DISS. ETH NO. 17690

Fast rates of convergence for adaptive classification

A dissertation submitted to
ETH ZURICH

for the degree of
Doctor of Sciences

presented by
BERNADETTA TARIGAN
MSc., Institut Teknologi Bandung
born September 26, 1972
citizen of Indonesia

accepted on the recomendation of
Prof. Dr. Sara A. van de Geer, examiner
Prof. Dr. Peter Bühlmann, co-examiner
Dr. Gilles Blanchard, co-examiner

2008

For the sake of a better learning

Contents

Abstract

Classification refers to the problem of predicting the category of an observation, based on the categories of previously observed examples and with as small an error as possible. A predictor is a function that assigns a data value to one out of a fixed number of mutually exclusive categories, called classes. Determining a suitable predictor is a statistical learning problem, since the properties of the data source are not known explicitly but have to be inferred from examples.

Let $x \in \mathcal{X}$ denote an observation and $y \in \mathcal{Y} = \{1, \ldots, m\}$ denote a category. The input-output pair (x, y) is a realization of a random pair (X, Y) governed by a joint probability distribution P on the space $\mathcal{X} \times \mathcal{Y}$. The assumption in the statistical learning is that the examples $\{(x_i, y_i)\}_{i=1}^n$ are drawn independently from the same distribution as (x, y). A predictor $g : \mathcal{X} \to \mathcal{Y}$ makes an error when $g(x) \neq y$. Under the 0–1 loss, the prediction error of g is $\widetilde{R}(g) := P(g(X) \neq Y)$. This prediction error is also called the standard risk or the true risk. Based on a given set of examples $D_n = (X_i, Y_i)_{i=1}^n$, the empirical risk minimization (ERM) method looks for the predictor $\hat{g}_n(X) = \hat{g}_n(X, D_n)$ that minimizes the empirical prediction error $(1/n) \sum_{i=1}^n \mathbb{1}(g(X_i) \neq Y_i)$ over a class $\mathcal{G}$ of candidate predictors. There exists a best theoretical predictor, the so-called Bayes predictor g^*, that has the smallest achievable risk, $\widetilde{R}(g^*)$. The performance of (a sequence of) predictors is now measured by the rates of convergence to zero of the excess prediction errors $\widetilde{R}(\hat{g}_n) - \widetilde{R}(g^*)$.

Recent results of statistical learning apply empirical process theory and concentration inequalities to show that convergence rates depend on the complexity condition of the class of candidate predictors $\mathcal{G}$ and the so-called *margin condition* or *noise condition*. To handle the complexity, we can apply either a regularized ERM-based method by defining the appropriate penalty for the class under consideration, or a complexity constraint on the class in terms of its entropy, parameterized by a constant $\rho \in (0, 1)$ (smaller ρ means simpler class).

The margin condition quantifies the identifiability of the Bayes predictor, and is parameterized by a constant $\kappa \geq 1$. When the underlying distributions behave well (low noise level), it places a small probability around the Bayes decision boundaries (κ small). The convergence rates obtained are functions of the unknown parameters ρ and κ.

Under the conditions above and for the binary case ($m = 2$), the rates of convergence to zero of $\widetilde{R}(\hat{g}_n) - \widetilde{R}(g^*)$ are adaptive to the unknown parameters and faster than the classical non-parametric rate $n^{-1/2}$. However, since the true 0–1 loss is a non-convex function, the ERM method is computationally infeasible. We replace the 0–1 loss with a convex upper bound surrogate loss $l(g(X), Y)$ so that the method can be implemented. To obtain fast convergence rates, two questions of concern are: (a) whether the Bayes predictor g^* also minimizes the surrogate l-risk $R(g) := \mathbb{E}[l(g(X), Y)]$ over all possible predictors (Bayes consistency of l), and (b) whether the minimization of the excess l-risk $R(\hat{g}_n) - R(g^*)$ of the predictors obtained implies to the minimization of the excess prediction error $\widetilde{R}(\hat{g}_n) - \widetilde{R}(g^*)$.

This thesis obtains fast convergence rates of the excess risk with respect to the hinge loss as the surrogate that adapt to the unknown margin and complexity parameters. These results help to statistically explain the practical efficiency of the support vector machine (SVM) algorithms that use the hinge loss. Furthermore, this thesis analyses the so-called classification with reject option, where we allow a predictor to reject taking a decision on the categories if the observation is too hard to classify. Under suitable margin condition and complexity constraint, we obtain fast convergence rates of the excess true risk.

In this work we do not assume that the model class contains the Bayes predictor. The results we obtain are in terms of the approximation error $\inf_{g \in \mathcal{G}} R(g) - R(g^*)$.

For the binary problem with hinge loss, we consider classifiers that are linear combinations of base functions. Instead of an ℓ_2-penalty, which is used by the SVM, we put an ℓ_1-penalty on the coefficients. Under certain conditions on the base functions, hinge loss with this complexity penalty is shown to lead to an oracle inequality involving both model complexity and margin.

While statistical properties of binary classifiers are quite well understood, their extensions to multicategory cases ($m > 2$) are not trivial and certainly more involved. However, the so-called multi-hinge loss—an extension of binary hinge loss that considers all of the categories at once—has been shown to be Bayes consistent. Furthermore, the convergence to zero (in probability) of the

excess multi-hinge risk implies the convergence to zero with the same rate (in probability) of the excess prediction error. In this thesis we show a moment bound for the so-called multi-hinge loss minimizers based on two kinds of complexity constraints: entropy with bracketing and empirical entropy. Obtaining such a result based on the latter is harder than finding one based on the former. We obtain fast rates of convergence that adapt to the unknown margin and complexity parameters, that is, $n^{-\kappa/(2\kappa-1+\rho)}$.

The reject option can improve performance in applications for which the cost of rejecting certain samples, and handling them with different procedures (for example, manual classification), is not larger than the cost of misclassifying. We can think of embedding the reject option as adding the rejection category into the output space. We consider the case in which acceptable misclassification is given as a parameter α (α-*reject loss*). Based on this reject loss, we investigate the margin condition and impose a complexity constraint on the class of predictors that lead the fast rates of convergence for the excess true risk.

Zusammenfassung

Der Begriff Klassifikation beschreibt das Problem, die Kategorie einer Beobachtung vorherzusagen, basierend auf den Kategorien gegebener Beobachtungsbeispiele. Dabei soll der Erwartungswert des mit einem Fehler verknüpften Verlusts so klein wie möglich sein. Ein Prädiktor ist eine Funktion, welche einen Datenpunkt einer von mehreren möglichen, paarweise disjunkten Kategorien zuordnet. Diese Kategorien werden als Klassen bezeichnet.

Sei $x \in \mathcal{X}$ eine Beobachtung und $y \in \mathcal{Y} = \{1, \ldots, m\}$ ein Kategorie-Bezeichner. Das Paar (x, y) ist eine Realisierung des Paares (X, Y) von Zufallsvariablen mit gemeinsamer Verteilung P auf dem Raum $\mathcal{X} \times \mathcal{Y}$. In der statistischen Lerntheorie werden die Beispiele $\{(x_i, y_i)\}_{i=1}^n$ als unabhängige Züge von der gemeinsamen Verteilung P vorausgesetzt. Ein Prädiktor $g : \mathcal{X} \to \mathcal{Y}$ erzeugt einen Fehler falls $g(x) \neq y$. Unter 0–1-Verlust ist der Vorhersagefehler von g gegeben durch $\widetilde{R}(g) := P(g(X) \neq Y)$. Dieser Vorhersagefehler wird auch als Standardrisiko oder wahres Risiko bezeichnet. Die Methode der empirischen Risiko-Minimierung (ERM) bestimmt einen Prädiktor $\hat{g}_n(X) = \hat{g}_n(X, D_n)$ welcher, gegeben eine Menge $D_n = (X_i, Y_i)_{i=1}^n$ von Beispielen, den empirischen Vorhersagefehler $(1/n) \sum_{i=1}^n \mathbb{1}(g(X_i) \neq Y_i)$ über eine Klasse $\mathcal{G}$ von möglichen Prädiktoren minimiert. Es existiert ein bester theoretischer Pädiktor, der sogenannte Bayes-Prädiktor g^*, welcher durch kleinstmögliches Risiko $\widetilde{R}(g^*)$ gekennzeichnet ist. Um die Qualität eines Prädiktors (bzw. einer Folge von Prädiktoren) zu messen, wird die Konvergenz der Risikodifferenz $\widetilde{R}(\hat{g}_n) - \widetilde{R}(g^*)$ gegen Null betrachtet.

Neuere Resultate der statistischen Lerntheorie verwenden die Theorie empirischer Prozesse und Konzentrations-Ungleichungen für Wahrscheinlichkeitsmasse, um die Abhängigkeit der Konvergenzraten von der Komplexität der Hypothesenklasse $\mathcal{G}$ und der sogenannten Margin-Bedingung (*margin condition*) oder Rausch-Bedingung (*noise condition*) zu zeigen. Um die Komplexität der

Klasse kann entweder mit Hilfe regularisierter ERM durch Formulierung eines passenden Straftermes kontrolliert werden, oder durch eine auf der Entropie der Klasse basierende Komplexitätsbedingung. Letztere wird durch einen Konstante $\rho \in (0, 1)$ parametrisiert, wobei ein kleiner Wert von ρ einer einfachen Klassenstruktur entspricht. Die Margin-Bedingung misst die Identifizierbarkeit des Bayes-Prädiktors, und wird durch einen Parameter $\kappa \geq 1$ kontrolliert. Für kleine Werte des Parameters treten entlang der Bayes-Entscheidungsgrenze nur kleine Wahrscheinlichkeiten auf, sofern das Rausch-Niveau der zugrundeliegenden Verteilungen hinreichend niedrig ist. Die resultierenden Konvergenzraten sind Funktionen der Parameter ρ und κ.

Unter den oben genannten Bedingungen und für den Fall binärer Klassenzuordnungen ($m = 2$) sind die Konvergenzraten von $\widetilde{R}(\hat{g}_n) - \widetilde{R}(g^*)$ gegen Null abhängig von den unbekannten Parametern ρ und κ, und höher als die klassische nicht-parametrische Rate von $n^{-1/2}$. Allerdings ist die ERM-Methode nicht rechnerisch durchführbar, da die 0–1-Verlustfunktion nicht konvex ist. Wir ersetzen den 0–1-Verlust durch eine konvexe Ersatzfunktion $l(g(X), Y)$, welche eine obere Schranke an den wahren Verlust darstellt, und die Methode praktisch durchführbar macht. Um hohe Konvergenzraten zu erhalten, sind zwei Fragen zu beantworten: (a) Minimiert der Bayes-Prädiktor g^* auch das dem Ersatzverlust l entsprechende l-Risiko $R(g) := \mathbb{E}[l(g(X), Y)]$, und (b) ob ein durch Minimierung des l-Differenzrisikos $R(\hat{g}_n) - R(g^*)$ bestimmter Prädiktor tatsächlich den Prädiktionsfehler $\widetilde{R}(\hat{g}_n) - \widetilde{R}(g^*)$ minimiert.

