delta) \tag{2.5}$$

for an increasing function $\phi_n : (0,\infty) \to \mathbb{R}$ such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$. Let $\theta_n \in \mathbb{H}$ and let δ_n satisfy

$$\begin{aligned} \phi_n(\delta_n) &\le \sqrt{n}\delta_n^2 && (2.6)\\ \delta_n^2 &\ge Pm_{\theta_n} - Pm_{\theta_0} && (2.7)\\ \delta_n &\ge \lambda_n\|\theta_n\|_{\mathbb{H}} && (2.8) \end{aligned}$$

then

$$Pm_{\hat{\theta}_n} - Pm_{\theta_0} = O_P^*(\delta_n^2).$$

The theorem thus states that the rate of convergence of $P(m_{\hat{\theta}_n} - m_{\theta_0})$, where $\hat{\theta}_n$ is the minimizer of the criterion function (2.4) and θ_0 is the minimizer of the true risk Pm_θ is given by the approximation error

$$\inf_\theta(P(m_\theta - m_{\theta_0}) + \lambda^2\|\theta\|_{\mathbb{H}}^2)$$

which corresponds to (2.7) and (2.8), and the square of the solution δ_n to the equation $\phi_n(\delta) \le \sqrt{n}\delta_n^2$.

In the first chapter, we have already studied the smallest achievable risk associated with 0-1 loss function, which is attained by the Bayes classifier θ_0. The same is true of the hinge loss function m_θ. This forms the content of the following lemma.

Lemma 4. *The risk associated with the hinge loss function m_θ is minimized over the set Θ of all measurable maps $\theta : \mathcal{X} \to \mathbb{R}$ by*

$$\theta_0(x) \ = \ 2\mathbf{1}_{\eta(x) > \frac{1}{2}} - 1 \quad \textit{where} \quad \eta(x) := \mathbb{P}(Y = 1|X = x).$$

Proof. Conditioning on the input we may write

$$Pm_\theta \;=\; \int (\eta(x)(1-\theta(x))_+ + (1-\eta(x))(1+\theta(x))_+) d\mathbb{P}_X(dx).$$

We minimize the integrand separately for each x. The integrand is minimized over $\theta(x) \in \mathbb{R}$ for the case $\eta(x) = 0$ at every $\theta(x) \leq -1$, for $\eta(x) \in (0, \frac{1}{2})$ it is minimal at $\theta(x) = -1$, for $\eta(x) = 1/2$ minimal at every $\theta(x) \in [-1, 1]$, for $\eta(x) \in (1/2, 1)$ minimal at $\theta(x) = 1$ and for $\eta(x) = 1$ minimal at every $\theta(x) \geq 1$. The function $\theta_0(x) = 2\mathbf{1}_{\eta(x)>\frac{1}{2}} - 1$ is a possible choice. □

Chapter 3

Convergence Rates for Support Vector Machines

In the first part of this chapter, we present and discuss recent results on convergence rates for SVM. Steinwart and Scovel in [9] derive fast rates for the case of Gaussian kernel RKHS under Tsybakov noise assumption and a geometric noise assumption they introduce. The second result proved by Blanchard, Bousquet and Massart in [2] is a very general model selection procedure with penalization. Their bound on excess risk admits an arbitrary choice of RKHS, thus no actual bound is provided for the approximation error. The second part of this chapter aims at obtaining convergence rates for the case of Gaussian kernel RKHS, first deriving the rates following the approach of Blanchard, Bousquet and Massart and next the approach of Van der Vaart and Wellner in [11]. Finally, we rederive the same rates using an entropy bound as in [9].

3.1 Rates for Support Vector Machines with a Gaussian Kernel

In [9] Steinwart and Scovel treat SVMs in the special case of Gaussian kernel RKHS and assume $\mathcal{X} \subset \mathbb{R}^d$ which is to be assumed throughout this section.

3.1.1 Assumptions

We first introduce Tsybakov's noise assumption, which describes the amount of "noise" in the labels. (By a noise-free distribution we understand $\mathbb{P}(Y = 1|X = x) \in \{0, 1\}$). This may be achieved in terms of the function

$$\min\{\eta(x), 1 - \eta(x)\} = \frac{1}{2} - |\eta(x) - \frac{1}{2}|.$$

In other words, we are interested in the behaviour of the function $\eta(x)$ around the critical level $\frac{1}{2}$. Observe that we may equivalently consider the function $|2\eta - 1|$. In regions where the function $|2\eta - 1|$ is close to 1, there is only a small amount of noise, whereas function values close to 0 only occur in regions with a high level of noise. The following assumption describes the size of the regions with a high level of noise.

Assumption 1. *(**Amount of noise**) Let $0 \leq \kappa \leq \infty$ and $\mathbb{P}$ be a probability measure on $\mathcal{X} \times \mathcal{Y}$. We say that $\mathbb{P}$ has Tsybakov noise exponent κ if there exists a constant*

$C > 0$ such that for all small $t > 0$ we have

$$\mathbb{P}_X(\{x \in \mathcal{X} : |2\eta(x) - 1| \leq t\}) \leq Ct^\kappa. \tag{3.1}$$

Observe that all distributions have Tsybakov noise exponent $\kappa = 0$. It is also easy to see that a distribution with Tsybakov noise exponent κ has Tsybakov noise exponent κ' for all $\kappa' < \kappa$. In the extreme case $\kappa = \infty$, the conditional probability η is bounded away from $1/2$.

We may note that (3.1) does not make any assumption on the location of the noise. This will be covered in the assumption to follow.

To formulate a geometric noise assumption which describes the location of noise of the $\mathbb{P}$ distribution, we first need the following definitions. Let

$$\begin{aligned}
\mathcal{X}_1 &:= \left\{x \in \mathcal{X} : \eta(x) > \frac{1}{2}\right\} \\
\mathcal{X}_0 &:= \left\{x \in \mathcal{X} : \eta(x) = \frac{1}{2}\right\} \\
\mathcal{X}_{-1} &:= \left\{x \in \mathcal{X} : \eta(x) < \frac{1}{2}\right\}
\end{aligned}$$

and for $x \in \mathcal{X}$ we define a distance function $x \mapsto \tau_x$ by

$$\tau_x := \begin{cases} d(x, \mathcal{X}_0 \cup \mathcal{X}_{-1}) & \text{if } x \in \mathcal{X}_1 \\ d(x, \mathcal{X}_0 \cup \mathcal{X}_1) & \text{if } x \in \mathcal{X}_{-1} \\ 0 & \text{otherwise,} \end{cases} \tag{3.2}$$

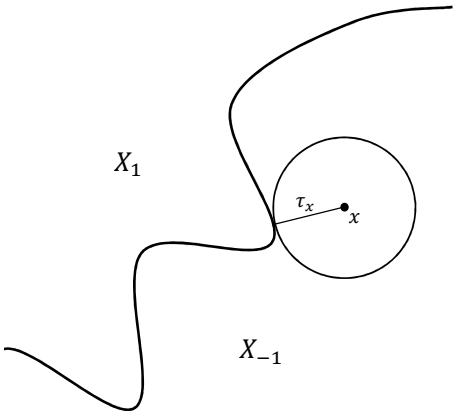


Figure 3.1: τ_x measures the distance to the decision boundary.

where $d(x, A)$ denotes the distance between x and the set A with respect to the Euclidean norm, i.e.

$$d(x, A) = \inf\{\|x - y\|_2 : y \in A\}.$$

The function τ_x measures the distance of x to the "decision boundary".
Now we may formulate the geometric noise assumption, which addresses the location of noise of the distribution $\mathbb{P}$. It allows one to estimate the approximation error for Gaussian RKHS and therein lies its main value.

Assumption 2. *(**Location of noise**) Let $\mathcal{X} \subset \mathbb{R}^d$ be compact and $\mathbb{P}$ be a probability measure on $\mathcal{X} \times \mathcal{Y}$. We say that $\mathbb{P}$ has geometric noise exponent $\alpha > 0$ if there exists a constant $C > 0$ such that*

$$\int_{\mathcal{X}} |2\eta(x) - 1| \exp\left(-\frac{\tau_x^2}{t}\right) \mathbb{P}_X(dx) \leq Ct^{\alpha d/2}, \quad t > 0. \tag{3.3}$$

We say that $\mathbb{P}$ has geometric noise exponent ∞ if it has geometric noise exponent α for all $\alpha > 0$.

The condition (3.3) describes the concentration of the measure $|2\eta - 1| d\mathbb{P}_X$ around the decision boundary. The less the measure is concentrated in this region, the larger the geometric noise exponent can be chosen. The assumption is illustrated in figure 3.2.

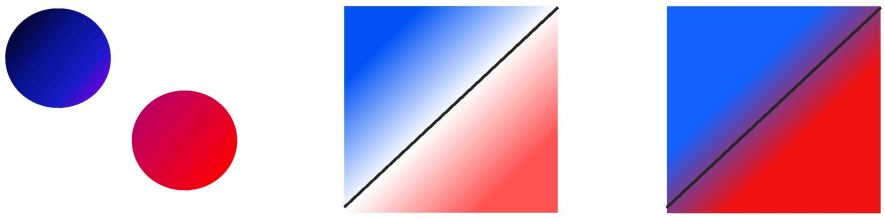


Figure 3.2: Illustration of the geometric noise assumption. Assume $X_0 = \emptyset$. From left to right, we can see three cases (1) $\alpha = \infty$. X_1 and X_{-1} are strictly separated, (2) P_X is lowly concentrated around the decision boundary, (3) $|2\eta - 1|$ is close to 0 near the decision boundary.

3.1.2 Results

Under assumptions 1 and 2 it is possible to obtain learning rates for Gaussian kernel SVMs. This is the content of the next theorem derived in [9].

Theorem 5. *Let $\mathcal{X}$ be the closed unit ball of $\mathbb{R}^d$ and $\mathbb{P}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$ with Tsybakov noise exponent $\kappa \in [0, \infty]$ and a geometric noise exponent $\alpha \in (0, \infty)$. Let $\hat{\theta}_n$ be the solution to the minimization problem*

$$\hat{\theta}_n := \arg \min_{\theta \in \mathbb{H}_{\sigma_n}(\mathcal{X})} \left(\mathbb{P}_n m_\theta + \lambda_n^2 \|\theta\|^2_{\mathbb{H}_{\sigma_n}(\mathcal{X})} \right).$$

Define

$$\beta := \begin{cases} \frac{\alpha}{2\alpha+1} & \text{if } \alpha \leq \frac{1}{2} + \frac{1}{\kappa} \\ \frac{2\alpha(\kappa+1)}{2\alpha(\kappa+2)+3\kappa+4} & \text{otherwise,} \end{cases}$$

and $\lambda_n := n^{-(\alpha+1)/(2\alpha)\beta}$ and $\sigma_n := n^{\beta/(\alpha d)}$. Then for all $\varepsilon > 0$ there exists a $C > 0$ such that for all $x \geq 1$ and $n \geq 1$ (the SVM classifier) $\hat{\theta}_n$ satisfies with probability at least $1 - e^{-x}$

$$Pl_{\hat{\theta}_n} - Pl_{\theta_0} \leq Cx^2 n^{-\beta+\varepsilon}.$$

If $\alpha = \infty$ the last assertion holds if $\sigma_n = \sigma$ is a constant with $\sigma > 2\sqrt{d}$.

First we may note that in the case when $\alpha \leq \frac{1}{2} + \frac{1}{\kappa}$, we do not get very fast rates, i.e. we get rates slower than $n^{-1/2}$. In the other case, i.e. when $\alpha > \frac{1}{2} + \frac{1}{\kappa}$, we may obtain rates that are faster than $n^{-1/2}$. This is the case if and only if $\alpha > \frac{3\kappa+4}{2\kappa}$.