Die vorliegende Arbeit bestimmt schnelle Konvergenzraten für das Differenzrisiko bei Ersatz des 0–1-Verlusts durch eine Hinge-Verlustfunktion (*hinge loss*). Die Konvergenzraten passen sich den unbekannten Margin- und Komplexitäts-Parametern an. Diese Resultate helfen, die praktische Effizienz von Support Vector Machine-Algorithmen mit Hinge-Verlust statistisch zu erklären. Desweiteren wird die sogenannte Klassifikation mit Zurückweisung (*reject option*) untersucht, bei welcher der Prädiktor für schwer klassifizierbare Beobachtungen eine Entscheidung verweigern kann. Unter geeigneten Margin- und Komplexitäts-Bedingungen erhalten wir schnelle Konvergenzraten für das wahre Differenzrisiko. Wir nehmen dabei nicht an, dass der Bayes-Prädiktor in der Modellklasse enthalten ist. Die vorgestellten Resultate werden in Bezug auf den Approximationsfehler $\inf_{g \in \mathcal{G}} R(g) - R(g^*)$ formuliert.

Für das Zwei-Klassen-Problem mit Hinge-Verlust betrachten wir Klassifikatoren, die als Linearkombinationen von Basisfunktionen definiert sind. Anstelle des von der Support Vector Machine verwendeten ℓ_2-Strafterms untersuchen wir eine ℓ_1-Bedingung an die Koeffizienten. Bei Verwendung des Hinge-Verlusts

können wir unter gewissen Anforderungen an die Basisfunktionen die Gültigkeit einer
Oracle-Ungleichung (*oracle inequality*) nachweisen, die von Modellkomplexität und Margin abhängt.

Während die statistischen Eigenschaften des Zwei-Klassen-Falls relativ gut verstanden sind, erweist sich deren Verallgemeinerung auf mehr Kategorien ($m > 2$) als nicht-trivial und oft schwerer handhabbar. Für den sogenannten multiplen Hinge-Verlust—eine Verallgemeinerung des Hinge-Verlusts, welche sämtliche Kategorien gleichzeitig berücksichtigt—konnte allerdings Bayes-Konsistenz nachgewiesen werden. Desweiteren folgt aus der Konvergenz des auf dem multiplen Hinge-Verlusts basierenden Differenzrisikos gegen Null die entsprechende Konvergenz des wahren Differenzrisikos, und zwar mit derselben Konvergenzrate. In der hier vorgestellten Arbeit zeigen wir Momenten-Schranken für die Minimierer des multiplen Hinge-Verlusts. Die Schranken basieren auf zwei Typen von Komplexitäts-Bedingungen, nämlich auf Entropie mit sogenanntem *bracketing*, und auf empirischer Entropie. Die Herleitung des letzteren Resultats erweist als deutlich aufwendiger. Wir erhalten schnelle Konvergenzraten in Abhängigkeit von den unbekannten Margin- und Komplexitäts-Parametern, nämlich $n^{-\kappa/(2\kappa-1+\rho)}$.

Klassifikation mit Zurückweisung kann die Leistung eines Prädiktors verbessern, sofern in einer gegebenen Anwendung die Kosten für die Klassifikation eines zurückgewiesenen Beispiels mit Hilfe eines Ausweichverfahrens (z.B. Klassifikation von Hand) kleiner sind als die Kosten einer Fehlklassifikation. Zurückweisung durch den Prädiktor lässt sich als zusätzliche Kategorie im Ausgaberaum beschreiben. Wir betrachten den Fall, in welchem akzeptablen Kosten einer Fehlklassifikation in Form eines Parameters α gegeben sind (α-*reject loss*). Wir untersuchen die Margin-Bedingung und geben eine Komplexitäts-Bedingung für die Hypothesenklasse an, die zu schneller Konvergenz des wahren Differnzrisikos führt.

Chapter 1

Introduction and overview

1.1 Motivation

Classification problems are real-world problems arising in many fields. The problem of email filtering, in which we want to filter spam emails, is a classification problem with two classes: spam or non-spam. The email filtering problem is an example of a text categorization problem. For example, say we have 4601 emails (the inputs) together with the information which emails are spam and which are not (the outputs), where the word "free" occurs 0.52% among the spam and 0.07% among the non-spam (see Hastie, Tibshirani, and Friedman, 2001, Tabel 1.1). Based on the above data, we might classify a new unseen email by a classification rule that assigns it to spam if the percentage of the word "free" in the email is more than say 0.40%, and otherwise assigns it to non-spam. A question then arises: how good is the rule in predicting the new unseen email as spam or non-spam? We want to incur the smallest possible error while classifying the new email. Note that the costs of misclassification might not be equal since the consequence of misclassifying a spam email as non-spam is less severe than that of misclassifying a non-spam email as spam.

Recognition of hand-written digits is a classification problem arising in determining postal codes. Here the inputs are images of hand-written digits of the postal codes that have been normalized to have more or less the same size and orientation, and the outputs are the numbers $0, 1, \ldots, 9$.

In computer vision applications, some examples of classification problems are face detection, pedestrian detection and facial expression classification. Biosequence analysis for gene expression and multiple tumor types are applications in bioinformatics, and breast cancer diagnosis and prognosis are classification problems in medicine. See, e.g., Christianini and Shawe-Taylor (2000) and Guyon (2008) at `http://www.clopinet.com/isabelle/Projects/SVM/applist.html` for more details of these problems. These examples are only a few among many other pattern recognition problems. Such problems are also called discriminant analysis problems or supervised learning problems, where we learn from the data to predict hitherto unseen observations.

1.2 Probabilistic setting and definitions

Classification is about predicting the unknown category of an observation, while making as small an error as possible. An observation x is a collection of measurements from an input space $\mathcal{X}$. If, for example, $\mathcal{X} = \mathbb{R}^d$, then x is represented by a d-dimensional vector. The observation belongs to a category, denoted by y, which takes values in the output set $\mathcal{Y} = \{1, \ldots, m\}$, with $m \in \mathbb{N}$ the number of categories. When $m = 2$, it is called *binary* classification. *Multicategory* classification refers to the case $m > 2$. In the hand-written digit recognition problem above, we have a multicategory case with $m = 10$, and x might be represented by a 256-dimensional real vector when the images are 16×16 grayscale maps. A *predictor* is a mapping $g : \mathcal{X} \to \mathcal{Y}$. Based on a given data set $\{(x_i, y_i)\}_{i=1}^n$, we construct a function $\hat{g}_n = \hat{g}_n(x, \{x_i, y_i\}_{i=1}^n)$ that presents our best prediction of the category y of a new observation x. With abuse of notation, the mapping $\hat{g}_n : \mathcal{X} \times \{\mathcal{X} \times \mathcal{Y}\}^n \to \mathcal{Y}$ is also called a predictor.

The pair (x, y) is called the input-output pair, and it is a realization of a random pair $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ governed by a joint probability distribution on the space $\mathcal{X} \times \mathcal{Y}$. The essential assumption in statistical learning is that all the data $\{(x_i, y_i)\}_{i=1}^n$ are drawn independently from the same distribution as (x, y).

We regard $\{(x_i, y_i)\}_{i=1}^n$ as a realization of a random sample $D_n = \{(X_i, Y_i)\}_{i=1}^n$. That is, (X_i, Y_i) are independent and identically distributed according to a joint distribution $P = P(x, y)$, where $X_i \in \mathcal{X} \subset \mathbb{R}^d$ and $Y_i \in \mathcal{Y} = \{1, \ldots, m\}$, for all $i = 1, \ldots, n$. We regard (x, y) as a realization of a random pair (X, Y) having the same unknown joint distribution P and independent of the sample D_n. Let

X be observed; our goal is to predict the category Y of X.

A predictor is a (measurable) mapping $g : \mathbb{R}^d \to \{1, \ldots, m\}$. A misclassification of g at (x, y) occurs when $g(x) \neq y$. When the cost of misclassification from one category to another category is the same, say 1, and the cost of correct classification is 0, then the 0–1 loss of g at (x, y) is defined as

$$\mathbb{1}(g(X) \neq Y) \ , \tag{1.1}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function taking value 1 if its argument is true and 0 otherwise. The performance of g is measured by its expected loss with respect to the joint distribution P: the smaller the expected loss is, the better. We call the expectation of the 0–1 loss of g,

$$P(g(X) \neq Y) = \mathbb{E}[\mathbb{1}(g(X) \neq Y)] \ , \tag{1.2}$$

the *prediction error* of g. Therefore, the problem in classification is to find the classifier which minimizes the prediction error.

Let $p_j(\cdot) := P(Y = j | X = \cdot)$ be the conditional probability of category j given X, $j = 1, \ldots, m$. Given the value of the observation X, there exists a best (theoretical) predictor, the so-called *Bayes predictor*, that predicts the most likely category Y,

$$g^* = \arg \max_{j=1,\ldots,m} p_j \ . \tag{1.3}$$

That is, g^* minimizes the prediction error (1.2) over all possible predictors. To see that the prediction error is smallest when using g^*, note that for any fixed j, the conditional probability of misclassification to class j is

$$\sum_{k=1, k\neq j}^{m} P(Y = k | X = x) = \sum_{k=1, k\neq j}^{m} p_k(x) = 1 - p_j(x) \ ,$$

as shown in Lee, Lin, and Wahba (2004). Hence for a fixed x, maximizing p_j gives the smallest prediction error. For the binary case ($m = 2$), see for example Devroye, Györfi, and Lugosi (1996, Theorem 2.1). Since the underlying distribution P is unknown (and hence the p_j's are unknown too), we do not know the Bayes predictor g^*; neither can we compute the corresponding smallest prediction error $P(g^*(X) \neq Y)$. Based on the sample $D_n = \{(X_i, Y_i)\}_{i=1}^n$, we want to construct a predictor $\hat{g}_n$ to approximate the Bayes predictor g^*.

How can we construct $\hat{g}_n$ based on D_n? A reasonable approach is to replace the theoretical prediction error $\mathbb{E}[\mathbb{1}(g(X) \neq Y)]$ by its *empirical prediction*

error

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(g(X_i)\neq Y_i)\ , \tag{1.4}$$

which is the proportion of misclassifications in the sample. We then can take $\hat{g}_n$ as the function which minimizes the empirical prediction error (1.4) over a class of candidate predictors $\mathcal{G}$:

$$\hat{g}_n(X, D_n) := \arg\min_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(g(X_i)\neq Y_i)\ . \tag{1.5}$$

This approach is called the *empirical risk minimization* (ERM) method. We will describe this approach more closely in the next section.

What conditions do we need to impose for the construction of $\hat{g}_n$? The minimum requirement is that it is *consistent*. From now on we consider $\mathcal{X} = \mathbb{R}^d$ with $d \geq 1$. We note that a mapping $\hat{g}_n : \mathbb{R}^d \times \{\mathbb{R}^d \times \{1, \ldots, m\}\}^n \to \{1, \ldots, m\}$ is a function of X and of the data D_n. Hence, the *conditional prediction error* $P(\hat{g}_n(X, D_n) \neq Y|D_n)$, as a measure of performance of $\hat{g}_n$, is a random variable since it depends on the data D_n. Let us write

$$\begin{aligned}\widetilde{R}(\hat{g}_n) &:= P(\hat{g}_n(X, D_n) \neq Y|D_n)\ ,\\ \widetilde{R}^* &:= P(g^*(X) \neq Y)\ .\end{aligned}$$

The Bayes predictor $\widetilde{R}^*$ is a deterministic quantity. Further, we call the difference

$$\widetilde{R}(\hat{g}_n) - \widetilde{R}^*$$

the *excess prediction error* of $\hat{g}_n$. We drop the word "conditional" and from now on it is understood that when dealing with $\hat{g}_n = \hat{g}_n(X, D_n)$ we have in mind this dependency on the data. A sequence of predictors $\{\hat{g}_n,\ n \geq 1\}$ is consistent if it converges to the target function g^* in the limit of infinite data. That is, it satisfies the consistency condition

$$\lim_{n\to\infty} \mathbb{P}(\widetilde{R}(\hat{g}_n) - \widetilde{R}^* \geq \epsilon) = 0\ ,\quad \text{for all } \epsilon > 0\ ,$$

or, equivalently,

$$\lim_{n\to\infty} \mathbb{E}[\widetilde{R}(\hat{g}_n)] = \widetilde{R}^*\ .$$

Here, the notation $\mathbb{P}$ and $\mathbb{E}$, respectively, is used for the probability and expectation operators with respect to the joint distribution of the infinite sequence $(X_1, Y_1), (X_2, Y_2), \ldots$. The quantity $\widetilde{R}(\hat{g}_n)$ is often called the *generalization error* of $\hat{g}_n$. It is the probability of making a mistake on an unseen sample in the future that does not depend on the training samples D_n, hence it is sometimes called the *probability of misclassification* of $\hat{g}_n$. We are interested in obtaining probability or moment type inequalities of the excess prediction error with convergence rates faster than the classical nonparametric rate $n^{-1/2}$.

1.3 Empirical risk minimization

We have defined the prediction error (1.2) as the expectation of 0–1 loss. We now consider a general loss function $l : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$. The expectation of a loss l induced by a predictor g is called the theoretical *l-risk* of g and is denoted by

$$R(g) := \mathbb{E}[l(g(X), Y)] \ , \tag{1.6}$$

where the expectation is taken with respect to the unknown joint distribution P of (X, Y). We remind the readers that the notation $\widetilde{R}(g)$ is the risk of g wrt. 0–1 loss. Let g_l^* denote the minimizer of the l-risk $R(g)$ over all possible predictors,

$$g_l^* := \arg\min_{\text{all } g} R(g) \ .$$

Further, let $R^* := R(g_l^*)$ denotes the smallest l-risk, which is a deterministic quantity. We add a lower case l in g_l^* to distinguish it from g^*, which was defined as the global minimizer of $\widetilde{R}(g)$. We define the corresponding *empirical l-risk* based on the sample D_n as

$$R_n(g) := \frac{1}{n} \sum_{i=1}^{n} l(g(X_i), Y_i) \ . \tag{1.7}$$

The ERM approach is to approximate g_l^*, which minimizes the unknown theoretical l-risk (1.6), by the function

$$\hat{g}_n := \arg\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \ l(g(X_i), Y_i),$$

which minimizes the empirical l-risk (1.7) over all g, in a *given* class $\mathcal{G}$ of candidate predictors. This is the general principle of minimizing a risk functional on the basis of empirical data (see e.g., Vapnik, 2000).

The question is then whether the l-risk of the $\hat{g}_n$ obtained gets closer to the smallest l-risk with a fast rate as n gets larger. That is, whether the *excess* l-risk $R(\hat{g}_n) - R^*$ goes to zero at a fast rate. We recall that $R(\hat{g}_n) = \mathbb{E}[l(\hat{g}_n(X, D_n), Y)|D_n]$ is a random variable and that the ERM procedure is conducted in a given class of candidate predictors $\mathcal{G}$. To see the influence of the model class $\mathcal{G}$, we decompose the excess l-risk of $\hat{g}_n$ as follows:

$$R(\hat{g}_n) - R^* = \left(R(\hat{g}_n) - \inf_{g \in \mathcal{G}} R(g) \right) + \left(\inf_{g \in \mathcal{G}} R(g) - R^* \right) . \tag{1.8}$$

For reasons we will see soon, the first and the second part of the decomposition are called *estimation error* and *approximation error*, respectively.

On one hand, if the model class $\mathcal{G}$ gets larger (or in other words, more complex), then the smallest risk in the class $\inf_{g \in \mathcal{G}} R(g)$ gets closer to the smallest risk over all possible predictors R^*. Hence, the second part of the decomposition (1.8) gets smaller. The approximation error is a deterministic value that measures how well predictors in $\mathcal{G}$ can approach the target function g_l^* (it would be zero if $g_l^* \in \mathcal{G}$).

On the other hand, when $\mathcal{G}$ is very rich then minimizing the empirical risk (1.7) would lead to an over-fitting situation. For example, with $\mathcal{X} = \mathbb{R}^d$ and X continuous, one can almost certainly construct a predictor $\hat{g}_n(X, D_n)$ which perfectly predicts the categories of the training data D_n (i.e., the empirical risk on the data is zero), but behaves on the other points as the opposite of the target function g_l^* (i.e., for 0–1 loss, $g(X) \neq Y$ for all $(X, Y) \notin D_n$, so that $P(g(X) \neq Y) = 1$). Hence, the first part of the decomposition (1.8) gets larger. The random estimation error measures how close $\hat{g}_n(X, D_n)$ is to the best predictor in $\mathcal{G}$, taking into account the use of a finite sample size n.

Clearly we need to trade off the two conflicting errors. For a given sufficiently rich model class $\mathcal{G}$, to prevent the over-fitting situation one can modify the criterion to be minimized by adding a penalty for "complicated" predictors. The penalty term depends on the complexity of the class $\mathcal{G}$. This first approach is known as *regularized* ERM or *penalized* ERM. That is,

$$\text{minimizing } \{R_n(g) \ + \ pen(g)\} \text{ over } g \in \mathcal{G}. \tag{1.9}$$

One can ask how to choose the penalty term such that the estimator obtained results in small excess l-risk. The second approach is to set some constraint on

the complexity of the model class. That is,

$$\begin{aligned} &\text{minimizing } R_n(g) \text{ over } g \in \mathcal{G} \\ &\text{subject to a complexity constraint on } \mathcal{G} \ . \end{aligned} \tag{1.10}$$

The complexity of a class of functions can be measured by the so-called ϵ-entropy number of the class. There are different definitions of entropies depending on the norm that is used, such as ϵ-entropy, ϵ-entropy with bracketing and random empirical ϵ-entropy. Roughly speaking, an ϵ-entropy number is the logarithm of the smallest number of balls with radius ϵ that are needed to cover the class. The formal definitions of these complexity measures are given in the Appendix. We do not detail the notion of complexity in this introduction—it will become clear in the sequel chapter—and we note that the complexity of a class can be summarized in a complexity parameter $\rho \in (0, 1)$.

To investigate the rate of convergence of the excess l-risk of a predictor g, we employ tools from empirical process theory. Here we give a brief outline of the main ideas behind the approach. More details are in the subsequent chapters on certain types of classification problems. First we need to introduce some notation.

We fix a loss function l and write

$$\nu_n(g) := \sqrt{n}\ (R_n(g) - R(g))\ ,\ g \in \mathcal{G}\ .$$

Let g^o be the function in $\mathcal{G}$ that achieves the minimum l-risk in the class. That is,

$$g^o := \arg\min_{g \in \mathcal{G}}\ R(g)\ .$$

Now, with the new notation, we can rewrite the decomposition (1.8) as

$$R(\hat{g}_n) - R^* = \frac{\nu_n(g^o) - \nu_n(\hat{g}_n)}{\sqrt{n}} + R_n(\hat{g}_n) - R_n(g^o) + R(g^o) - R^*\ .$$

Let $\hat{g}_n$ be an ERM estimator, i.e., $\hat{g}_n$ is the minimizer of the empirical l-risk (1.7) in the class. It means $R_n(\hat{g}_n) - R_n(g^o)$ is non-positive. Hence, bounding $\nu_n(g^o) - \nu_n(\hat{g}_n)$ by its absolute gives

$$R(\hat{g}_n) - R^*\ \leq\ \frac{|\nu_n(\hat{g}_n) - \nu_n(g^o)|}{\sqrt{n}}\ +\ R(g^o) - R^*\ . \tag{1.11}$$

We call (1.11) a *basic inequality*, following van de Geer (2000). This is the point when we enter empirical process theory. We consider the empirical process indexed by $\mathcal{G}$,

$$\nu_n = \{\nu_n(g) \ : \ g \in \mathcal{G}\} \ ,$$

and we study the asymptotic equicontinuity of the process to obtain an exponential tail probability of the empirical process. With complexity constraint or regularized ERM, fast rates of convergence can be obtained under the so-called *margin condition*. This is a condition needed to identify the Bayes predictor wrt. the loss being used, and it basically ensures that the variance of the excess loss is upper bounded in terms of the expectation of the excess loss. Originally, the terminology "margin condition" comes from the binary case of the prediction error considered in the work of Mammen and Tsybakov (1999) and Tsybakov (2004), where the behaviour of p_1, the conditional probability of category 1, is restricted near $\{x : p_1(x) = 1/2\}$. The "margin" set $\{x : p_1(x) = 1/2\}$ identifies the Bayes predictor which assigns a new x to class 1 if $p_1(x) > 1/2$ and class 2 otherwise. The margin condition is also often called the *condition on the noise level*, and it is summarized in a margin parameter κ. In this thesis the fast rates obtained are adaptive wrt. both the unknown complexity and margin parameters.

1.4 Large margin-based proxy losses

As a class $\mathcal{G}$ may be a rich set, minimization of the empirical prediction error (1.4) over $\mathcal{G}$ is computationally infeasible because 0–1 loss is a non-convex function. This is known as an NP-hard problem. To solve this problem, one can replace 0–1 loss with a convex proxy/surrogate loss.

Recently the research in the machine learning community has focused on the so-called *large margin-based losses* which basically are *convex upper bounds* of 0–1 loss. Originally these losses were designed for *binary classification* ($m = 2$) under 0–1 loss. In the binary case, the categories $\{1, 2\}$ are conveniently encoded as $\{\pm 1\}$. Instead of $g : \mathbb{R}^d \to \{\pm 1\}$, we use a mapping

$$f : \mathbb{R}^d \to \mathbb{R} \ , \tag{1.12}$$

which we call a *classifier*, so that the predictor induced by f is

$$g = \operatorname{sign}(f) \ . \tag{1.13}$$

An error then occurs when $Yf(X) \leq 0$. Equivalently, a correct classification happens if and only if $Yf(X)$ is positive (we use the convention that $\text{sign}(0) = 0$). Here the term *margin* comes in: the quantity $Yf(X)$ is called the margin. Written as a function of the margin, 0–1 loss is $\mathbb{1}(Yf(X) \leq 0)$. We remind the readers not to be confused with the term "margin condition" introduced in Section 1.3. The risk of a classifier f wrt. a loss function l is

$$R(f) := \mathbb{E}[l(Yf(X))] \ ,$$

which is equal to the risk of the predictor g induced by f. Similarly, the classification error of f wrt. 0–1 loss is

$$\widetilde{R}(f) := \mathbb{E}[\mathbb{1}(Yf(X) \leq 0)] \ ,$$

which is equal to the prediction error of the predictor g induced by f.