It is also interesting to consider what happens in the limiting cases. In the limiting case $\kappa \to \infty$, we have a strong assumption on the amount of noise, but a weak assumption on the location of noise. In this situation we get

$$\frac{2\alpha(\kappa+1)}{2\alpha(\kappa+2)+3\kappa+4} \to \frac{2\alpha}{2\alpha+3}.$$

Thus we have rates faster than $n^{-1/2}$ if and only if $\alpha > 3/2$.

In the limit case $\alpha \to \infty$, we have a strong assumption on the location of noise and a weak assumption on the amount of noise. Then we get

$$\frac{2\alpha(\kappa+1)}{2\alpha(\kappa+2)+3\kappa+4} \to \frac{\kappa+1}{\kappa+2}.$$

3.2 A General Bound on Excess Risk

In this section, we consider an arbitrary reproducing kernel Hilbert space $\mathbb{H}$ associated with a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Bousquet, Blanchard and Massart in [2] prove a very general result that may in particular be applied to bound the excess risk for a support vector machine classifier.

3.2.1 Assumptions

We present two different settings that are treated in [2]. The difference lies in the way the capacity of the RKHS is analyzed. We shall consider the following assumptions.

(A1) $\mathbb{H}$ is a separable space. (Separability of $\mathbb{H}$ is ensured e.g. when $\mathcal{X}$ is a compact topological space, k is continuous on $\mathcal{X} \times \mathcal{X}$ and $k(x,x) \leq M^2 < \infty$ for all $x \in \mathcal{X}$.)

(A2) ("Low noise" condition) for all $x \in \mathcal{X}$: $|\eta(x) - \frac{1}{2}| \geq \eta_0$.

(A3) for all $x \in \mathcal{X}$: $\min\{\eta(x), 1-\eta(x)\} \geq \eta_1$.

We shall consider two different settings employing the assumptions above.

Setting S1. Suppose that the assumptions (A1), (A2) and (A3) are all satisfied. In this setting, the capacity of the RKHS is analyzed via the spectral properties of the kernel integral operator $L_k : L^2(\mathbb{P}_X) \to L^2(\mathbb{P}_X)$ given by

$$(L_k f)(x) = \int k(x,y) f(y) d\mathbb{P}_X(y),$$

which is positive, self-adjoint and trace-class (see [2] for full details). The operator L_k can be diagonalized in an orthogonal basis of $L^2(\mathbb{P}_X)$, it has discrete spectrum

$\lambda_1 \geq \lambda_2 \geq \ldots$ (the eigenvalues are considered with repeated multiplicities) and satisfies $\sum_{j\geq 0} \lambda_j < \infty$. For $n \in \mathbb{N}$ define

$$\gamma_n := \eta_1^{-1} \frac{1}{\sqrt{n}} \inf_{d\in\mathbb{N}} \left(\frac{d}{\sqrt{n}} + \frac{\eta_1}{M} \sqrt{\sum_{j>d} \lambda_j} \right).$$

Setting S2. Suppose that the assumptions (A1) and (A2) are satisfied. Define

$$\phi_n(\delta) := \int_0^\delta \sqrt{N(\varepsilon, B_{\mathbb{H}}, \|\cdot\|_\infty)} \mathrm{d}\varepsilon. \tag{3.4}$$

Let δ_n be the solution to the equation

$$\phi_n(\delta_n) = M^{-1}\sqrt{n}\delta_n^2. \tag{3.5}$$

For $n \in \mathbb{N}$ define

$$\gamma_n := M^{-2}\delta_n^2.$$

3.2.2 Results

Theorem 6. *Consider either setting* (S1) *under assumptions* (A1), (A2) *and* (A3)*, or setting* (S2) *under assumptions* (A1) *and* (A2)*. Define the constant* $w_1 = \eta_1$ *for setting* (S1) *and* $w_1 = 1$ *for setting* (S2)*. Let* $\nu > 0$ *be a fixed real number, and let* $\lambda_n > 0$ *be a real number satisfying*

$$\lambda_n^2 \geq c \left(\gamma_n + \frac{1}{w_1} \frac{\log(\frac{\log n}{\nu}) \vee 1}{n} \right), \tag{3.6}$$

where c *is a universal constant. Let* $\hat{\theta}_n$ *be the solution to the optimization problem*

$$\hat{\theta}_n := \arg\min_{\theta\in\mathbb{H}} \frac{1}{n} \sum_{i=1}^n (1 - \theta(X_i)Y_i)_+ + \lambda_n^2 \|\theta\|_{\mathbb{H}}^2. \tag{3.7}$$

Then

$$Pm_{\hat{\theta}_n} - Pm_{\theta_0} \leq 2 \inf_{\theta\in\mathbb{H}} \left[Pm_\theta - Pm_{\theta_0} + 8\lambda_n^2 M^2 \|\theta\|_k^2 \right] + 4\lambda_n^2 \left(16 + \frac{cw_1}{\eta_0} \right).$$

The reader might have noticed the resemblance of Theorem 6, Setting (S2), to Theorem 4 in the sense that both give a bound on the excess risk in terms of the approximation error

$$a(\lambda) := \inf_{\theta\in\mathbb{H}} (P(m_\theta - m_{\theta_0}) + \lambda_n^2 \|\theta\|_{\mathbb{H}}^2)$$

and the square of the worst solution δ_n to the equation $\phi_n \leq \sqrt{n}\delta_n^2$. The function ϕ_n in (3.4) corresponds to an entropy bound on the supremum of the empirical process in view of section 2.2 on entropy bounds.
A slight difference in the formulations of the theorems lies in the bound (3.6) that gives a restriction on the smoothing parameter λ_n in Theorem 6.

Remark 2. Let us remark that the relative classification error $Pl_{\hat{\theta}_n} - Pl_{\theta_0}$ is upper-bounded by the relative hinge loss error $Pm_{\hat{\theta}_n} - Pm_{\theta_0}$ as shown in [11], lemma .4.23, hence, the theorem results in a bound on the relative classification error as well.

3.3 Deriving Convergence Rates for Support Vector Machines With a Gaussian Kernel

Thus far we have presented two general results that give rates for empirical minimization classifiers with regularization, namely Theorem 4 and 6. Next we specialize to SVM with a Gaussian kernel RKHS, which we briefly introduced in section 1.2.2.

3.3.1 Following the Approach of Bousquet et al. [2]

In this section we shall assume that the Tsybakov noise exponent κ is infinite; as is implied by assumption (A2). We shall assume that the distribution $\mathbb{P}$ has a geometric noise exponent $\alpha > 0$, which will help us bound the approximation error. Under these assumptions, we obtain rates for SVM as stated in the following theorem.

Theorem 7. *Let $\mathcal{X}$ be the closed unit ball of $\mathbb{R}^d$, and $\mathbb{P}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$ with Tsybakov noise exponent $\kappa = \infty$ and a geometric noise exponent $\alpha \in (0, \infty)$. Let assumptions* (A1) and (A2) *be satisfied. Let $\sigma_n := n^{1/((2+\alpha)d)}$ and $\lambda_n := n^{-(1+\alpha)/(2+\alpha)}$. Let $\mathbb{H}_{\sigma_n}$ be the reproducing kernel Hilbert space on $\mathcal{X}$ with a Gaussian kernel of width $1/\sigma_n$. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a given training set and $\mathbb{P}_n$ the corresponding empirical measure. Finally, let $\hat{\theta}_n$ be the solution to the minimization problem*

$$\hat{\theta}_n := \arg \min_{\theta \in \mathbb{H}_{\sigma_n}(\mathcal{X})} \left(\mathbb{P}_n m_\theta + \lambda_n^2 \|\theta\|_{\mathbb{H}}^2 \right).$$

Then the SVM classifier $\hat{\theta}_n$ satisfies

$$P(m_{\hat{\theta}_n} - m_{\theta_0}) = O_P^*(n^{-\frac{\alpha}{\alpha+2}} (\log n)^{d+1}).$$

Before we give the proof of theorem 7, we will need one technical lemma and a bound on the approximation error. We follow the approach presented in [9], which gives a bound on the approximation error

$$a_\sigma(\lambda) := \inf_{\theta \in \mathbb{H}_\sigma} P(m_\theta - m_{\theta_0}) + \lambda^2 \|\theta\|_{\mathbb{H}_\sigma}^2$$

by a suitable combination of σ and λ. More precisely, for all $\lambda > 0$ it holds

$$a_\sigma(\lambda) \lesssim \sigma^d \lambda^2 + \sigma^{-\alpha d}. \tag{3.8}$$

Note that in order for the right hand side of (3.8) to converge to zero it is necessary to assume that $\lambda \to 0$ and $\sigma \to \infty$.

Theorem 8. *Let $\sigma > 0$, $\mathcal{X}$ be the closed unit ball of the Euclidean space $\mathbb{R}^d$ and $a_\sigma(\cdot)$ be the approximation error function with respect to $\mathbb{H}_\sigma(\mathcal{X})$. Furthermore, let $\mathbb{P}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$ that has a geometric noise exponent $0 < \alpha < \infty$ with constant C as in Assumption 2. Then there exists a constant $c_d > 0$ depending only on the dimension d such that for all $\lambda > 0$ we have*

$$a_\sigma(\lambda) \leq c_d(\sigma^d \lambda^2 + C(2d)^{\alpha d/2} \sigma^{-\alpha d}).$$

Proof. We first rewrite the approximation error by introducing a linear operator $V_\sigma : L^2(\mathbb{R}^d) \to \mathbb{H}_\sigma(\mathbb{R}^d)$ defined by

$$V_\sigma g(x) = \frac{(2\sigma)^{d/2}}{\pi^{d/4}} \int_{\mathbb{R}^d} e^{-2\sigma^2\|x-y\|_2^2} g(y)dy$$

for $g \in L^2(\mathbb{R}^d), x \in \mathbb{R}^d$ (here $L^2(\mathbb{R}^d)$ denotes the space of measurable functions whose rth powers are Lebesgue integrable). This operator is an isometric isomorphism, i.e. a surjective linear map such that $\|V_\sigma g\|_{\mathbb{H}_\sigma(\mathbb{R}^d)} = \|g\|_{L^2(\mathbb{R}^d)}$ for all $g \in L^2(\mathbb{R}^d)$. Thus we obtain

$$\begin{aligned} a_\sigma(\lambda) &= \inf_{g \in L^2(\mathbb{R}^d)} \lambda^2 \|V_\sigma g\|^2_{\mathbb{H}_\sigma(\mathbb{R}^d)} + P(m_{V_\sigma g} - m_{\theta_0}) && (3.9) \\ &= \inf_{g \in L^2(\mathbb{R}^d)} \lambda^2 \|g\|^2_{L^2(\mathbb{R}^d)} + P(m_{V_\sigma g} - m_{\theta_0}). && (3.10) \end{aligned}$$

We find a bound on the approximation error $a_\sigma(\lambda)$ by making a specific choice of $g \in L^2(\mathbb{R}^d)$. However, first we need a lemma that enlarges the support of $\mathbb{P}$ to ensure all balls of the form $B(x, \tau_x)$, where τ_x is the distance to the decision boundary as defined in (3.2), are contained in the enlarged support. This is crucial to enable to control the behaviour of $V_\sigma g$ by tails of Gaussian distributions.