Using the prediction rule (1.13), the Bayes predictor (1.3) for the binary case can be written as

$$g^*(x) = \text{sign}(2p_1(x) - 1) \ , \tag{1.14}$$

or equivalently, the Bayes classifier is

$$f^*(x) = 2 \cdot \mathbb{1}\{p_1(x) > 1/2\} - 1 \ , \tag{1.15}$$

almost surely on the set $\{x : p_1(x) \neq 1/2\}$. Among popular and successful convex large margin-based losses $l(Yf(X))$, some examples are *exponential* loss $l(Yf(X)) = \exp(-Yf(X))$ used in the AdaBoost algorithm, *logistic* loss $l(Yf(X)) = \log(1 + \exp(-Yf(X)))$ used in the logit regression algorithm and *hinge* loss

$$(0, 1 - Yf(X))_+ \tag{1.16}$$

used in support vector machines (SVMs) algorithm, where $(a)_+ := \max\{0, a\}$. We refer to Lin (2004, 2002), Zhang (2004c) and Bartlett, Jordan, and McAuliffe (2006) and the references therein for more examples of binary large margin-based loss functions. We intuitively see that the idea of these convex relaxations of the 0–1 loss is to obtain functions that favor large values of positive margin (i.e., correct classification, $Yf(X)$ positive).

The use of such binary convex large margin-based losses was first motivated by the computational advantage of convexity for obtaining tractable solutions of the corresponding empirical loss minimization procedures. After their big success in practice, statistical learning theory then started to investigate the

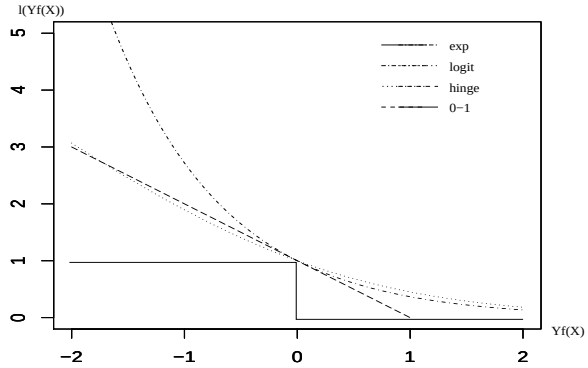

Figure 1.1: *Large margin-based losses.*

statistical behaviour of the consequences of replacing binary 0–1 loss by a proxy binary convex upper bound loss. Such procedures can be cast as regularized ERM methods. Two questions of concern are:

- Whether the proxy loss l is Bayes consistent: $f_l^* = f^*$. That is, whether the Bayes classifier f^*, which globally minimizes the 0–1 risk, is also a global minimizer f_l^* of the l-risk.
- Whether the minimization of the excess proxy l-risk $R(f) - R(f_l^*)$ implies the minimization of the excess prediction error $\widetilde{R}(f) - \widetilde{R}(f^*)$.

Recent results in statistical learning theory confirm a positive answer for both questions. Many authors have obtained these results in the binary case. Independently, Lin (2004) and Zhang (2004c) show that exponential loss, logit loss and hinge loss are almost surely Bayes consistent on the set $\{x : p_1(x) \notin \{0, 1/2, 1\}\}$. Furthermore, they also show that minimizing such excess proxy risks leads to minimization of the excess prediction error. Bartlett et al. (2006) refine the above results and in particular give the explicit relation

$$\widetilde{R}(f) - \widetilde{R}(f^*) \;\leq\; R(f) - R(f_l^*) \tag{1.17}$$

for the hinge loss (1.16). This inequality is optimal in the sense that no better bound is possible. Clearly, for any upper bound loss l of 0–1 loss, we have

$\widetilde{R}(f) \leq R(f)$ for any measurable f. The Bayes consistency of such a proxy loss gives $R(f_l^*) = R(f^*)$. For hinge loss, one has $R(f^*) = 2\widetilde{R}(f^*)$ (see, e.g., Lin, 2002). Hence, the trivial upper bound is $\widetilde{R}(f) - \widetilde{R}(f^*) \leq R(f) - (1/2)R(f_l^*)$.

Bayes consistency and an inequality such as (1.17) for the hinge loss mean that fast rates of convergence of the excess hinge risk imply fast rates of convergence of the excess prediction error. Hence, as we mentioned at the end of Section 1.2, our interest is now in probability and moment type inequalities of the excess hinge risk of $\hat{f}_n(X, D_n)$ obtained from a minimization of empirical hinge loss.

While statistical properties of binary large margin classifiers are quite well understood, their extensions to multicategory cases are not trivial. In general, the extension of ERM-based binary classification methods to solve multicategory cases is more involved. In Chapter 2 we discuss some existing extensions and the recent results on formulating the empirical multicategory proxy loss that preserve Bayes consistency and the property that excess proxy-risk minimization leads to excess prediction error minimization.

1.5 Outline of the thesis

In Section 1.3 we have seen that to obtain fast convergence rates of excess risk using the ERM method we need to: (1) identify the margin condition wrt. the loss being used; (2) use either the complexity penalty approach (1.9) or the complexity-constraint approach (1.10). When the true loss $\tilde{l}$ is computationally infeasible, we can replace it with a convex upper bound proxy loss l; hence we need to identify the margin condition wrt. the proxy loss. In summary, classification prediction rates based on the ERM method depend on the pair *(loss,complexity)*. We mainly focus on hinge loss as a proxy.

Chapter 2 gives the geometrical background of hinge loss, also known as support vector machine (SVM) loss, which was originally intended to solve the binary classification problem. The use of this loss can be cast as a penalized ERM method. We briefly discuss some variants of the loss and the penalties that lead to different names. We also discuss an algorithm that gives the entire solution path of the method based on the hinge loss and the so-called ℓ_1 penalty, and its application to real data on postal codes from the United States Postal Service (USPS) database. We then discuss some extensions of binary large margin-based losses to the multicategory case, in particular binary hinge loss.

In Chapter 3, we apply the regularized approach (1.9) where we obtain some asymptotic statistical properties of the pair (binary hinge loss, ℓ_1 penalty) with a particular choice of the class of classifiers. This chapter is the article "Classifiers of support machine type with ℓ_1 complexity regularization" (Tarigan and van de Geer, 2006).

The extension of binary hinge to the multicategory case described in Chapter 2 is used in Chapter 4 to obtain a moment bound with fast convergence rates of the excess hinge-risk. Here we impose a complexity constraint on the class of candidate classifiers, i.e. the approach (1.10). Chapter 4 is a slightly revised version of a research report (Tarigan and van de Geer, 2007) that has been accepted (subject to minor revisions) for publication in the Journal of Machine Learning Research, Dec. 2007.

In Chapter 5 we discuss the multicategory case *with reject option*, where we allow a predictor to reject taking a decision on the categories if the observation is too hard to classify. The reject option can improve performance in applications for which the cost of rejecting certain samples, and handling them with different procedures (e.g., manual classification), is not larger than the cost of wrong classification. We investigate the margin condition with respect to the reject-loss, together with some complexity constraint on the class of predictors, that lead to fast rates of convergence of the excess true risk.

Chapter 2

Support Vector Machines

2.1 Introduction

The Support Vector Machine (SVM) is a learning machine that looks for a maximum margin classifier. It is originally based on a geometrical interpretation of the linearly separable *binary* classification problem wrt. the 0–1 true loss by seeing it as a problem of finding the optimal separating hyperplane (OSH) in a dot product space (Hilbert vector space). That is, the hyperplane that not only separates/classifies all the data correctly but also maximizes the minimum distance (the margin) between the data and the hyperplane. This concept is then extended to the linearly non-separable case where the OSH is now allowed to make some mistakes on the data. A further extension is mapping the input space to a high dimensional reproducing kernel Hilbert space (RKHS) where the linear classification procedure can be done with a cheap computational cost via the dot product of the kernel function.

We divide this chapter into two parts, the binary case (Sections 2.2–2.6) and the multicategory case (Sections 2.7–2.8). First we focus on the binary case. We briefly discuss the geometrical interpretation of the SVM algorithm, the view of the algorithm as a regularization method or penalized empirical risk minimization (ERM) method, the relation of the algorithm to the Bayes rule (that is, the Bayes consistency of SVMs), and how to choose the hyperparameters in the model. We also introduce the so-called *hinge loss with ℓ_1-norm penalty* model and give an illustration and a simulation on a real data set from

the US Postal Service. Some results on asymptotic statistical properties of this model under some class of classifiers are given in Chapter 3.

Extending the binary classification procedure of SVM to multicategory classification is not trivial. In general, an extension from the binary case to the multicategory case is not obvious, regardless of the loss being used. We briefly give an overview of some existing large margin-based methods in the literature and the relationships to the Bayes predictor. We focus on an extension of binary hinge loss to multicategory case, which we call *multi-hinge* loss. Based on this loss, we obtain some asymptotic statistical results for the multicategory case that are presented in Chapter 4.

2.2 Geometric interpretation of binary SVM

We start with the linearly separable and linearly non-separable cases, and then discuss the generalization of SVM. Most material of this chapter can be found in Burges (1998), Hofmann (2003), Schölkopf and Smola (2002), and Hastie et al. (2001).

2.2.1 Linearly separable case

Let $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{\pm 1\}$ and $D_n = \{(x_i, y_i)_{i=1}^n\}$ be the training data. In Machine Learning, the training data are often called the *examples*. Positive examples are examples with positive class $y = 1$ and negative examples are examples with negative class $y = -1$.

The linearly separable case refers to the situation that the positive examples can be perfectly separated from the negative ones by a hyperplane classifier in $\mathbb{R}^d$. That is, there exists a normal weight vector $\beta \in \mathbb{R}^d$ ($\|\beta\|_2 = 1$) and a threshold $\beta_0 \in \mathbb{R}$ such that $y_i(\beta^t x_i + \beta_0) > 0$ for all $i = 1, \ldots, n$, where β^t is the transpose of $\beta \in \mathbb{R}^d$. Let $f(x) := \beta^t x + \beta_0$. The separating hyperplane is $H = H(\beta, \beta_0) = \{x : f(x) = 0\}$. The predictor induced by the classifier $f(x)$ is $g(x) = \text{sign}(f(x))$. The SVM algorithms look for a hyperplane that not only perfectly separates the positive examples from the negative examples but also maximizes the *margin* of the two classes. The margin is the distance from the hyperplane to the closest examples of either class. The hyperplane is called the optimum separating hyperplane (OSH) and is unique.

Let $\gamma = \gamma(\beta, \beta_0) := \min_{1 \le i \le n} \; y_i(\beta^t x_i + \beta_0) > 0$. The OSH is given by the choice of (β_0, β) that maximizes γ and satisfies the constraints $y_i(\beta^t x_i + \beta_0) \ge \gamma$, for all $i = 1, \ldots, n$. That is, every example is at least γ away from the decision boundary H. For any solution of the optimization problem, any positively scaled multiple is a solution as well. Thus, we can set $\|\beta\|_2 = 1/\gamma$. This is a canonical parameterization of OSH with respect to a set of training data, i.e., $\min_{1 \le i \le n} |\beta^t x_i + \beta_0| = 1$. The separating margin is simply $1/\|\beta\|_2$. Figure 2.1 shows this canonical picture, the solid line corresponds to the solution that puts the positive examples maximally apart from the negative examples. Thus, the OSH is the solution of the convex quadratic optimization problem:

$$\begin{array}{ll} \min_{\beta, \beta_0} & \frac{1}{2} \; \|\beta\|_2^2 \\ \text{subject to} & y_i(\beta^t x_i + \beta_0) \;\; \ge \;\; 1 \;\; , \;\; \text{for all } i = 1, \ldots, n \;\; . \end{array} \tag{2.1}$$

The OSH can efficiently be found by solving the associated Lagrangian dual problem, which is an easier optimization problem than (2.1).

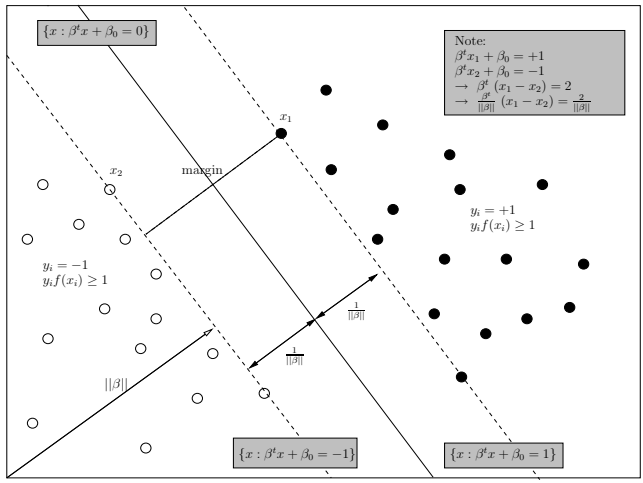

Figure 2.1: *Canonical OSH for the linearly separable case.*

Let $\alpha_i \ge 0$, $i = 1, \ldots, n$, be the Lagrange multipliers for the linear inequality constraints in (2.1). Minimizing the Lagrangian primal function $L_P(\beta, \beta_0, \alpha) = \frac{1}{2}\|\beta\|_2^2 - \sum_{i=1}^n \alpha_i \; [y_i(\beta^t x_i + \beta_0) - 1]$ wrt. the primal variables β and β_0 gives $\beta = \sum_{i=1}^n \alpha_i y_i x_i$ and $\sum_{i=1}^n \alpha_i y_i = 0$. Inserting the optimality condition for β

and β_0 back into L_P, we obtain the Lagrangian dual function

$$L_D(\alpha) = -(1/2)\sum_i \sum_j y_i y_j \alpha_i \alpha_j < x_i, x_j > \ + \ \sum_i \alpha_i$$

which is a lower bound for L_P. The Lagrangian dual problem of (2.1) is then

$$\begin{array}{ll} \max_{\beta,\beta_0} & L_D \\ \text{subject to} & \alpha_i \geq 0 \text{ for all } i \ , \text{ and } \sum_{i=1}^n \alpha_i y_i = 0 \ , \end{array} \tag{2.2}$$

which is also a convex quadratic optimization in α but has a somewhat simpler set of constraints.

An important fact, which explains the name "support vectors", is linked to the Karush-Kuhn-Tucker (KKT) conditions that require that

$$\alpha_i \ [y_i(\beta^t x_i + \beta_0) - 1] = 0 \ , \text{ for all } i = 1, \ldots, n \ .$$

For any data point (x_i, y_i) for which its functional margin satisfies $y_i(\beta^t x_i + \beta_0) > 1$, we then have $\alpha_i = 0$. Thus, $\alpha_i > 0$ holds only for those examples that are "on" the margin. It means that the optimal weight vector β can be written as an expansion of the training data, but those data points for which $\alpha_i = 0$ will not contribute. The data points for which $\alpha_i > 0$ are called the *support vectors* (SVs). Assuming that only a relatively small fraction of data points will have $\alpha_i > 0$, this will result in a *sparse* expansion of the optimal weight vector. Hence, it indicates the possibility of substantial data compression (see Schölkopf, Burges, and Vapnik, 1995).

Let $\hat{\alpha}_1, \ldots, \hat{\alpha}_n$ be the solution of the Lagrangian dual problem (2.2). The resulting parameters of the OSH are given by $\hat{f}(x) = \hat{\beta}^t x + \hat{\beta}_0$, where

$$\begin{array}{rcl} \hat{\beta} & = & \sum_{i=\text{SVs}} \hat{\alpha}_i y_i x_i \\ \hat{\beta}_0 & = & y_i - \hat{\beta}^t x_i \ , \text{ for any } i \text{ with } \alpha_i > 0 \ . \end{array}$$

The value of $\hat{\beta}_0$ above is obtained due to the KKT conditions $y_i(\beta^t x_i + \beta_0) - 1 = 0$ for all $\alpha_i > 0$. Since $y_i \in \{\pm 1\}$ for all i, multiplying the conditions by y_i gives the result. Any i such that $\alpha_i > 0$ will do, but it is common in practice to take the average of all the solutions for numerical stability. To simplify the exposition, we use only hyperplanes containing the origin, i.e., β_0 is set to zero. Hence, the prediction of a new observation x is performed by the equation $\hat{g}(x) = \text{sign}(\hat{f}(x))$, where $\hat{f}(x) = \sum_{i=SVs} \hat{\alpha}_i \ y_i \ < x_i, x >$.

2.2.2 Linearly nonseparable case

Assuming that the data is perfectly separable is not always a reasonable thing because we may have some examples on the wrong side of their margin, as shown in Figure 2.2. This is called the linearly non-separable case. Here the constraints in (2.1) are violated. To generalize the maximum margin discriminant approach, we need to soften the so-called hard-margin constraints (2.1) but only when necessary. Hence, we allow a violation of the constraints and add a penalty for every violation. This can be done by introducing the so-called slack variables $\epsilon_i \geq 0$, $i = 1, \ldots, n$, that quantify the extent of the violation, and a constant $C > 0$:

$$\begin{array}{ll} \min_{\beta,\beta_0} & \|\beta\|_2^2 + C \sum_{i=1}^n \epsilon_i \\ \text{subject to} & y_i(\beta^t x_i + \beta_0) \ \geq \ 1 - \epsilon_i \text{ for all } i \\ & \epsilon_i \geq 0 \text{ for all } i \ . \end{array} \tag{2.3}$$

In Machine Learning the formulation (2.3) above is called the L_1 *soft-margin SVM.*

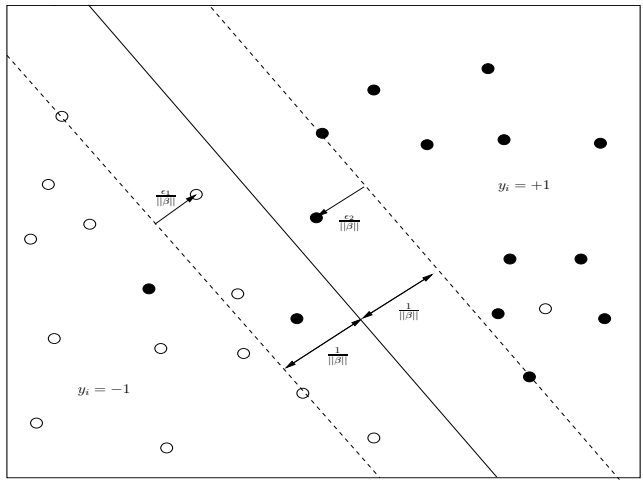

Figure 2.2: *Linear SVM in the non-separable case.*

The formulation (2.3) is based on the idea of finding the hyperplane that simultanously minimizes the number of training errors and separates the remaining correctly classified examples with maximum margin. A training error

occurs when $\epsilon_i > 1$, so $\sum_i \epsilon_i$ is an upper bound for the number of training errors. The positive constant C is a pre-chosen *tuning parameter*. It controls the trade-off between the margin and the training error. If we bound $\sum_i \epsilon_i$ at some value C', then $C = 1/C'$. As $C' \to 0$, or equivalently $C \to \infty$, one gets back to the hard-margin formulation in the separable case (2.1).

The resulting problem is still a quadratic program and can be solved along the same lines as the hard-margin maximization problem. A minor modification occurs in the constraints of the Lagrangian dual problem (2.2) where we replace the constraint $\alpha_i \geq 0$ with $0 \leq \alpha_i \leq C$ for all i. These are called box constraints, since the vector α has to be in a box of side length C. Notice that this doubles the number of inequality constraints. The support vectors are now not only the examples on the margin, but also the examples inside the margin.

There is an alternative formulation of the L_1 soft-margin SVM that uses quadratic penalties on the slack variables:

$$\begin{array}{ll} \min_{\beta,\beta_0} & \|\beta\|_2^2 + \frac{2}{C}\sum_{i=1}^n \epsilon_i^2 \\ \text{subject to} & y_i(\beta^t x_i + \beta_0) \;\geq\; 1 - \epsilon_i \text{ for all } i \; . \end{array} \tag{2.4}$$

It is called the L_2 *soft-margin SVM.*

2.2.3 Nonlinear case or general SVMs

Note that the dual objective function L_D in (2.2) depends on the data only through the dot products between examples, $< x_i, x_j > \; = x_i^t x_j$. Hence, it allows us to generalize the linear OSH concept to a nonlinear decision surface by replacing the inner products with more general kernel functions (see Boser, Guyon, and Vapnik, 1992).

The idea is to map the original input space $\mathcal{X}$ to a higher dimensional Euclidean space $\mathcal{H}_K$ via a map $\Phi : \mathcal{X} \to \mathcal{H}_K$ such that in $\mathcal{H}_K$ we can construct the linear SVM algorithm. The corresponding Lagrangian dual problem for the general soft-margin SVM is

$$\begin{array}{ll} \max_{\beta,\beta_0} & -(1/2)\sum_i \sum_j y_i y_j \alpha_i \alpha_j < \Phi(x_i), \Phi(x_j) > \; + \; \sum_i \alpha_i \\ \text{subject to} & 0 \leq \alpha_i \leq C \text{ for all } i \\ & \sum_{i=1}^n \alpha_i y_i = 0 \; . \end{array} \tag{2.5}$$

The space $\mathcal{H}_K$ is determined by a kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that measures the similarity between any two objects in $\mathcal{X}$ and their corresponding maps in $\mathcal{H}_K$

through the relation

$$K(x,\tilde{x}) =< \Phi(x), \Phi(\tilde{x}) > \ , \quad x,\tilde{x} \in \mathcal{X} \ . \tag{2.6}$$

Such a kernel K is called a reproducing kernel and the space $\mathcal{H}_K$ is called a reproducing kernel Hilbert space (RKHS). Kernels that give the relation (2.6) are those kernels that fulfill Mercer's condition (see e.g., Vapnik, 1998, Sect. 10.5.2). In solving the Lagrangian dual problem (2.5), the relation (2.6) allows us not to explicitly compute $< \Phi(x), \Phi(\tilde{x}) >$ that are elements in a possibly infinite dimensional space. We compute $K(x,\tilde{x})$ that are elements in a finite dimensional space instead. This is known as the *kernel trick*. It will work for *any* algorithm in which the data only appear as dot products. Hence, mapping back $\mathcal{H}_K$ to $\mathcal{X}$ results in a nonlinear discriminant function

$$\hat{f}(x) = \sum_{i \in SVs} \hat{\alpha}_i y_i K(x, x_i) \ ,$$

where $\hat{\alpha}_i$ are the solutions of the Lagrangian dual problem (2.5). More details about RKHS and kernel methods in machine learning can be found in, e.g., Wahba (1990) and Hofmann, Schölkopf, and Smola (2007).

Some examples of popular kernels used in practice are:

- linear: $K(x,x') =< x,x' >$
- polynomial: $K(x,x') =< x,x' >^d$
- inhomogeneous polynomial: $K(x,x') = (1+ < x,x' >)^d$
- Gaussian or Radial Basis Function: $K(x,x') = \exp(-\|x-x'\|_2^2/2\sigma^2)$
- exponential: $K(x,x') = \exp(-\|x-x'\|_2/2\sigma^2)$
- hybrid: $K(x,x') = (1+ < x,x' >)^d \ \exp(-\gamma\|x-x'\|_2^k)$.