Lemma 5. *Let X be a closed unit ball of $\mathbb{R}^d$ and $\mathbb{P}$ be a probability measure on $X \times Y$ with regular conditional probability $\eta(x) = \mathbb{P}(Y = 1|x), x \in X$. Let $\tilde{X} := 3X$, i.e. $\tilde{X} := \{3x : x \in X\}$. Define*

$$\tilde{\eta}(x) := \begin{cases} \eta(x) & \text{if } \|x\| \le 1, \\ \eta\left(\frac{x}{\|x\|}\right) & \text{otherwise.} \end{cases}$$

Define also $\tilde{X}_{-1} = \{x \in \tilde{X} : \tilde{\eta}(x) < \frac{1}{2}\}$ and $\tilde{X}_1 = \{x \in \tilde{X} : \tilde{\eta}(x) > \frac{1}{2}\}$. Let $B(x, r)$ denote the open ball of radius r about x in $\mathbb{R}^d$. Then for $x \in X_1 = \{x \in X : \eta(x) > \frac{1}{2}\}$ we have $B(x, \tau_x) \subset \tilde{X}_1$ and for $x \in X_{-1} = \{x \in X : \eta(x) < \frac{1}{2}\}$ we have $B(x, \tau_x) \subset \tilde{X}_{-1}$.

Proof. Let $x \in X_1$ and $\tilde{x} \in B(x, \tau_x)$. If $\tilde{x} \in X$ we have $\|x - \tilde{x}\| < \tau_x$ which implies $\tilde{\eta}(\tilde{x}) = \eta(\tilde{x}) > 1/2$ by the definition of τ_x. This shows $\tilde{x} \in \tilde{X}_1$. Now assume $\tilde{x} \notin X$, i.e. $\|\tilde{x}\| > 1$. Since $\|\langle x, \tilde{x}\rangle\| \le \|\tilde{x}\|$ and using Pythagoras' theorem we obtain

$$\left\| \frac{\tilde{x}}{\|\tilde{x}\|} - x \right\|^2 \le \left\| \tilde{x} - \frac{\langle x, \tilde{x}\rangle \tilde{x}}{\|\tilde{x}\|^2} \right\|^2 + \left\| \frac{\langle x, \tilde{x}\rangle \tilde{x}}{\|\tilde{x}\|^2} - x \right\|^2 = \|x - \tilde{x}\|^2.$$

Thus $\left\| \frac{\tilde{x}}{\|\tilde{x}\|} - x \right\| < \tau_x$, which implies $\tilde{\eta}(\tilde{x}) = \eta\left(\frac{\tilde{x}}{\|\tilde{x}\|}\right) > 1/2$. □

Proof of Theorem 8 (continued). We first define a measurable map $\tilde{\theta} : \tilde{X} \to [-1, 1]$ which satisfies $\tilde{\theta} = 1$ on $\tilde{X}_1$, $\tilde{\theta} = -1$ on $\tilde{X}_{-1}$ and $\tilde{\theta} = 0$ otherwise. Set $g := \left(\frac{\sigma^2}{\pi}\right)^{d/4} \tilde{\theta}$. Then we obtain

$$\|g\|_{L^2(\mathbb{R}^d)} \le \left(\frac{81\sigma^2}{\pi}\right)^{d/4} B^d, \tag{3.11}$$

where B^d denotes the volume of X.

By lemma (4) we may rewrite

$$P(m_\theta - m_{\theta_0}) = P_X(\eta(1-\theta)_+ + (1-\eta)(1+\theta)_+ - \eta(1-\theta_0)_+ - (1-\eta)(1+\theta_0)_+)$$

where denote $P_X f = \int f(x)\mathbb{P}_X(dx)$. For measurable $\theta \in [-1,1]$ it then follows

$$\begin{aligned} P(m_\theta - m_{\theta_0}) &= P_X(\eta(1-\theta) + (1-\eta)(1+\theta) + \\ &\quad -\eta(1-\theta_0) - (1-\eta)(1+\theta_0)) \\ &= P_X((\theta-\theta_0)(1-2\eta)) = P_X(|\theta-\theta_0||1-2\eta|). \end{aligned} \tag{3.12}$$

Since $-1 \leq \tilde{\theta} \leq 1$, then we get also $-1 \leq V_\sigma g \leq 1$. Since $\mathbb{P}_X$ has support in X, then by (3.12), we obtain

$$Pm_{V_\sigma g} - Pm_{\theta_0} = \mathbb{E}_{\mathbb{P}_X}(|2\eta - 1||V_\sigma g - \theta_0|). \tag{3.13}$$

To bound $|V_\sigma g(x) - \theta_0(x)|$ for $x \in X_1$, observe

$$\begin{aligned} V_\sigma g &= \left(\frac{2\sigma^2}{\pi}\right)^{d/2} \int_{\mathbb{R}^d} e^{-2\sigma^2\|x-y\|_2^2} \tilde{\theta}(y) dy \\ &= \left(\frac{2\sigma^2}{\pi}\right)^{d/2} \int_{\mathbb{R}^d} e^{-2\sigma^2\|x-y\|_2^2} (\tilde{\theta}(y)+1) dy - 1 \\ &\geq \left(\frac{2\sigma^2}{\pi}\right)^{d/2} \int_{B(x,\tau_x)} e^{-2\sigma^2\|x-y\|_2^2} (\tilde{\theta}(y)+1) dy - 1. \end{aligned}$$

Lemma 5 showed that for all $x \in X_1$ it holds $B(x,\tau_x) \subset \tilde{X}_1$ thus we get

$$\begin{aligned} V_\sigma g &\geq 2\left(\frac{2\sigma^2}{\pi}\right)^{d/2} \int_{B(x,\tau_x)} e^{-2\sigma^2\|x-y\|_2^2} dy - 1 \\ &= 1 - 2\mathbb{P}_{\gamma_\sigma}(|u| \geq \tau_x), \end{aligned}$$

where $\gamma_\sigma := (2\sigma^2/\pi)^{d/2} e^{-2\sigma^2|u|^2} du$ is a spherical Gaussian in $\mathbb{R}^d$. According to the tail bound in [7], inequality (3.5) on page 59, we have

$$1 \geq V_\sigma g(x) \geq 1 - 8e^{-\sigma^2\tau_x^2/2d}, \quad x \in X_1.$$

Since for $x \in X_{-1}$ we can obtain an analogous estimate, we have

$$|V_\sigma g(x) - \theta_0(x)| \leq 8e^{-\sigma^2\tau_x/2d}$$

for all $x \in X_1 \cup X_{-1}$. Using 3.13, the last observation and the geometric noise assumption for $t := 2d/\sigma^2$ yields

$$\begin{aligned} Pm_{V_\sigma g} - Pm_{\theta_0} &\leq 8\mathbb{E}_{x\sim\mathbb{P}_X}(|2\eta(x) - 1|e^{-\sigma^2\tau_x/2d}) \\ &\leq 8C(2d)^{\alpha d/2}\sigma^{-\alpha d}, \end{aligned} \tag{3.14}$$

where C is the constant in the theorem. The result follows combining (3.11), (3.14) and (3.9). □

Proof of Theorem 7. Lemma 2 gives us a bound on the entropy number of a unit ball in a Gaussian RKHS as follows

$$\phi_n(\delta) = \int_0^\delta \sqrt{\log N(\varepsilon, B_{\mathbb{H}_\sigma}, \|.\|_\infty)} \mathrm{d}\varepsilon \lesssim \int_0^\delta \sigma^{d/2} \left(\log \frac{1}{\varepsilon}\right)^{\frac{1+d}{2}} \mathrm{d}\varepsilon.$$

We would like to obtain the solution to equality (3.5). Thus we consider the following equality (note that we are only interested in the asymptotic behaviour (as $n \to \infty$) of the solution, thus we leave out constants and write $\sim$ instead of equality)

$$\int_0^{\delta_n} \sigma_n^{d/2} \left(\log \frac{1}{\varepsilon}\right)^{\frac{1+d}{2}} \mathrm{d}\varepsilon \sim \sqrt{n}\delta_n^2.$$

We now need the following lemma.

Lemma 6. $\int_0^\delta \left(\log \frac{1}{\varepsilon}\right)^{\frac{1+d}{2}} d\varepsilon \sim \delta \left(\log \frac{1}{\delta}\right)^{\frac{d+1}{2}}$ *for* $\delta \to 0$.

Proof. Since the function $\log \frac{1}{\delta}$ is decreasing, the integral is lower bounded by

$$\int_0^\delta \left(\log \frac{1}{\varepsilon}\right)^{\frac{1+d}{2}} \mathrm{d}\varepsilon \geq \delta \left(\log \frac{1}{\delta}\right)^{\frac{1+d}{2}}.$$

Integration by parts and again using that $\log \frac{1}{\delta}$ is decreasing yields

$$\begin{aligned}\int_0^\delta \left(\log \frac{1}{\varepsilon}\right)^{\frac{d+1}{2}} \mathrm{d}\varepsilon &= \delta \left(\log \frac{1}{\delta}\right)^{\frac{d+1}{2}} + \frac{d+1}{2} \int_0^\delta \left(\log \frac{1}{\varepsilon}\right)^{\frac{d-1}{2}} \mathrm{d}\varepsilon \\ &\leq \delta \left(\log \frac{1}{\delta}\right)^{\frac{d+1}{2}} + \frac{d+1}{2} \frac{1}{\log \frac{1}{\delta}} \int_0^\delta \left(\log \frac{1}{\varepsilon}\right)^{\frac{d+1}{2}} \mathrm{d}\varepsilon\end{aligned}$$

which gives an upper bound

$$\int_0^\delta \left(\log \frac{1}{\varepsilon}\right)^{1+d} \mathrm{d}\varepsilon \leq \frac{1}{1 - \frac{d+1}{2} \frac{1}{\log \frac{1}{\delta}}} \delta \left(\log \frac{1}{\delta}\right)^{\frac{d+1}{2}}.$$

The assertion follows. □

Hence (3.5) reduces to

$$\sigma_n^{d/2} \delta_n \left(\log \frac{1}{\delta_n}\right)^{\frac{d+1}{2}} \sim \sqrt{n}\delta_n^2. \tag{3.15}$$

It can be shown that the solution to equality (3.15) is of order $\frac{\sigma_n^{d/2}}{\sqrt{n}} (\log n)^{\frac{d+1}{2}}$. Consider substitution $y_n := \delta_n^{2/(d+1)}$ that will transform equation (3.15) into a more suitable form

$$\kappa_n \log \frac{1}{y_n} = y_n, \tag{3.16}$$

where we used the notation $\kappa_n := (\frac{\sigma_n^d}{n})^{1/(d+1)}$ to simplify the manipulation. If we define a function $f : x \mapsto x e^{\kappa_n^{-1} x}$, then the solution y_n to (3.16) satisfies $f(y_n) = 1$. For every $a \in \mathbb{R}$ define

$$y_n^a := \kappa_n \log n - a\kappa_n \log\log n.$$

Then for every fixed a, it is straightforward to show that

$$\lim_{n\to\infty} f(y_n^a)(\log n)^{a-1} = 1.$$

Thus for every $a > 1$ there exists $n_0 \in \mathbb{N}$ such that $y_n \geq y_n^a$ for all $n > n_0$, and for every $a < 1$ there exists $n_1 \in \mathbb{N}$ such that $y_n \leq y_n^a$ for every $n > n_1$. Hence we conclude that for $n \to \infty$

$$y_n = \kappa_n \log n - \kappa_n \log\log n + o\left(\kappa_n \log\log n\right).$$

Thus $y_n \sim \kappa_n \log n = \left(\frac{\sigma_n^d}{n}\right)^{1/(d+1)} \log n$, which implies $\delta_n \sim \frac{\sigma_n^{d/2}}{\sqrt{n}} \left(\log n\right)^{\frac{d+1}{2}}$.

Substituting the solution δ_n into (3.6) we obtain

$$\lambda_n^2 \gtrsim \frac{\sigma_n^d}{n} \left(\log n\right)^{d+1} + \frac{\log(\log n)}{n}. \tag{3.17}$$

By Theorem 6, Theorem 8 and by (3.17) it follows that (as $n \to \infty$)

$$\begin{aligned} Pm_{\hat{\theta}_n} - Pm_{\theta_0} &\lesssim a(\lambda_n^2) + \lambda_n^2 \\ &\lesssim \sigma_n^d \lambda_n^2 + \sigma_n^{-\alpha d} + \frac{\sigma_n^d}{n} \left(\log n\right)^{d+1} + \frac{\log(\log n)}{n} \\ &\lesssim \sigma_n^d \left(\frac{\sigma_n^d}{n} \left(\log n\right)^{d+1} + \frac{\log(\log n)}{n} \right) + \sigma_n^{-\alpha d} \\ &\quad + \frac{\sigma_n^d}{n} \left(\log n\right)^{d+1} + \frac{\log(\log n)}{n} \\ &\lesssim \frac{\sigma_n^{2d}}{n} \left(\log n\right)^{d+1} + \sigma_n^{-\alpha d} \end{aligned} \tag{3.18}$$

We obtained a rate in (3.18) that is valid for an arbitrary choice of σ_n (note that the parameter σ_n may be chosen arbitrarily in our estimation procedure provided that $\sigma_n \to \infty$ as $n \to \infty$). But we are mainly interested in a rate that is optimal with respect to the choice of σ_n, hence we may assume $\sigma_n = n^x$. Then the rates generated by the terms in (3.18) are $2dx - 1$ and $-\alpha dx$ respectively (neglecting the influence of the logarithm since it may be dominated by n to the power ε for any small $\varepsilon > 0$). To obtain the optimal rate, we minimize the maximum of $2dx - 1$ and $-\alpha dx$ to get the optimal solution

$$x^* = \frac{1}{d(2+\alpha)}.$$

The optimal choice of $\sigma_n = 1/(d(2+\alpha))$ and $\lambda_n = -(1+\alpha)/(2+\alpha)$ leads to a rate $P(m_{\hat{\theta}_n} - m_{\theta_0}) = O_P^*(n^{-\frac{\alpha}{2+\alpha}} \left(\log n\right)^{d+1})$. □

3.3.2 Following the Approach of Van der Vaart and Wellner [11]

We adopt the approach presented in section 2.3 and assume that the SVM uses a Gaussian kernel RKHS. In this case there is no obvious restriction on the smoothing parameter λ. In this section we also assume that the distribution $\mathbb{P}$ has a geometric noise exponent $\alpha > 0$ and a Tsybakov noise exponent $\kappa \geq 0$.

Theorem 9. *Let $\mathcal{X}$ be the closed unit ball of $\mathbb{R}^d$, and $\mathbb{P}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$ with Tsybakov noise exponent $\kappa \in [0, \infty)$ and a geometric noise exponent $\alpha \in (0, \infty)$. Let $\sigma_n = n^{\beta/(\alpha d)}$ and $\lambda_n := n^{-(1+\alpha)/(2\alpha)\beta}$. Let $\mathbb{H}_{\sigma_n}$ be the reproducing kernel Hilbert space on $\mathcal{X}$ with a Gaussian kernel of width $1/\sigma_n$. Finally, let $\hat{\theta}_n$ be the solution to the minimization problem*

$$\hat{\theta}_n := \arg \min_{\theta \in \mathbb{H}_{\sigma_n}(\mathcal{X})} \left(\mathbb{P}_n m_\theta + \lambda_n^2 \|\theta\|_{\mathbb{H}}^2 \right).$$

Then the SVM classifier $\hat{\theta}_n$ satisfies

$$P(m_{\hat{\theta}_n} - m_{\theta_0}) = O_P^*(n^{-\beta} (\log n)^{2(d+1)}),$$

where β is given by

$$\beta = \begin{cases} \frac{2\alpha}{2\alpha+3} & \text{if } \alpha \leq \frac{\kappa}{2} + \frac{1}{2} \\ \frac{\alpha(\kappa+1)}{\kappa+1+\alpha(\kappa+2)} & \text{if } \alpha > \frac{\kappa}{2} + \frac{1}{2}. \end{cases}$$

Proof. Theorem 2 that guarantees us a bound on (2.5) provided that we can uniformly bound the class of functions over which the supremum is taken and that we can bound the $L^2(\mathbb{P})$-norm $\|m_\theta - m_{\theta_0}\|_{P,2}$ in terms of $P(m_\theta - m_{\theta_0})$.

Fix $\delta > 0$ and $n \in \mathbb{N}$ (sufficiently large). Define

$$\mathcal{F}_{n,\delta} := \left\{ m_\theta - m_{\theta_0} : \|\theta\|_{\mathbb{H}_{\sigma_n}} \leq \frac{\delta}{\lambda_n}, P(m_\theta - m_{\theta_0}) < \delta^2 \right\}.$$

It is shown in [11], lemma .4.23, that $\|m_\theta - m_{\theta_0}\|_{P,2}$ may be bounded above by $P(m_\theta - m_{\theta_0})$ under the assumption that the distribution $\mathbb{P}$ has a Tsybakov noise exponent κ. Then

$$\|m_\theta - m_{\theta_0}\|_{P,2} \leq (1 + \|\theta - \theta_0\|_\infty^{1/2}) P(m_\theta - m_{\theta_0})^{\frac{1}{2}} + C^{1/(\kappa+1)} P(m_\theta - m_{\theta_0})^{\frac{\kappa}{2(\kappa+1)}}. \quad (3.19)$$

Since by the definition of $\mathcal{F}_{n,\delta}$ it holds $P(m_\theta - m_{\theta_0}) < \delta^2$, then we can bound (3.19) as follows

$$\|m_\theta - m_{\theta_0}\|_{P,2} \leq (1 + \|\theta - \theta_0\|_\infty^{1/2})\delta + C^{1/(\kappa+1)} \delta^{\kappa/(\kappa+1)}.$$

We realize that the uniform norm $\|\cdot\|_\infty$ can be bounded with the norm corresponding to a reproducing kernel Hilbert space, as shown in [2]. This essentially follows by the reproducing kernel property (1.10), but we have to guarantee a bound on the kernel function k as well (for the Gaussian kernel, this is satisfied). For all $f \in \mathbb{H}$ and for all $x \in \mathcal{X}$ it then follows by (1.10) and the Cauchy-Schwarz inequality

$$\begin{aligned} |f(x)| &= |\langle f, k(x, \cdot) \rangle_{\mathbb{H}}| \leq \|f\|_{\mathbb{H}} \|k(x, \cdot)\|_{\mathbb{H}} \\ &= \|f\|_{\mathbb{H}} \sqrt{k(x, x)} \leq M \|f\|_{\mathbb{H}}, \end{aligned}$$

where M^2 is a (constant) bound on $k(x,x)$ for every $x \in \mathcal{X}$. Thus $\|f\|_\infty \leq M\|f\|_{\mathbb{H}}$.

Next by the triangle inequality and using the bound on $\|\theta\|_{\mathbb{H}_{\sigma_n}}$ in the definition of $\mathcal{F}_{n,\delta}$, we obtain

$$\|\theta - \theta_0\|_\infty \leq \|\theta\|_\infty + \|\theta_0\|_\infty \leq M\frac{\delta}{\lambda_n} + 1$$

Hence we finally get a bound (denote it $f_n(\delta)$) on the $L^2(\mathbb{P})-$norm $\|m_\theta - m_{\theta_0}\|_{P,2}$ given by

$$f_n(\delta) := \left[1 + \left(1 + M\frac{\delta}{\lambda_n}\right)^{1/2}\right]\delta + C^{1/(\kappa+1)}\delta^{\kappa/(\kappa+1)}. \tag{3.20}$$

The Lipschitz property of the map $\theta \mapsto m_\theta$ gives us a bound on

$$\|m_\theta - m_{\theta_0}\|_\infty \leq L\|\theta - \theta_0\|_\infty \leq L\left(1 + \frac{\delta}{\lambda_n}\right). \tag{3.21}$$

By Theorem 2 it follows

$$\begin{aligned}\mathbb{E}_P^* \sup_{m_\theta - m_{\theta_0} \in \mathcal{F}_{n,\delta}} |\mathbb{G}_n(m_\theta - m_{\theta_0})| \;\lesssim\; & J_{[]}(f_n(\delta), \mathcal{F}_{n,\delta}, L^2(P)) \\ & + \frac{J_{[]}^2(f_n(\delta), \mathcal{F}_{n,\delta}, L^2(P))}{f_n(\delta)^2\sqrt{n}} L\left(1 + \frac{\delta}{\lambda_n}\right). \end{aligned} \tag{3.22}$$

Next we bound the entropy integral

$$J_{[]}(f_n(\delta), \mathcal{F}_{n,\delta}, L^2(P)) = \int_0^{f_n(\delta)} \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}_{n,\delta}, L^2(P))}\mathrm{d}\varepsilon.$$

The bracketing number $N_{[]}(\varepsilon, \mathcal{F}, L^2(P))$ is dominated by the covering number corresponding to the uniform norm in the following way

$$N_{[]}(\varepsilon, \mathcal{F}, L^2(P)) \leq 2N(\varepsilon, \mathcal{F}, \|\cdot\|_\infty).$$

This can be seen as follows: a ball of radius $\varepsilon > 0$ with respect to the uniform norm, centered at $f \in \mathcal{F}$, can be covered by two brackets $[f - \varepsilon, f]$ and $[f, f + \varepsilon]$ of size $\|f - (f - \varepsilon)\|_2 = \|(f + \varepsilon) - f\|_2 = \varepsilon$.

Next, we realize that (1) since $\theta \mapsto m_\theta$ is Lipschitz (with some constant L) we can bound the entropy number of $\mathcal{F}_{n,\delta}$ in terms of a corresponding class of functions θ (2) a covering number is an increasing function with respect to set inclusion and (3) the $\varepsilon-$covering number of a ball with radius R equals $\varepsilon/R-$covering number of a unit ball. These observations yield

$$\begin{aligned} N(L\varepsilon, \mathcal{F}_{n,\delta}, \|\cdot\|_\infty) &\leq N\left(\varepsilon, \left\{\theta \in \mathbb{H}_{\sigma_n} : \|\theta\|_{\mathbb{H}_{\sigma_n}} \leq \frac{\delta}{\lambda_n}, P(m_\theta - m_{\theta_0}) < \delta^2\right\}, \|\cdot\|_\infty\right) \\ &\leq N\left(\varepsilon, \left\{\theta \in \mathbb{H}_{\sigma_n} : \|\theta\|_{\mathbb{H}_{\sigma_n}} \leq \frac{\delta}{\lambda_n}\right\}, \|\cdot\|_\infty\right) \\ &= N\left(\frac{\varepsilon\lambda_n}{\delta}, B_{\mathbb{H}_{\sigma_n}}, \|\cdot\|_\infty\right). \end{aligned}$$

A bound on entropy number of a unit ball in a Gaussian RKHS derived in [12], lemma 4.5, yields

$$\log N\left(\frac{\varepsilon\lambda_n}{\delta}, B_{\mathbb{H}_{\sigma_n}}, \|\cdot\|_\infty\right) \lesssim \sigma_n^d \left(\log\frac{\delta}{\varepsilon\lambda_n}\right)^{d+1}.$$

To obtain a bound on the entropy integral $J_{[]}(f_n(\delta), \mathcal{F}, L^2(P))$ we realize by lemma (6) that

$$\int_0^{f_n(\delta)} \sigma_n^{d/2}\left(\log\frac{\delta}{\varepsilon\lambda_n}\right)^{(d+1)/2} d\varepsilon \sim \sigma_n^{d/2} f_n(\delta)\left(\log\frac{\delta}{\lambda_n f_n(\delta)}\right)^{(d+1)/2}.$$