It is worth to notice that the SVM is empirically a very stable algorithm, in the sense that different kernels empirically lead to similar classification accuracies and support vector sets (Schölkopf et al., 1995).

2.3 SVM as a penalized ERM

The SVM in Machine Learning is actually a penalized ERM method (1.9) in Statistical Learning Theory. Consider the L_1 soft-margin formulation as in

(2.3). If the i-th margin constraint is not violated, then the optimal choice for ϵ_i is 0. On the other hand, if $y_i(\beta^t x_i + \beta_0) < 1$, then one has to choose $\epsilon_i \geq 1 - y_i(\beta^t x_i + \beta_0)$ but ϵ_i should be as small as possible, hence resulting in $\epsilon_i = 1 - y_i(\beta^t x_i + \beta_0)$. In summary, $\epsilon_i = \max(0, 1 - y_i f(x_i))$ where $f(x) = \beta^t x + \beta_0$. Thus the solution of (2.3) is equivalent to the solution of

$$\min_{\beta,\beta_0} \quad \frac{1}{n}\sum_{i=1}^{n}(1 - y_i f(x_i))_+ + \lambda\|\beta\|_2^2 \ , \tag{2.7}$$

where $\lambda = 1/(2nC)$ and $(a)_+ = \max(0, a)$ for $a \in \mathbb{R}$. In Statistical Learning, the L_1 soft-margin SVM (2.3) is called the *penalized ERM with hinge loss and ℓ_2-norm penalty*, and the L_2 soft-margin SVM (2.4) is called penalized ERM with *squared* hinge loss and ℓ_2-norm penalty.

In general, the soft-margin SVMs with reproducing kernel K seek the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}[(1 - y_i f(x_i))_+]^q \ + \ \lambda\|h\|_{\mathcal{H}_K}^2 \ , \tag{2.8}$$

where $f(x) = h(x) + \text{constant} \in \mathcal{H}_K + \{1\}$ with $\mathcal{H}_K$ the RKHS, $h \in \mathcal{H}_K$, $\|\cdot\|_{\mathcal{H}_K}$ is the norm in $\mathcal{H}_K$ and $q = 1, 2$. The case $q = 1$ and $q = 2$ are called L_1 and L_2 soft margin SVMs respectively, see (2.3) and (2.4). The tuning parameter λ balances the data fit and the complexity penalty of $f(x)$ measured by $\|h\|_{\mathcal{H}_K}^2$. The expression in (2.7) is a special case of SVM methodology (2.8) when using the linear kernel in $\mathbb{R}^d$ and $q = 1$.

2.4 Good behaviour of SVM

The success of the SVM methodology (2.8) can be explained by two approaches. The first, and the original, explanation is represented by theoretical justification of the SVM in Vapnik's structural risk minimization approach (see Vapnik, 1998). Vapnik's arguments are based on upper bounds of the generalization error in terms of the Vapnik-Chervonenkis (VC) dimension.

The second explanation is based on the identification of the relationship between SVMs and the Bayes rule. Lemmas 2.1 and 3.1 in Lin (2002) show that the asymptotic target functions of SVMs (2.8) are directly related to the Bayes rule: the sign of the minimizer of the theoretical hinge risk

$$\mathbb{E}[((1 - Yf(X))_+)^q] \tag{2.9}$$

equals the sign of $p(x) - 1/2$, with $q = 1, 2$. This is often called Bayes consistency of a loss function. To see this, for $q = 1$, we condition on X to obtain

$$\mathbb{E}_{Y|X}[(1 - Yf(X))_+|X] = p_1(X)(1 - f(X))_+ + p_2(X)(1 + f(X))_+ \ .$$

Clearly, the Bayes predictor (1.14) minimizes the theoretical hinge risk (2.9). Zhang (2004c) shows that minimizing the excess hinge risk also indirectly minimizes the excess prediction error, for the case $q = 1$. This result has been extended by Bartlett et al. (2006), who obtain the following explicit relation:

$$P(Y \neq \text{sign}(f(X))) - P(Y \neq \text{sign}(f^*(X)))$$

$$\leq \ \mathbb{E}[(1 - Yf(X))_+] - \mathbb{E}[(1 - Yf^*(X))_+] \ ,$$

where f^* is the Bayes classifier (1.15).

2.5 Hyperparameters and model selection

The tuning parameter λ (equivalently C) and the parameter appearing in the kernel function (such as d for the polynomial kernel or σ for the Gaussian kernel) are called the *hyperparameters*. Typically they are chosen before solving the risk minimization problem. The problem of determining the optimal hyperparameter values is known as the *model selection* problem. It is usually solved by choosing the hyperparameter values that minimize an estimate of the generalization error. The generalization error is another name for the true risk (1.6).

An estimate of the generalization error can be obtained by several ways. Some examples are: the k-fold and leave-one-out cross validation (CV) methods; the generalized approximate cross validation bound as an estimate of the generalized comparative Kullback-Leibler measure (see Wahba, Lin, and Zhang, 2000); and the VC bound based on the VC dimension (see Vapnik, 1998). The most widely used method in practice is the k-fold CV method (see Hastie et al., 2001, Sect. 7.1). It is applicable to all learning algorithms. Duan, Keerthi, and Poo (2003) confirmed the optimal performance of this method in tuning SVM hyperparameters.

The k-fold CV method is based on resampling the sample. The data D_n are randomly divided into k subsamples of (approximately) equal size n/k. Of the k subsamples, we train the model in $k - 1$ subsamples and leave out the remaining single subsample for validating/testing the model to obtain a test

error. The test error is the proportion of misclassification in the test data. The process is repeated k times leaving out a different subsample each time. An estimate of the generalization error is the average of the k test errors. The optimum hyperparameter values are those that correspond to the estimated generalization error. Clearly, the larger k is, the more computations are needed; but the smaller the difference of the estimated error to the true error, because we use a larger sample size $(k-1)n/k$ for training. In practice, the most common choice for k is either 5 or 10. When $k = n$, it is called leave-one-out CV.

2.6 SVM with ℓ_1-norm penalty

The most commonly used SVM in practice is the L_1 soft-margin formulation (2.3), which is equivalent to hinge loss with ℓ_2-norm penalty (2.7). In this section we consider the hinge loss with ℓ_1-norm penalty:

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} \left[1 - y_i \left(\beta_0 + \sum_{j=1}^{q} \beta_j h_j(x_i) \right) \right]_+ + \lambda \|\beta\|_1 \quad , \tag{2.10}$$

where $\{h_1(x), \ldots, h_q(x)\} =: D$ is a dictionary of basis functions from $\mathbb{R}^d$ to $\mathbb{R}$, and q and λ are the hyperparameters. The solution is denoted as $\hat{\beta}_0(\lambda)$ and $\hat{\beta}(\lambda)$, and the fitted model is

$$\hat{f}(x) = \hat{\beta}_0 + \sum_{j=1}^{q} \hat{\beta}_j \; h_j(x) \; . \tag{2.11}$$

The problem above is equivalent to the following linear optimization problem:

$$\begin{array}{ll} \min_{\beta_0,\beta} & \sum_{i=1}^{n} \left[1 - y_i \left(\beta_0 + \sum_{j=1}^{q} \beta_j h_j(x_i) \right) \right]_+ \\ \text{subject to} & \|\beta\|_1 = |\beta_1| + \cdots + |\beta_q| \le s \; , \end{array} \tag{2.12}$$

where we replace the ℓ_2-norm penalty $\|\beta\|_2^2$ in (2.7) with the ℓ_1-norm penalty $\|\beta\|_1$, with $s = s(\lambda)$ is a hyperparameter.

The main motivation of using the ℓ_1-norm penalty is that it leads to a sparsity scenario of the true model (sparse solution in feature selection). In addition, it has piecewise linear solution paths $\hat{\beta}(s)$, the solution to (2.12) as a function of s.

2.6.1 Feature selection and piecewise linear solution paths

The ℓ_1-norm penalty was first introduced in the regression problem as the LASSO penalty (see Tibshirani, 1996), and in the classification problem as SVM $L1$ problem (see Bradley and Mangasarian, 1998). In the regression problem, in contrast with the ℓ_2-norm penalty which forces all coefficients β_j's to be non-zero and hence selects all the q features, the ℓ_1-norm penalty allows an automatic reduction of the features as λ changes. Setting $\lambda = 0$, or equivalently $s = \infty$, reverses the problem to minimizing non-regularized empirical loss. On the other hand, a very large λ, or equivalently $s = 0$, will shrink the vector $\hat{\beta}$ to 0 and thus leads to an empty model. Making λ sufficiently large, or equivalently s sufficiently small, will cause some of the coefficients $\hat{\beta}_j$ to be exactly zero. In the case that the true model is indeed sparse and/or there are redundant noise variables, ℓ_1-regularization will capture the underlying model much better than ℓ_2-regularization.

Besides the unique property of performing automatic feature selection, the ℓ_1-norm penalty, in combination with the hinge loss, generates piecewise linear solution paths $\hat{\beta}(s)$. That is, $\hat{\beta}(s)$ is linear in each coordinate. This nice property has been shown by Zhu, Rosset, Hastie, and Tibshirani (2004). It facilitates the adaptive selection of the hyperparameter s. This is important, since to get a good fitted model that performs well on future (unseen) data, we need to select an appropriate tuning parameter s. They also propose an efficient algorithm to compute the exact entire solution path $\{\hat{\beta}(s),\ 0 \le s \le \infty\}$. Let us briefly describe their algorithm. Given a training data $\{(x_i, y_i)\}_{i=1}^n$, this algorithm finds some joints $(1, \ldots, J)$. These joints correspond to increasing finite sequences $(s_k)_{k=1}^J$ and $(\hat{\beta}_0(s_k), \hat{\beta}(s_k))_{k=1}^J$ that have the following properties:

(i) $(\hat{\beta}_0(s_k), \hat{\beta}(s_k)) = \arg\min_{\beta_0,\beta} \ \sum_{i=1}^n [1 - y_i(\beta_0 + \sum_{j=1}^q \beta_j h_j(x_i))]_+$

subject to $\|\beta\|_1 \le s_k$,

(ii) $\|\hat{\beta}(s_k)\|_1 = s_k$,

(iii) $\hat{\beta}(s)$ is linear in each coordinate for $s \in (s_{k-1}, s_k]$ with $s_0 = 0$.

Remark 1 We note that recently there are more results on obtaining the piecewise linear solution paths of some risk optimization problems. Efron, Hastie, Johnstone, and Tibshirani (2004) have shown that the LASSO (i.e., regression with least square error loss and ℓ_1-norm penalty) has piecewise linear solution paths too. Motivated by this, Hastie, Rosset, Tibshirani, and Zhu (2004) then

show that the same holds for the hinge loss with ℓ_2-norm penalty. Rosset and Zhu (2007) generalize the characterization of the properties of (loss,penalty) pairs which give piecewise linear solution paths. They also suggest a robust model for classification, namely the so-called *huberized squared hinge loss* in combination with ℓ_1-norm penalty.

Remark 2 Although the ℓ_1-norm penalty has such good properties, it has some limitations as well. Zou and Hastie (2005) argue that there are two major limitations, especially in dealing with microarray data: (1) in the case of several highly correlated and relevant variables, it tends to choose only one or a few of them and shrink the rest to zero, hence it fails to detect a group; (2) in the case that $q > n$, it fails to identify more than n variables since it can assign at most n nonzero coefficients. Zou and Hastie (2005) then propose the so-called *elastic net* regularization that combines good properties of both the ℓ_1-norm and the ℓ_2-norm penalties. Motivated by this, Wang, Zhu, and Zou (2006) propose the so-called *doubly regularized SVM*, i.e., the hinge loss in combination with the elastic net penalty, and develop efficient algorithms to compute the whole solution paths of the model. Wang, Zhu, and Zou (2007) then propose another model, which was motivated by the result of Rosset and Zhu (2007). The model is called *hybrid huberized SVM*, i.e., the huberized hinge loss in combination with the elastic net penalty, and develop the algorithm to compute the entire solution paths of the model.

2.6.2 A toy example

Here, for the sake of illustration, we reproduce the toy simulation in Zhu et al. (2004). We generate 50 training data in each of the two classes: the first class has two standard normal independent inputs x_1 and x_2, with class label $y = 1$; the second class also has two standard normal independent inputs, but conditioned on $4.5 \leq x_1^2 + x_2^2 \leq 8$, with class label $y = -1$. For this toy example, the distribution $P(x, y)$ is known. We can compute the conditional probability of the first class $p_1(x_1, x_2)$ and check that it stays away from $1/2$ (the margin condition is satisfied, see the end of Section 1.3). The scatter plot of the training data is shown in Figure 2.3, an empty circle for the first class $y = 1$ and a solid circle for the second class $y = -1$. Note that the first class almost completely surrounds the second class in a 2-dimensional subspace.

In the original input space $\mathbb{R}^2$, a hyperplane cannot separate the classes, but one can see that the classes are separable in $\mathbb{R}^5$. We use an enlarged feature space corresponding to the polynomial up to second degree kernel, as

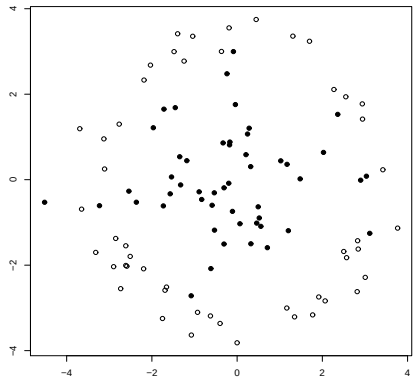

Figure 2.3: *Scatter plot of the 100 training data, empty circle for the class $y = 1$ and solid circle for the class $y = -1$.*

suggested by the representation of the data in Figure 2.3. The dictionary of basis functions is $D = \{\sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2\}$. Here we use the basis representation (2.11) rather than the inhomogeneous polynomial kernel representation $K(x, \tilde{x}) = (< x, \tilde{x} > +1)^2$. This basis representation is the same as the linear kernel representation.

Figure 2.4 shows the piecewise linear solution path $\hat{\beta}(s)$ for our training data with $n = 100$. The two solid lines are for x_1^2 and x_2^2, which are the two relevant features, and the vertical lines indicate the value of $(s_k)_{k=1}^J$.

To select an optimal value of $s \in (s_k)_{k=1}^J$, we use a test data set $\{(x_i', y_i')\}_{i=1}^m$ and choose s_{opt} that gives the smallest test error amongst the fitted models (2.11). We generate $m = 1000$ test data. Figure 2.5 shows the value of the test error along the solution path (s_k). The vertical dashed line indicates the optimal value of the tuning parameter $s_{opt} = s_9 = 0.49$, corresponding to $\hat{\beta}_0 = -1.04$ and $\hat{\beta} = (0, 0, 0, 0.27, 0.23)$. This gives $\hat{f}(x) = -1.04 + 0.27\, x_1^2 + 0.23\, x_2^2$. The horizontal dashed line is the corresponding (minimal) test error 0.07.

Table 2.1 shows the average minimum test errors and their standard error (SE) over 50 simulations of test data and training data. We also conduct the procedure when some noise inputs are added. The table confirms that the ℓ_1-norm penalty performs better than the ℓ_2-norm penalty in the presence of

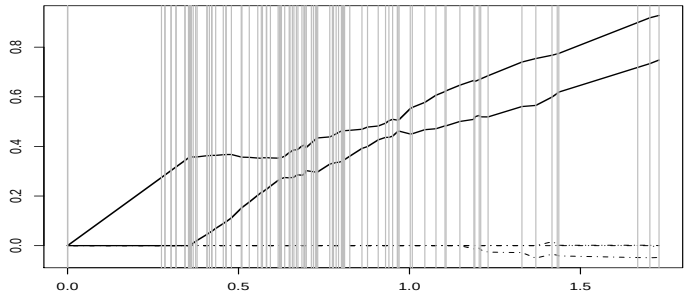

Figure 2.4: *The piecewise linear solution paths* $\hat{\beta}(s)$ *as a function of* $s = \sum_{j=1}^{5} |\hat{\beta}_j|$ *for the training data.*

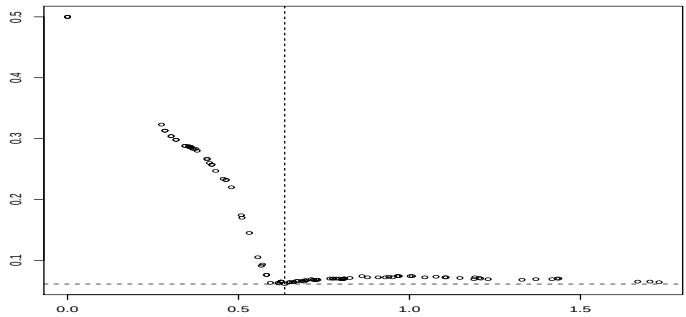

Figure 2.5: *The test error along the solution path* s*, where the vertical dashed line indicates the* s_{opt} *and the horizontal dashed line is the corresponding test error.*

noise.

# noise input	q	ℓ_2-norm penalty		ℓ_1-norm penalty	
		error×1%	# selected q	error×1%	# selected q
0	5	7.3 (0.9)	2.7 (1.0)	7.2 (0.9)	2.7 (0.9)
2	14	7.4 (1.0)	4.7 (3.8)	7.4 (1.0)	4.7 (3.2)
4	27	7.9 (1.4)	7.8 (7.3)	7.3 (0.9)	4.5 (3.7)
6	54	10.9 (1.1)	28.7 (4.1)	7.5 (0.8)	3.6 (2.8)
8	65	13.3 (5.6)	38.2 (4.1)	7.3 (1.0)	3.2 (3.7)
10	88	12.4 (1.7)	41.4 (5.0)	7.9 (2.2)	3.5 (3.7)

Table 2.1: *Simulation on ℓ_1-norm and ℓ_2-norm SVMs*

2.6.3 The relationship between s and λ

Recall that in view of the optimization problems (2.10) and (2.12), we see that there is a 1-1 correspondence between s and λ. That is, $s = 0$ is equivalent to $\lambda = \infty$, and an increasing value of s corresponds to a decreasing value of λ. We now want to answer the following question: Given some s, what is the corresponding λ? Unfortunately we cannot analytically find the explicit relationship between s and λ for any value of s. However, given the sequence $(s_k)_{k=1}^J$, we can calculate the corresponding $(\lambda_k)_{k=1}^J$.

Given $(s_k)_{k=1}^J$ from training data, we have $(\hat{\beta}_0(s_k), \hat{\beta}(s_k))_{k=1}^J$ with properties (i)-(iii) (see Subsection 2.6.1). Define

$$C_n(s) = \sum_{i=1}^{n} [1 - y_i(\hat{\beta}_0(s) + \sum_{j=1}^{q} \hat{\beta}_j(s) h_j(x_i))]_+ \ ,$$

with $s \in (s_k)_{k=1}^J$. We observe that $C_n(s)$ is decreasing in s. Since $\hat{\beta}(s)$ is piecewise linear in s, $C_n(s)$ is also piecewise linear in s. That is,

$$C_n(s) = C_n(s_{k-1}) + \frac{C_n(s_k) - C_n(s_{k-1})}{s_k - s_{k-1}} \quad s \ , \quad s \in (s_{k-1}, s_k] \ .$$

Moreover, $C_n(s)$ is continuous and convex (since hinge function $(x)_+ = \max(0, x)$ is convex).

Note that $s_1 = 0$ and $\hat{\beta}_0(s_1) = \hat{\beta}(s_1) = 0$. Thus, $\lambda_1 = \infty$ and $C_n(s_1) = n$. Since $\hat{\beta}_0(s_k)$ and $\hat{\beta}(s_k)$ solve (2.10) and $\|\hat{\beta}(s_k)\|_1 = s_k$, we have

$$(\hat{\beta}_0(s_k), \hat{\beta}(s_k)) = \arg\min_s \{C_n(s) + \lambda s\} \ .$$

Setting the derivative of $C_n(s)+\lambda s$ (with respect to s) equal to zero, we simply get $\lambda_k = -\,[C_n(s_k) - C_n(s_{k-1})]/[s_k - s_{k-1}],\ k = 2, \ldots, J$.

2.6.4 Simulation on the USPS data set

The US Postal Service (USPS) database contains 7291 training samples and 2007 test samples of handwritten digits collected from mail envelopes in Buffalo (see Schölkopf and Smola, 2002). Each digit is a 16×16 vector with gray-scale values between -1 and 1. The database is publicly available, e.g. at, `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets`. Here we do two binary classification tasks using the model (2.10): to separate 3 from 5, and to separate 3 from 8. Figure 2.6 shows some examples of the digits. We notice that the gray area is rather large in comparison with the white area showing the digits.

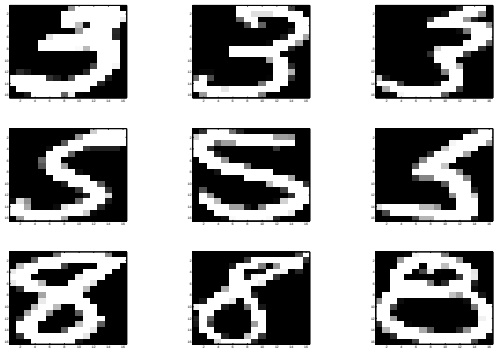

Figure 2.6: *Examples of digits 3, 5 and 8*

We employ the 10-fold CV method and the one-way grid-search method to choose the optimum tuning parameter s_{opt} using the training data only; we keep the test data untouched for estimating the generalization error. Since the algorithm generates the entire solution paths $\hat{\beta}(s)$, the one-way grid-search costs less computationally and is easy to obtain. We first fix a search range S, say $\{1 + i/100\ :\ i = 0, 1, \ldots, 600\}$. In the inner 10-fold CV loop, we calculate the validation errors $e_j(S),\ j = 1, \ldots, 10$. We then take the average of the errors as an estimate of the test error. That is, $s_{\text{opt}} = \arg\min_{s \in S}(1/10)\sum_{j=1}^{10} e_j(S)$. We fit the model to all the training samples to obtain the $\hat{\beta}(s_{\text{opt}})$, then we test the

USPS	# training samples	# test samples	CV error	s_{opt}	# selected features	test error
3 vs 5	731 + 652	198 + 200	0.038	3.76	12 (256)	0.055
3 vs 8	731 + 645	198 + 147	0.006	6.05	31 (256)	0.014

Table 2.2: *Binary SVMs with the ℓ_1-norm penalty on USPS database*

fitted model with the test samples to obtain an estimate for the generalization error. Figure 2.7 gives the solution paths of the selected features for both cases. Table 2.2 gives the results on both tasks. We see that in this sparse scenario, the ℓ_1-norm penalty does a good job, reducing more than 85% of the number of the features in the original input space $\mathbb{R}^{256}$. The test errors are somewhat large, but it is known that the USPS test set is rather difficult. The human error rate for multicategory classification of the digits in the USPS test set is 2.37% (see Chaaban and Scheessele, 2007). The *R*-code used in conducting the simulations is kindly provided by Li Wang and Ji Zhu at University of Michigan (see also Zhu et al., 2004).