It follows by (3.22) and the last observation

$$\begin{aligned}\mathbb{E}_P^* \sup_{m_\theta - m_{\theta_0}\in\mathcal{F}_{n,\delta}} |\mathbb{G}_n(m_\theta - m_{\theta_0})| \lesssim\ & \sigma_n^{d/2} f_n(\delta)\left(\log\frac{\delta}{\lambda_n f_n(\delta)}\right)^{(d+1)/2} \\ & + \sigma_n^d \left(\log\frac{\delta}{\lambda_n f_n(\delta)}\right)^{(d+1)} \left(1+\frac{\delta}{\lambda_n}\right)\frac{1}{\sqrt{n}}.\end{aligned}$$

Note that the right hand side of the last expression depends on the asymptotic relationship between δ_n and λ_n, thus in order to simplify expressions, let us first consider what can we deduce about the ratio δ_n/λ_n. By the bound on the approximation error it follows that δ_n^2 cannot be smaller that the approximation error for the best possible choice of σ. This implies

$$\delta_n^2 \gtrsim \inf_\sigma \left(\sigma^d\lambda^2 + \sigma^{-\alpha d}\right) = \lambda^{\frac{2\alpha}{1+\alpha}},$$

since infimum is attained for $\sigma \sim \lambda^{-\frac{2}{(1+\alpha)d}}$. Thus

$$\delta_n \gtrsim \lambda_n^{\frac{\alpha}{1+\alpha}}. \tag{3.23}$$

Since for $\lambda \downarrow 0$ it holds $\lambda^{\frac{\alpha}{1+\alpha}} \gtrsim \lambda$, it follows that $\frac{\delta_n}{\lambda_n} \gtrsim 1$. Hence

$$f_n(\delta) = \left(1 + \left(1 + M\frac{\delta_n}{\lambda_n}\right)^{\frac{1}{2}}\right)\delta_n + C^{\frac{1}{\kappa+1}}\delta_n^{\frac{\kappa}{\kappa+1}} \lesssim \frac{\delta_n^{\frac{3}{2}}}{\sqrt{\lambda_n}} + \delta_n^{\frac{\kappa}{\kappa+1}}.$$

and

$$\begin{aligned}\mathbb{E}_P^* \sup_{m_\theta - m_{\theta_0}\in\mathcal{F}_{n,\delta}} |\mathbb{G}_n(m_\theta - m_{\theta_0})| \lesssim\ & \sigma_n^{d/2} f_n(\delta)\left(\log\frac{\delta}{\lambda_n f_n(\delta)}\right)^{(d+1)/2} \\ & + \sigma_n^d \left(\log\frac{\delta}{\lambda_n f_n(\delta)}\right)^{(d+1)} \frac{\delta}{\lambda_n\sqrt{n}} =: \phi_n(\delta).\end{aligned}$$

We want to find a solution δ_n such that $\phi(\delta_n) \lesssim \sqrt{n}\delta_n^2$, where by ϕ_n we denoted the right hand side of the last inequality.

Without any further assumptions, we cannot deduce which of the terms in

$$f_n(\delta) = \delta^{\frac{3}{2}}/\sqrt{\lambda} + \delta^{\frac{\kappa}{\kappa+1}}$$

dominates for $\delta \to 0$. Let us therefore first consider the two cases separately.

Assume first that $\delta^{\frac{3}{2}}/\sqrt{\lambda} \gtrsim \delta^{\frac{\kappa}{\kappa+1}}$, which is the case if and only if $\delta^{\frac{\kappa+3}{\kappa+1}} \gtrsim \lambda$. Then $f_n(\delta)$ may be dominated by $\delta^{\frac{3}{2}}/\sqrt{\lambda}$. To find δ_n that satisfies $\phi_n(\delta_n) \lesssim \sqrt{n}\delta_n^2$, we are looking for δ_n that satisfies (note here that the entropy integral is an increasing function of $f_n(\delta)$ hence we may plug in a bound on $f_n(\delta)$)

$$\sigma_n^{d/2} \frac{\delta_n^{\frac{3}{2}}}{\sqrt{\lambda_n}} \left(\log \frac{1}{\sqrt{\delta_n \lambda_n}} \right)^{(d+1)/2} \lesssim \sqrt{n}\delta_n^2 \tag{3.24}$$

$$\sigma_n^{d} \left(\log \frac{1}{\sqrt{\delta_n \lambda_n}} \right)^{d+1} \frac{\delta}{\lambda_n \sqrt{n}} \lesssim \sqrt{n}\delta_n^2. \tag{3.25}$$

To solve (3.24) and (3.25) we may use substitutions $\tau := \sqrt{\delta\lambda}, \omega := \delta\lambda$, respectively. This yields

$$\delta_n \sim \frac{\sigma_n^d}{n\lambda_n} (\log n)^{d+1}.$$

By theorem 4 a bound on the rate to the Bayes risk is given by

$$P(m_{\hat{\theta}_n} - m_{\theta_0}) = O_P^* \left(\frac{\sigma_n^{2d}}{n^2\lambda_n^2} (\log n)^{2(d+1)} + \sigma_n^d \lambda_n^2 + \sigma_n^{-\alpha d} \right).$$

To find the optimal rate with respect to the choice of σ_n, λ_n, we substitute powers of n for $\sigma_n := n^x$ and $\lambda_n = n^y$, which reduces the problem to a $2-$dimensional optimization problem. Thus we would like to minimize the function of two variables

$$f : (x, y) \mapsto \max\{2dx - 2y - 2, dx + 2y, -\alpha dx\}. \tag{3.26}$$

It is relatively easy to see that f is lower bounded. Simple calculations show that the three hyperplanes in the argument of maximum (3.26) have a single point of intersection for any admissible values of parameters d, α given by

$$(x^*, y^*) = \left(\frac{2}{d(2\alpha+3)}, -\frac{1+\alpha}{2\alpha+3} \right).$$

By lower boundedness of f and given the geometric interpretation (of mutual position of hyperplanes in $\mathbb{R}^3$), it follows that the point of intersection is a global minimum of f. The value of f at the point of minimum is $f(x^*, y^*) = -\frac{2\alpha}{2\alpha+3}$.

Our initial assumption $\delta^{\frac{3}{2}}/\sqrt{\lambda} \gtrsim \delta^{\frac{\kappa}{\kappa+1}}$ gives us (after plugging in the solution δ_n and λ_n) the following relationship between α and κ

$$\alpha < \frac{\kappa}{2} + \frac{1}{2}.$$

Thus in the above case, the rate of convergence to the Bayes risk is $n^{-\frac{2\alpha}{2\alpha+3}}$.

Next we consider the case $\delta^{\frac{3}{2}}/\sqrt{\lambda} \lesssim \delta^{\frac{\kappa}{\kappa+1}}$, which is the case if and only if $\delta^{\frac{\kappa+3}{\kappa+1}} \lesssim \lambda$. Then $f_n(\delta)$ may be dominated by $\delta^{\frac{\kappa}{\kappa+1}}$. It follows

$$\phi_n(\delta) \lesssim \sigma_n^{d/2} \delta^{\frac{\kappa}{\kappa+1}} \left(\log \frac{\delta}{\lambda_n \delta^{\frac{\kappa}{\kappa+1}}} \right)^{(d+1)/2} + \sigma_n^d \frac{\delta}{\lambda_n \sqrt{n}} \left(\log \frac{\delta}{\lambda_n \delta^{\frac{\kappa}{\kappa+1}}} \right)^{(d+1)}.$$

We want to find δ_n such that

$$\sigma_n^{\frac{d}{2}} \delta_n^{\frac{\kappa}{\kappa+1}} \left[\log \frac{\delta_n}{\lambda_n \delta_n^{\frac{\kappa}{\kappa+1}}} \right]^{\frac{d+1}{2}} \lesssim \sqrt{n}\delta_n^2 \tag{3.27}$$

$$\sigma_n^d \frac{\delta_n}{\lambda_n \sqrt{n}} \left[\log \frac{\delta_n}{\lambda_n \delta_n^{\frac{\kappa}{\kappa+1}}} \right]^{d+1} \lesssim \sqrt{n}\delta_n^2. \tag{3.28}$$

Solving (3.27)-(3.28) yields

$$\delta_n \gtrsim \frac{\sigma_n^{\frac{(\kappa+1)}{2(\kappa+2)}d}}{n^{\frac{(\kappa+1)}{2(\kappa+2)}}} (\log n)^{\frac{(\kappa+1)}{2(\kappa+2)}(d+1)} \tag{3.29}$$

$$\delta_n \gtrsim \frac{\sigma_n^d}{n\lambda_n} (\log n)^{d+1}, \tag{3.30}$$

where (3.29) corresponds to (3.27) and (3.30) corresponds to (3.28). Thus the rate is given by

$$P(m_{\hat{\theta}_n} - m_{\theta_0}) = O_P^* \left(\frac{\sigma_n^{\frac{\kappa+1}{\kappa+2}d}}{n^{\frac{\kappa+1}{\kappa+2}}} (\log n)^{\frac{\kappa+1}{\kappa+2}(d+1)} + \frac{\sigma_n^{2d}}{n^2\lambda_n^2} (\log n)^{2(d+1)} + \sigma_n^d \lambda_n^2 + \sigma_n^{-\alpha d} \right).$$

Optimizing over $\lambda_n = n^y, \sigma_n = n^x$ reduces to minimizing the maximum of $\frac{\kappa+1}{\kappa+2}(dx - 1), 2dx - 2y - 2, dx + 2y, -\alpha dx$. Observe that we cannot attain a better bound from the approximation error than when $\sigma^d\lambda^2 \sim \sigma^{-\alpha d}$, i.e. $dx^* + 2y = -\alpha dx^*$. Thus we conclude that if we can find x^*, y^* such that the following is satisfied

$$\frac{\kappa+1}{\kappa+2}(dx^* - 1) \geq 2dx^* - 2y^* - 2 = dx^* + 2y = -\alpha dx^*, \tag{3.31}$$

then the rate $-\alpha dx^*$ is optimal. Condition (3.31) corresponds to a solution

$$(x^*, y^*) = \left(\frac{2}{d(2\alpha+3)}, -\frac{1+\alpha}{2\alpha+3} \right),$$

and the inequality yields $\alpha \leq \frac{\kappa}{2} + \frac{1}{2}$. The optimal rate is then $f^* = -\frac{2\alpha}{2\alpha+3}$. However, our initial assumption $\delta^{\frac{\kappa+3}{\kappa+1}} \lesssim \lambda$ yields $\alpha \geq \frac{\kappa}{2} + \frac{1}{2}$.

Similarly, if we can find x^*, y^* such that

$$2dx^* - 2y^* - 2 \geq \frac{\kappa+1}{\kappa+2}(dx^* - 1) = dx^* + 2y = -\alpha dx^*, \tag{3.32}$$

then the rate $-\alpha dx^*$ is optimal. Condition (3.32) corresponds to a solution

$$(x^*, y^*) = \left(\frac{\kappa+1}{d(\kappa+1+\alpha(\kappa+2))}, -\frac{(1+\alpha)(\kappa+1)}{2(\kappa+1+\alpha(\kappa+2))} \right)$$

and the inequality yields $\alpha \geq \frac{\kappa}{2} + \frac{1}{2}$. Our initial assumption $\delta^{\frac{\kappa+3}{\kappa+1}} \lesssim \lambda$ yields $\alpha \geq \frac{\kappa}{2} + \frac{1}{2}$. The optimal rate is then $f^* = -\frac{\alpha(\kappa+1)}{\kappa+1+\alpha(\kappa+2)}$. This finishes the proof. □

3.3.3 Steinwart and Scovel's Entropy Bound

Steinwart and Scovel in [9] use a different bound on the entropy number of a unit ball in a Gaussian kernel RKHS. We believed that the reason why we obtained faster convergence rates might be caused by the type of bound they use, however, it turns out that their bound is, in a sense, a limiting case of the bound in lemma (2) and we obtain (in limit) the same rates as in the previous section. The bound they suggest is as follows.

Lemma 7. *Let $\sigma \geq 1, 0 < p < 2$ and $X \subset \mathbb{R}^d$ be a compact subset with nonempty interior. Then there is a constant $c_{p,d}$ independent of σ such that for all $\varepsilon > 0$ we have*

$$\log N(\varepsilon, B_{\mathbb{H}_\sigma(X)}, \|\cdot\|_\infty) \leq c_{p,d}\sigma^{(1-\frac{p}{4})d}\varepsilon^{-p}.$$

We obtain the following result.

Theorem 10. *Let $\mathcal{X}$ be the closed unit ball of $\mathbb{R}^d$, and $\mathbb{P}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$ with Tsybakov noise exponent $\kappa \in [0, \infty)$ and a geometric noise exponent $\alpha \in (0, \infty)$. Fix a constant $p \in (0, 2)$. Let $\sigma_n := n^{\beta/(\alpha d)}$ and $\lambda_n := n^{-(1+\alpha)/(2\alpha)\beta}$. Let $\mathbb{H}_{\sigma_n}$ be the reproducing kernel Hilbert space on $\mathcal{X}$ with a Gaussian kernel of width $1/\sigma_n$. Finally, let $\hat{\theta}_n$ be the solution to the minimization problem*

$$\hat{\theta}_n := \arg\min_{\theta \in \mathbb{H}_{\sigma_n}(\mathcal{X})} \left(\mathbb{P}_n m_\theta + \lambda_n^2 \|\theta\|_{\mathbb{H}}^2\right).$$

Then the SVM classifier $\hat{\theta}_n$ satisfies

$$P(m_{\hat{\theta}_n} - m_{\theta_0}) = O_P^*(n^{-\beta}(\log n)^{2(d+1)}),$$

where β is given by

$$\beta = \begin{cases} \frac{2\alpha}{(2+p)\alpha+3} & \alpha \leq \frac{\kappa}{2} + \frac{1}{2} \\ \frac{\alpha(\kappa+1)}{\kappa+1+\alpha(\kappa+2)+\frac{p}{2}\left[\kappa(\alpha+\frac{1}{2})+\frac{1}{2}\right]} & \alpha > \frac{\kappa}{2} + \frac{1}{2}. \end{cases}$$

Proof. Using similar techniques as in Theorem 9 we obtain for the class of functions

$$\mathcal{F}_{n,\delta} := \left\{ m_\theta - m_{\theta_0} : \|\theta\|_{\mathbb{H}_{\sigma_n}} \leq \frac{\delta}{\lambda_n}, P(m_\theta - m_{\theta_0}) < \delta^2 \right\}$$

by Theorem 2

$$\mathbb{E}_P^* \sup_{m_\theta - m_{\theta_0} \in \mathcal{F}_{n,\delta}} |\mathbb{G}_n(m_\theta - m_{\theta_0})| \lesssim J_{[]}(f_n(\delta), \mathcal{F}_{n,\delta}, L^2(P)) \left(1 + \frac{J_{[]}(f_n(\delta), \mathcal{F}_{n,\delta}, L^2(P))}{f_n(\delta)^2\sqrt{n}} A\right)$$

with $A := L(1 + \frac{\delta}{\lambda_n})$ as in (3.22) and $f_n(\delta) = \delta^{\frac{3}{2}}/\sqrt{\lambda} + \delta^{\frac{\kappa}{\kappa+1}}$. Next we bound the entropy integral

$$J_{[]}(f_n(\delta), \mathcal{F}_{n,\delta}, L^2(P)) = \int_0^{f_n(\delta)} \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}_{n,\delta}, L^2(P))} \mathrm{d}\varepsilon.$$

The entropy bound in lemma (7) yields

$$N\left(\frac{\varepsilon\lambda}{\delta}, B_{\mathbb{H}_\sigma}, \|\cdot\|_\infty\right) \lesssim \sigma^{(1-\frac{p}{4})d}\left(\frac{\varepsilon\lambda}{\delta}\right)^{-p}.$$

By the same arguments as in Theorem 9 it follows

$$\begin{aligned} J_{[]}(f_n(\delta), \mathcal{F}_{n,\delta}, L^2(P)) &\lesssim \left(\frac{\delta}{\lambda}\right)^{\frac{p}{2}} \sigma^{\frac{1}{2}(1-\frac{p}{4})d} \int_0^{f(\delta)} \varepsilon^{-\frac{p}{2}} d\varepsilon \\ &\lesssim \left(\frac{\delta}{\lambda}\right)^{\frac{p}{2}} \sigma^{\frac{1}{2}(1-\frac{p}{4})d} f(\delta)^{1-\frac{p}{2}}. \end{aligned}$$

We obtain a bound

$$\begin{aligned} \mathbb{E}_P^* \sup_{m_\theta - m_{\theta_0} \in \mathcal{F}_{n,\delta}} |\mathbb{G}_n(m_\theta - m_{\theta_0})| &\lesssim J_{[]}(f_n(\delta), \mathcal{F}_{n,\delta}, L^2(P)) + \frac{J_{[]}(f_n(\delta), \mathcal{F}_{n,\delta}, L^2(P))}{f^2(\delta)\sqrt{n}} \frac{\delta}{\lambda} \\ &\lesssim \left(\frac{\delta}{\lambda}\right)^{\frac{p}{2}} \sigma^{\frac{1}{2}(1-\frac{p}{4})d} f(\delta)^{1-\frac{p}{2}} + \left(\frac{\delta}{\lambda}\right)^{p+1} \sigma^{(1-\frac{p}{4})d} f(\delta)^{-p} \frac{1}{\sqrt{n}} \\ &=: \phi_n(\delta). \end{aligned}$$

Again we consider the two cases based on the dominating term in $f(\delta)$.

First consider the case when $\frac{\delta^{\frac{3}{2}}}{\sqrt{\lambda}}$ dominates. This is the case if and only if $\delta^{\frac{\kappa+3}{\kappa+1}} \gtrsim \lambda$. Then the solution to $\phi_n(\delta_n) \lesssim \sqrt{n}\delta_n^2$ satisfies

$$\begin{aligned} \left(\frac{\delta_n}{\lambda_n}\right)^{\frac{p}{2}} \sigma_n^{\frac{1}{2}(1-\frac{p}{4})d} \left(\frac{\delta_n^{\frac{3}{2}}}{\sqrt{\lambda_n}}\right)^{1-\frac{p}{2}} &\lesssim \sqrt{n}\delta_n^2 \\ \left(\frac{\delta_n}{\lambda_n}\right)^{p+1} \sigma_n^{(1-\frac{p}{4})d} \left(\frac{\delta_n^{\frac{3}{2}}}{\sqrt{\lambda_n}}\right)^{-p} \frac{1}{\sqrt{n}} &\lesssim \sqrt{n}\delta_n^2. \end{aligned}$$

Algebraic manipulations show that the two equations above are identical and that the solution δ_n is given by

$$\delta_n^2 \gtrsim \sigma_n^{2\frac{1-\frac{p}{4}}{1+\frac{p}{2}}d} \lambda_n^{-2} n^{-\frac{2}{1+\frac{p}{2}}}.$$

We use the same approximation error bound as in the previous section and apply theorem 4. Thus to find the optimal rate with respect to σ and λ we want to minimize the function of two variables

$$f : (x, y) \mapsto \max\left\{\left(2\left(1-\frac{p}{4}\right)dx - 2y - 2\right)\frac{1}{1+\frac{p}{2}}, dx + 2y, -\alpha dx\right\}.$$

The minimum is attained for

$$(x^*, y^*) = \left(\frac{2}{d((2+p)\alpha+3)}, -\frac{1+\alpha}{(2+p)\alpha+3}\right)$$

and the value at the point of minimum is $f^* = -\frac{2\alpha}{(2+p)\alpha+3}$.

Our initial assumption $\delta^{\frac{\kappa+3}{\kappa+1}} \gtrsim \lambda$ then yields

$$\alpha < \frac{\kappa}{2} + \frac{1}{2}.$$

Next consider the situation when $\delta^{\frac{\kappa}{\kappa+1}}$ dominates $\frac{\delta^{\frac{3}{2}}}{\sqrt{\lambda}}$, i.e. $\delta^{\frac{\kappa+3}{\kappa+1}} \lesssim \lambda$. Then the solution to $\phi_n(\delta_n) \lesssim \sqrt{n}\delta_n^2$ satisfies

$$\begin{aligned}\left(\frac{\delta_n}{\lambda_n}\right)^{\frac{p}{2}} \sigma_n^{\frac{1}{2}(1-\frac{p}{4})d} \delta_n^{\frac{\kappa}{\kappa+1}\left(1-\frac{p}{2}\right)} &\lesssim \sqrt{n}\delta_n^2 \\ \left(\frac{\delta_n}{\lambda_n}\right)^{p+1} \sigma_n^{\left(1-\frac{p}{4}\right)d} \delta_n^{-\frac{\kappa}{\kappa+1}p} \frac{1}{\sqrt{n}} &\lesssim \sqrt{n}\delta_n^2.\end{aligned}$$

This is, equivalently stated,

$$\begin{aligned}\left[\sigma_n^{\left(1-\frac{p}{4}\right)d} \lambda_n^{-p} n^{-1}\right]^{\frac{\kappa+1}{\kappa+2-\frac{p}{2}}} &\lesssim \delta_n^2 \\ \left[\sigma_n^{2\left(1-\frac{p}{4}\right)d} \lambda_n^{-2(p+1)} n^{-2}\right]^{\frac{\kappa+1}{\kappa+1-p}} &\lesssim \delta_n^2.\end{aligned}$$

The rate to the Bayes risk is given by the approximation error and the solution to the last two equations. To find the optimal rate with respect to σ and λ we want to minimize

$$\begin{aligned}f : (x, y) \;\mapsto\; &\max\{\left[\left(1-\frac{p}{4}\right)dx - py - 1\right]\frac{\kappa+1}{\kappa+2-\frac{p}{2}}, \\ &\left[2\left(1-\frac{p}{4}\right)dx - 2(p+1)y - 2\right]\frac{\kappa+1}{\kappa+1-p}, dx+2y, -\alpha dx\}.\end{aligned}$$

We follow the same procedure as in previous parts. If we can find x^*, y^* such that the following is satisfied

$$\begin{aligned}dx^* + 2y^* = -\alpha dx^* &= \left[2\left(1-\frac{p}{4}\right)dx^* - 2(p+1)y^* - 2\right]\frac{\kappa+1}{\kappa+1-p} \\ &\leq \left[\left(1-\frac{p}{4}\right)dx^* - py^* - 1\right]\frac{\kappa+1}{\kappa+2-\frac{p}{2}}, \qquad (3.33)\end{aligned}$$

then the rate $-\alpha dx^*$ is optimal, since we cannot attain a better bound from the approximation error than when $\sigma^d\lambda^2 \sim \sigma^{-\alpha d}$, i.e. $dx^* + 2y = -\alpha dx^*$. Condition (3.33) corresponds to the solution

$$(x^*, y^*) = \frac{2(\kappa+1)}{(\kappa+1)(2\alpha+3) + \alpha(\kappa+2) + \frac{p}{2}\left[\kappa(\alpha+\frac{1}{2})+\frac{1}{2}\right]}\left(\frac{1}{d}, -\frac{1}{2}(1+\alpha)\right)$$

and the inequality yields $\alpha \leq \frac{\kappa}{2} + \frac{1}{2}$. The optimal rate is then $f^* = -\frac{2\alpha}{2\alpha+3+\frac{p}{2}\left[\kappa(\alpha+\frac{1}{2})+\frac{1}{2}\right]}$. However, our initial assumption $\delta^{\frac{\kappa+3}{\kappa+1}} \lesssim \lambda$ yields $\alpha \geq \frac{\kappa}{2} + \frac{1}{2}$.

If we can find x^*, y^* such that

$$\begin{aligned}dx^* + 2y^* = -\alpha dx^* &= \left[\left(1-\frac{p}{4}\right)dx^* - py^* - 1\right]\frac{\kappa+1}{\kappa+2-\frac{p}{2}} \\ &\leq \left[2\left(1-\frac{p}{4}\right)dx^* - 2(p+1)y^* - 2\right]\frac{\kappa+1}{\kappa+1-p}, \qquad (3.34)\end{aligned}$$

then the rate $-\alpha dx^*$ is optimal Condition (3.34) corresponds to a solution

$$(x^*, y^*) = \frac{\kappa + 1}{\kappa + 1 + \alpha(\kappa + 2) + \frac{p}{2}\left[\kappa(\alpha + \frac{1}{2}) + \frac{1}{2}\right]} \left(\frac{1}{d}, -\frac{1}{2}(1 + \alpha)\right)$$

and the inequality and initial assumption yield $\alpha \geq \frac{\kappa}{2} + \frac{1}{2}$.

The optimal rate is then $f^* = -\frac{\alpha(\kappa+1)}{\kappa+1+\alpha(\kappa+2)+\frac{p}{4}[\kappa(2\alpha+1)+1]}$. This finishes the proof. □

In both cases, the optimal value of $p \in (0, 2)$ is $p \to 0$ and taking the limit we obtain for all $\varepsilon > 0$ it holds $P(m_{\hat{\theta}_n} - m_{\theta_0}) = O_P^*(n^{-\beta+\varepsilon})$ where

$$\beta = \begin{cases} \frac{2\alpha}{2\alpha+3} & \alpha \leq \frac{\kappa}{2} + \frac{1}{2} \\ \frac{\alpha(\kappa+1)}{\kappa+1+\alpha(\kappa+2)} & \alpha > \frac{\kappa}{2} + \frac{1}{2}. \end{cases}$$

We obtained (in limit) the same rates as in the previous section, where we used the entropy bound in lemma (2).

3.3.4 Steinwart and Scovel's Entropy Bound II

Steinwart and Scovel in [9] also present a seemingly "better" bound for covering numbers of a unit ball in Gaussian RKHS than we used above. It is given in the following lemma. By $\mathbb{P}_{n,X}$ we denote here the marginal distribution of the empirical measure $\mathbb{P}_n$ with respect to X.

Lemma 8. *Let $\sigma \geq 1, 0 < p \leq 2$, $\delta > 0$ and $X \subset \mathbb{R}^d$ be a compact subset with nonempty interior. Then there is a constant $c_{p,d,\delta}$ independent of σ such that for all $\varepsilon > 0$ we have*

$$\sup_{\mathbb{P}_{n,\mathcal{X}}} \log N(\varepsilon, B_{\mathbb{H}_\sigma(X)}, L^2(\mathbb{P}_{n,\mathcal{X}})) \leq c_{p,d,\delta}\sigma^{(1-\frac{p}{2})(1+\delta)d}\varepsilon^{-p}.$$

Theorem 11. *Let $\mathcal{X}$ be the closed unit ball of $\mathbb{R}^d$, and $\mathbb{P}$ be a distribution on $\mathcal{X} \times \mathcal{Y}$ with Tsybakov noise exponent $\kappa \in [0, \infty)$ and a geometric noise exponent $\alpha \in (0, \infty)$. Fix a constant $p \in (0, 2)$. Let $\sigma_n := n^{\beta/(\alpha d)}$ and $\lambda_n := n^{-(1+\alpha)/(2\alpha)\beta}$. Let $\mathbb{H}_{\sigma_n}$ be the reproducing kernel Hilbert space on $\mathcal{X}$ with a Gaussian kernel of width $1/\sigma_n$. Finally, let $\hat{\theta}_n$ be the solution to the minimization problem*

$$\hat{\theta}_n := \arg\min_{\theta \in \mathbb{H}_{\sigma_n}(\mathcal{X})} \left(\mathbb{P}_n m_\theta + \lambda_n^2 \|\theta\|_{\mathbb{H}}^2\right).$$

Then the SVM classifier $\hat{\theta}_n$ satisfies

$$P(m_{\hat{\theta}_n} - m_{\theta_0}) = O_P^*(n^{-\beta}(\log n)^{2(d+1)}),$$

where β is given by

$$\beta = \begin{cases} \frac{2\alpha}{2\alpha+3+p\left[\alpha+\frac{1}{2}\right]} & \alpha \leq \frac{\kappa}{2} + \frac{1}{2} \\ \frac{\kappa+1}{\kappa+1+\alpha(\kappa+2)+\frac{p}{2}[\kappa(\alpha+1)+1]} & \alpha > \frac{\kappa}{2} + \frac{1}{2}. \end{cases}$$

Proof. Fix $n \in \mathbb{N}, \delta > 0$. Let

$$\mathcal{F}_{n,\delta} := \{m_\theta - m_{\theta_0} : \theta \in \mathbb{H}, \|\theta\|_{\mathbb{H}} \leq \frac{\delta}{\lambda_n}, P(m_\theta - m_{\theta_0}) < \delta^2\}.$$

By 3.21 it follows that for all $f \in \mathcal{F}_{n,\delta}$ it holds $|f(x)| \leq \left(1 + \frac{\delta}{\lambda_n}\right) L$ for all $x \in \mathcal{X}$. If we define

$$\mathcal{F}'_{n,\delta} := \left\{ \frac{f}{\left(1 + \frac{\delta}{\lambda_n}\right) L} : f \in \mathcal{F}_{n,\delta} \right\}.$$

then it follows $|f'(x)| \leq 1$ for all $f' \in \mathcal{F}'_{n,\delta}$ and $x \in \mathcal{X}$. Provided that for all $f' \in \mathcal{F}'_{n,\delta}$ it holds $Pf'^2 < D^2$ for some $D \in (0, 1)$ then Theorem 1 implies (for $F \equiv 1$)

$$\mathbb{E}_P^* \sup_{f' \in \mathcal{F}'_{n,\delta}} |\mathbb{G}_n f'| \lesssim J(D, \mathcal{F}'_{n,\delta}, L^2) \left(1 + \frac{J(D, \mathcal{F}'_{n,\delta}, L^2)}{D^2\sqrt{n}}\right).$$

It follows that

$$\begin{aligned} \mathbb{E}_P^* \sup_{f \in \mathcal{F}_{n,\delta}} |\mathbb{G}_n f| &= \left(1 + \frac{\delta}{\lambda_n}\right) L \mathbb{E}_P^* \sup_{f' \in \mathcal{F}'_{n,\delta}} |\mathbb{G}_n f'| \\ &\lesssim \left(1 + \frac{\delta}{\lambda_n}\right) L J(D, \mathcal{F}'_{n,\delta}, L^2) \left(1 + \frac{J(D, \mathcal{F}'_{n,\delta}, L^2)}{D^2\sqrt{n}}\right). \quad (3.35) \end{aligned}$$

By 3.20 for every $f \in \mathcal{F}_{n,\delta}$ it holds

$$\left(Pf^2\right)^{1/2} < \left[1 + \left(1 + M\frac{\delta}{\lambda_n}\right)^{1/2}\right]\delta + C^{1/(\kappa+1)}\delta^{\kappa/(\kappa+1)},$$

which implies

$$\begin{aligned}(Pf'^2)^{1/2} &= \frac{1}{\left(1+\frac{\delta}{\lambda_n}\right)L}(Pf^2)^{1/2} \\ &< \frac{1}{\left(1+\frac{\delta}{\lambda_n}\right)L}\left(\left[1 + \left(1 + M\frac{\delta}{\lambda_n}\right)^{1/2}\right]\delta + C^{1/(\kappa+1)}\delta^{\kappa/(\kappa+1)}\right).\end{aligned}$$

Again we shall use the bound on approximation error as in theorem 8 thus $\delta \gtrsim \lambda$ which implies

$$\begin{aligned}D &:= \frac{1}{\left(1+\frac{\delta}{\lambda_n}\right)L}\left(\left[1 + \left(1 + M\frac{\delta}{\lambda_n}\right)^{1/2}\right]\delta + C^{1/(\kappa+1)}\delta^{\kappa/(\kappa+1)}\right) \\ &\lesssim \frac{1}{\frac{\delta}{\lambda_n}}\left(\frac{\delta^{3/2}}{\sqrt{\lambda_n}} + \delta^{\kappa/(\kappa+1)}\right) \\ &= \delta^{1/2}\sqrt{\lambda_n} + \delta^{-\frac{1}{\kappa+1}}\lambda_n \to 0 \quad \text{as } n \to \infty.\end{aligned}$$

Thus eventually (for all n large enough) $D \in (0,1)$ as required.

Next we realize that since the ε-covering number of the set of functions $\mathcal{F}/c$ equals $c\varepsilon$ covering number of the set of functions $\mathcal{F}$ and by similar arguments as have already been used previously it follows for an arbitrary measure Q on $\mathcal{X} \times \mathcal{Y}$

$$\begin{aligned}N(L\varepsilon, \mathcal{F}'_{n,\delta}, L^2(Q)) &\lesssim N(L\varepsilon, \mathcal{F}_{n,\delta}, L^2(Q)) \\ &\lesssim N(\varepsilon, \{\theta \in \mathbb{H} : \|\theta\|_{\mathbb{H}} \le \frac{\delta}{\lambda_n}\}, L^2(Q_X)) \\ &= N\left(\frac{\varepsilon\lambda_n}{\delta}, B_{\mathbb{H}_\sigma}, L^2(Q_X)\right).\end{aligned}$$

By increasingness of the map $x \mapsto \sqrt{1+x}$ $(x > 0)$, lemma 8 (note here that $\sup_{\mathbb{P}_n} N(\varepsilon, \mathcal{F}, L^2(\mathbb{P}_n))$ tends to $\sup_Q N(\varepsilon, \mathcal{F}, L^2(Q))$ as $n \to \infty$, where Q is a discrete measure), the last observation and the fact that $\delta \gtrsim \lambda$ it follows

$$\begin{aligned}J(D, \mathcal{F}'_{n,\delta}, L^2) &= \int_0^D \sup_Q \sqrt{1 + N(\varepsilon, \mathcal{F}'_{n,\delta}, L^2(Q))}d\varepsilon \\ &\lesssim \int_0^D \sqrt{1 + \sigma^{(1-\frac{p}{2})(1+\delta)d}\left(\frac{\varepsilon\lambda_n}{\delta}\right)^{-p}}d\varepsilon \\ &\lesssim \int_0^D \sigma^{(1-\frac{p}{2})(1+\delta)\frac{d}{2}}\left(\frac{\varepsilon\lambda_n}{\delta}\right)^{-\frac{p}{2}}d\varepsilon \\ &\lesssim \sigma^{(1-\frac{p}{2})(1+\delta)\frac{d}{2}}\left(\frac{\lambda_n}{\delta}\right)^{-\frac{p}{2}}D^{1-\frac{p}{2}}.\end{aligned}$$

We distinguish two cases. For $D \lesssim \delta^{1/2}\sqrt{\lambda_n}$, which is the case if and only if $\delta^{\frac{\kappa+3}{\kappa+1}} \gtrsim \lambda$ we obtain $J(D, \mathcal{F}'_{n,\delta}, L^2) \lesssim \sigma_n^{\left(1-\frac{p}{2}\right)\frac{d}{2}} \delta_n^{\frac{p}{4}+\frac{1}{2}} \lambda_n^{-\frac{3}{4}p+\frac{1}{2}}$. Plugging this into (3.35) and solving the corresponding equation $\phi_n(\delta_n) \lesssim \sqrt{n}\delta_n^2$ (ϕ_n is the bound as usual) yields

$$\delta_n^2 \gtrsim \left(\sigma_n^{\left(1-\frac{p}{2}\right)d} \lambda_n^{-1-\frac{3}{2}p} n^{-1}\right)^{\frac{2}{1-\frac{p}{2}}}. \tag{3.36}$$

To obtain the optimal rate to the Bayes risk (given by theorem 4), we want to minimize the maximum of

$$\left(1-\frac{p}{2}\right)dx - \left(1+\frac{3}{2}p\right)y - 1, dx + 2y, -\alpha dx.$$

The optimal (x^*, y^*) is given by

$$(x^*, y^*) = \left(\frac{2}{d(2\alpha+3+p\left[\alpha+\frac{1}{2}\right])}, -\frac{1+\alpha}{2\alpha+3+p\left[\alpha+\frac{1}{2}\right]}\right).$$

Initial condition $\delta^{\frac{\kappa+3}{\kappa+1}} \gtrsim \lambda$ yields $\alpha \leq \frac{\kappa}{2}+\frac{1}{2}$.

If $D \lesssim \delta^{\frac{\kappa}{\kappa+1}}$ then $J(D, \mathcal{F}'_{n,\delta}, L^2) \lesssim \sigma_n^{\left(1-\frac{p}{2}\right)\frac{d}{2}} \delta_n^{\frac{\kappa}{\kappa+1}+\frac{1}{\kappa+1}\frac{p}{2}} \lambda_n^{-\frac{p}{2}}$. Solving $\phi_n(\delta_n) \lesssim \sqrt{n}\delta_n^2$ yields

$$\begin{aligned}
\delta_n^2 &\gtrsim \left[\sigma_n^{\left(1-\frac{p}{2}\right)d} \lambda_n^{-2p} n^{-1}\right]^{\frac{\kappa+1}{(1+\frac{p}{2})(\kappa+2)}} \\
\delta_n^2 &\gtrsim \left[\sigma_n^{2\left(1-\frac{p}{2}\right)d} \lambda_n^{-4p-2} n^{-2}\right]^{\frac{\kappa+1}{\kappa+1-p(\kappa+2)}}.
\end{aligned}$$

To obtain the optimal rate to the Bayes risk, we want to minimize the maximum of

$$\begin{aligned}
&\left[\left(1-\frac{p}{2}\right)dx - 2py - 1\right]\frac{\kappa+1}{(\kappa+2)(1+\frac{p}{2})}, \\
&\left[2\left(1-\frac{p}{2}\right)dx - 2(2p+1)y - 2\right]\frac{\kappa+1}{(\kappa+1+p(\kappa+2)}, dx+2y, -\alpha dx.
\end{aligned}$$

Calculations similar as in previous sections yield that for $\alpha \geq \frac{\kappa}{2}+\frac{1}{2}$ the optimal (x^*, y^*) is given by

$$(x^*, y^*) = \frac{\kappa+1}{\kappa+1+\alpha(\kappa+2)+\frac{p}{2}\left[\kappa(\alpha+1)+1\right]}\left(\frac{1}{d}, -\frac{1}{2}(1+\alpha)\right)$$

and the optimal rate is $-\frac{\alpha(\kappa+1)}{\kappa+1+\alpha(\kappa+2)+\frac{p}{2}[\kappa(\alpha+1)+1]}$. □

We again obtained (in the limit $p \to 0$) the same rates as in the previous section.

3.4 Comparison of Results

We compare and summarize the results for the support vector machine with a Gaussian kernel under the assumption that the underlying distribution $\mathbb{P}$ (of the i.i.d. training set) has a Tsybakov noise exponent κ and a geometric noise exponent α. Define the SVM classifier $\hat{\theta}_n$ by

$$\hat{\theta}_n := \arg \min_{\theta \in \mathbb{H}_{\sigma_n}(\mathcal{X})} \left(\mathbb{P}_n m_\theta + \lambda_n^2 \|\theta\|^2_{\mathbb{H}_{\sigma_n(\mathcal{X})}} \right).$$

Let us first list the rates obtained in [9]. They obtain $Pl_{\hat{\theta}_n} - Pl_{\theta_0} = O_P^*(n^{-\beta+\varepsilon})$ for any $\varepsilon > 0$ where

$$\beta = \begin{cases} \frac{\alpha}{2\alpha+1} & \text{if } \alpha \leq \frac{1}{2} + \frac{1}{\kappa} \\ \frac{2\alpha(\kappa+1)}{2\alpha(\kappa+2)+3\kappa+4} & \text{otherwise.} \end{cases}$$

In our results, we specify rates of convergence of $Pm_{\hat{\theta}_n} - Pm_{\theta_0}$, however, in view of remark 2, our bounds result in a bound on the excess risk $Pl_{\hat{\theta}_n} - Pl_{\theta_0}$ as well. The results we obtained following the general approach presented in [2] (and $\kappa = \infty$), i.e.

$$Pl_{\hat{\theta}_n} - Pl_{\theta_0} = O_P^*(n^{-\frac{\alpha}{\alpha+2}} (\log n)^{d+1})$$

are not better than those presented in [9]. We see the cause of this in the additional condition that was imposed on the smoothing parameter λ.
Following the second approach in [11] proved to be more fruitful; we obtained better bounds for convergence rates than those given in [9] for an arbitrary value of κ and for $\alpha > 1/2$. These are given by $Pl_{\hat{\theta}_n} - Pl_{\theta_0} = O_P^*(n^{-\beta}(\log n)^{2(d+1)})$ where

$$\beta = \begin{cases} \frac{2\alpha}{2\alpha+3} & \text{if } \alpha \leq \frac{\kappa}{2} + \frac{1}{2} \\ \frac{\alpha(\kappa+1)}{\kappa+1+\alpha(\kappa+2)} & \text{if } \alpha > \frac{\kappa}{2} + \frac{1}{2}. \end{cases}$$

One may also note that our rates tend to (the limits of) the rates obtained in [9] in the limiting case $\alpha \to \infty$, i.e. $(\kappa+1)/(\kappa+2)$. In the limiting case $\kappa \to \infty$ we obtain the rate $2\alpha/(2\alpha+3)$.

Another interesting observation is that we obtained (in limit) the same convergence rates also using the entropy bound as suggested in [9] (actually they are slightly worse since we can only approach the optimal rate arbitrarily close but never reach it). The rates are given by $n^{-\beta}$ where

$$\beta = \begin{cases} \frac{2\alpha}{(2+p)\alpha+3} & \xrightarrow{p\to 0} \frac{2\alpha}{2\alpha+3} & \alpha \leq \frac{\kappa}{2} + \frac{1}{2} \\ \frac{\alpha(\kappa+1)}{\kappa+1+\alpha(\kappa+2)+\frac{p}{2}\left[\kappa(\alpha+\frac{1}{2})+\frac{1}{2}\right]} & \xrightarrow{p\to 0} \frac{\alpha(\kappa+1)}{\kappa+1+\alpha(\kappa+2)} & \alpha > \frac{\kappa}{2} + \frac{1}{2}. \end{cases}$$

Summarizing overall results for support vector machines with a Gaussian kernel, we obtain that for all $\varepsilon > 0$ it holds $P(l_{\hat{\theta}_n} - l_{\theta_0}) = O_P^*(n^{-\beta}(\log n)^{2(d+1)})$, where β is given by

$$\beta = \begin{cases} \frac{\alpha}{2\alpha+1} - \varepsilon & \text{if } \alpha \leq \frac{1}{2} \\ \frac{2\alpha}{2\alpha+3} & \text{if } \frac{1}{2} < \alpha \leq \frac{\kappa}{2} + \frac{1}{2} \\ \frac{\alpha(\kappa+1)}{\kappa+1+\alpha(\kappa+2)} & \text{if } \alpha > \frac{\kappa}{2} + \frac{1}{2}. \end{cases}$$

Conclusion

In the first part of this work we discussed and related some recent results on convergence rates for support vector machines. We presented a general bound on excess risk derived in [2] that may be applied in particular to SVM and a general result for penalized empirical contrast procedures derived in [11], noting the link between the two. Next we specialized to support vector machines with a Gaussian kernel and presented a bound on the rates obtained in [9]. Consequently, for the Gaussian kernel support vector machine we derived convergence rates following both approaches [2] and [11] under Tsybakov noise assumption and a geometric noise assumption on the underlying distribution. The latter approach resulted in an improvement of the bound obtained in [9].

Let us briefly remark that to guarantee the optimal rates our techniques produce, one has to specify the values of the Tsybakov noise exponent and the geometric noise exponent, which are typically not available. To obtain these values from the data, techniques such as cross validation might be used to assess for which values of α and κ the algorithm performs best.

Bibliography

[1] J.Y. Audibert and A.B. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.* 35 608633, 2007.

[2] G. Blanchard, O. Bousquet, P. Massart. Statistical Performance of Support Vector Machines. *Ann. Statist.* Vol. 36, No. 2, 489-531, 2008.

[3] B. E. Boser, I. M. Guyon and V. N. Vapnik. A training algorithm for optimal margin classifiers. In Haussler, David (editor), 5th Annual ACM Workshop on COLT, pages 144152, Pittsburgh, PA, 1992. ACM Press.

[4] O. Bousquet, S. Boucheron and G. Lugosi. Introduction to Statistical Learning Theory. *Springer.* 169-207, 2004.

[5] C. Cortes and V. Vapnik. Support-Vector Networks. *Machine Learning*, 20, 1995.

[6] L. Devroye, L. Gyorfi and G. Lugosi. A Probabilistic Theory of Pattern Recognition, *Springer*, New York. MR1383093. 1996.

[7] M. Ledoux and M. Talagrand. Probability in Banach Spaces. Isoperimetry and Processes. *Springer, Berlin.* 1991.

[8] I. Steinwart. Support Vector Machines Are Universally Consistent. *Journal of Complexity* 18, 768-791, 2002.

[9] I. Steinwart, C. Scovel. Fast Rates for Support Vector Machines Using Gaussian Kernels. *Ann. Statist.* Vol 35, No. 2, 575-607, 2007.

[10] A. Tsybakov. Optimal Aggregation of Classifiers in Statistical Learning. *Ann. Statist.* Vol. 32, No. 1, 135166. 2004.

[11] A.W. van der Vaart and J.A. Wellner. Weak Convergence and Empirical Processes. *Springer.*

[12] A.W. van der Vaart and J.H. van Zanten. Adaptive Bayesian Estimation Using a Gaussian Random Field With Inverse Gamma Bandwidth. *Ann. Statist.* Vol. 37, No. 5B, 2655-2675.

[13] A.W. van der Vaart. Asymptotic Statistics. *Cambridge University Press* 1998.

[14] A.W. van der Vaart. Support Vector Machines, lecture notes. Vrije Universiteit Amsterdam. November 2010.

[15] V. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24, 774780, 1963.

[16] T. Zhang. Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization. *Ann. Statist.* 32 56-85, 2004.

[17] www.wikipedia.org

List of Figures