2.7 From binary to multicategory SVM

There are many ways to extend binary SVM methods to solve multicategory problems. Recent research shows that not all the binary SVM extensions preserve the desired Bayes consistency property. In general, the extension of ERM-based binary classification methods to solve multicategory cases is not trivial and more involved. In this section we discuss some existing extensions of binary SVM loss to the multicategory case. In particular, we are interested in a SVM binary extension that has the desired Bayes consistency property proposed by Lee et al. (2004). We discuss it in Section 2.8. Their result then leads to the study of Bayes consistency of a wider class of risk minimization formulations, namely large margin-based losses.

Since the success of binary SVM algorithms, many attempts have been made to extend it to multicategory cases. In summary there are two strategies: (1) solving a series of binary problems; (2) considering all of the categories at once. For the first strategy, some popular methods are the one-versus-rest method and the one-versus-one method. The *one-versus-rest* method constructs m binary SVM classifiers. The j-th classifier f_j is trained taking the examples from class j as positive and the examples from all other categories as negative.

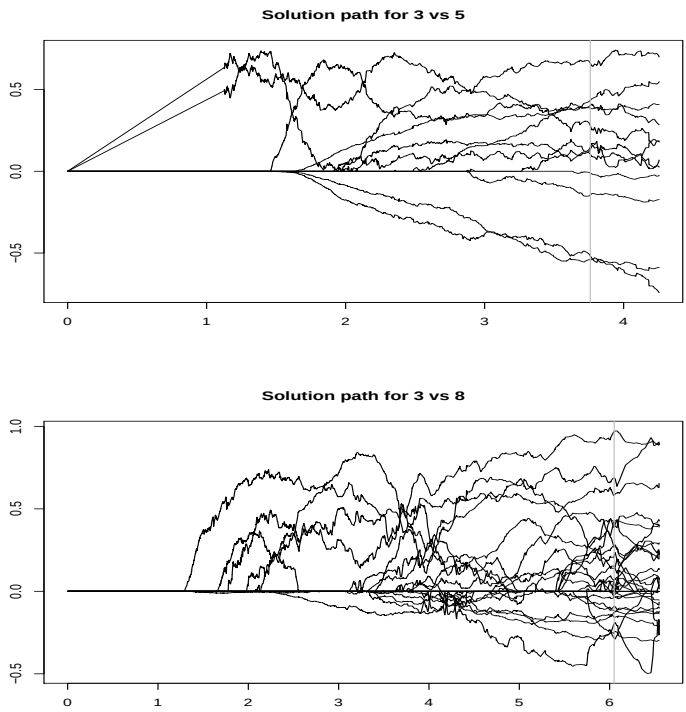

Figure 2.7: *Solution paths of 3 vs 5 and 3 vs 8*

A new example x is assigned to the category with the largest values of $f_j(x)$. The *one-versus-one* method constructs one binary SVM classifier for every pair of distinct categories, that is, all together $m(m-1)/2$ binary SVM classifiers are constructed. The classifier f_{ij} is trained taking the examples from category i as positive and the examples from category j as negative. For a new example x, if f_{ij} classifies x into category i, then the vote for category i is increased by one. Otherwise the vote for category j is increased by one. After each of the $m(m-1)/2$ classifiers makes its vote, x is assigned to the category with the largest number of votes. See Duan and Keerthi (2005) and the references therein for an empirical study of the performance of these methods and theirs variants.

An *all-at-once* strategy for SVM loss has been proposed by some authors.

For example, see Vapnik (2000), Weston and Watkins (1999), Crammer and Singer (2000, 2001), and Guermeur (2002). Roughly speaking, the idea is similar to the one-versus-rest approach but all the m classifiers are obtained by solving one problem. (See Hsu and Lin, 2002, for details of the formulations.) Lee, Lin, and Wahba (2004) (see also Lee, 2002) showed that the relationship of the formulations of the approaches above to the Bayes predictor is not clear from the literature and that they do not always implement the Bayes predictor. They propose a new approach that has good theoretical properties. That is, the defined loss is Bayes consistent and it provides a unifying framework for both equal and unequal misclassification costs, which we discuss in the next section.

2.8 A multi-hinge loss and Bayes consistency

We recall that in the binary case with 0–1 loss, Y is relabeled to be either 1 or -1, the classifier (1.12) is a real valued function and the predictor (1.13) is the sign of the classifier. This can be generalized to an m-category problem as follows: for any $x \in \mathbb{R}^d$, a classifier is a real-valued vector function

$$f = (f_1, \ldots, f_m) \; ; \; f_j : \mathbb{R}^d \to \mathbb{R} \; , \tag{2.13}$$

and a predictor induced by the vector classifier is

$$g = \arg \max_{j=1,\ldots,m} f_j \; . \tag{2.14}$$

The idea in Lee et al. (2004) is to carry over the symmetric representation of the category Y, which now is defined as an m-dimensional vector with 1 in the jth coordinate and $-1/(m-1)$ elsewhere, whenever Y indicates category j. Hence, the zero-sum constraint on the vector classifier is imposed,

$$\sum_{j=1}^{m} f_j = 0 \; .$$

Let $T(Y)$ be the jth row of the general cost matrix (5.1), when $Y = j$. That is, $T(Y)$ is the general misclassification cost vector from category j to the other categories. Analogous to the binary case, when applying RKHS-regularization, each component $f_j(x)$ is considered as an element of a RKHS $\overline{\mathcal{H}}_K = \{1\} + \mathcal{H}_K$, for all $j = 1, \ldots, m$. That is, $f_j(x)$ is expressed as $h_j(x) + b_j$ with $h_j \in \mathcal{H}_K$ and

b_j some constant. To find $f(\cdot) = (f_1(\cdot), \ldots, f_m(\cdot)) \in \Pi_{j=1}^m \overline{\mathcal{H}}_K$ with the zero-sum constraint, the extension of binary SVM methodology to multicategory SVM (MSVM) is to minimize

$$\frac{1}{n}\sum_{i=1}^{n} T(Y_i)^t (f(X_i) - Y_i)_+ \; + \; \frac{\lambda}{2}\sum_{j=1}^{m} ||h_j||^2_{\mathcal{H}_K} \; , \tag{2.15}$$

where $(f(X) - Y)_+$ means $((f_1(X) - Y_1)_+, \ldots, (f_m(X) - Y_m)_+)$ by taking the truncate function $(\cdot)_+ := \max\{0, \cdot\}$ componentwise.

For the *equal cost scenario*, where the misclassification vector $T(Y)^t$ is an m-dimensional vector with 0 in the jth row and 1 elsewhere, whenever $Y = j$, based on (2.15), the loss is now given by

$$l(Y, f(X)) := \sum_{j=1, j\neq Y}^{m} (f_j(X) + \frac{1}{m-1})_+ \; , \tag{2.16}$$

which we call *multi-hinge* loss, and the corresponding theoretical multi-hinge risk is

$$R(f) = \mathbb{E}[\sum_{j=1, j\neq Y}^{m} (f_j(X) + \frac{1}{m-1})_+] \; . \tag{2.17}$$

The binary SVM loss (2.7) is a special case of (2.15) by taking $m = 2$. When $Y = 1$, it is relabeled by $(1, -1)$ and

$$l(1, f(X)) = (0, 1)^t \; ((f_1(X) - 1)_+, (f_2(X) + 1)_+) = (1 - f_1(X))_+ \; .$$

Similarly, when $Y = -1$, it is relabeled by $(-1, 1)$ and $l(-1, f(X)) = (1 + f_1(X))_+$. Thus, (2.16) is identical with the binary SVM loss $(1 - Yf(X))_+$, where f_1 plays the same role as f.

The RKHS-regularization (2.15) has attracted some interest. For example, Lee and Cui (2006) study an algorithm that fits the entire regularization path of the MSVM and Wang and Shen (2007) study the use of the ℓ_1 penalty in place of the ℓ_2 penalty in (2.15). In Chapter 4, we will not study the RKHS-regularization, but we take the minimizer of the empirical multi-hinge loss

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1, j\neq Y_i}^{m} (f_j(X_i) + \frac{1}{m-1})_+$$

over a given class of candidate classifiers $\mathcal{F}$ satisfying a complexity constraint. That is, we do not invoke a penalized ERM approach but the complexity-constraint approach (1.10).

Let $f^* = (f_1^*, \ldots, f_m^*)$ with

$$f_j^* = \begin{cases} 1 & ; \quad \text{if } j = \arg\max_{k=1,\ldots,m} \ p_k \\ -1/(m-1) & ; \quad \text{else .} \end{cases} \tag{2.18}$$

In Lemma 2.8.1 below we show that the vector classifier f^* in (2.18) is the minimizer of multi-hinge risk (2.17) over all possible vector classifiers. We recall that the Bayes predictor g^* in (1.3) for the true multicategory 0–1 loss assigns the new observation x to the category having the largest conditional probability. Using the prediction rule (2.14), we have $g(f^*) = g^*$. That is, the multi-hinge loss (2.16) is Bayes consistent. Clearly, when $m = 2$, (2.18) reduces to (1.15) where f_1 plays the same role as f.

Lemma 2.8.1 can be found in Lee et al. (2004), Zhang (2004a,b), Tewari and Bartlett (2005) and Zou, Zhu, and Hastie (2006). We give our own proof that will be needed in Chapter 4.

Lemma 2.8.1. *The Bayes classifier f^* minimizes multi-hinge risk $R(f)$.*

Proof. We write $L(f(x)) = \mathbb{E}_{Y|X}[l(Y, f(X))|X = x]$ and recall that $p_j(x) = P(Y = j|X = x)$ for all $j = 1, \ldots, m$, and that $f = (f_1, \ldots, f_m)$ with $\sum_{j=1}^m f_j = 0$. Definition (2.16) of the loss and the fact that $\sum_{j=1}^m p_j = 1$ give

$$L(f) = \sum_{j=1}^m p_j \Big(\sum_{k=1, \ k \neq j}^m (f_k + \frac{1}{m-1})_+ \Big) = \sum_{j=1}^m (1 - p_j)(f_j + \frac{1}{m-1})_+ \ .$$

Let $p_k = \max_{j \in \{1,\ldots,m\}} p_j$. Then by (2.18), $f_j^* = -1/(m-1)$ for all $j \neq k$, and $f_k^* = 1$. Let $J^+(k) = \{j \neq k \ : \ f_j \geq -1/(m-1) \ , \ j = 1, \ldots, m\}$ and $J^-(k) = \{j \neq k \ : \ f_j < -1/(m-1) \ , \ j = 1, \ldots, m\}$. Write

$$\begin{aligned} \Delta(f) \ &:= \ L(f) - L(f^*) \\ &= \ \sum_{j \neq k} (1 - p_j)(f_j + \frac{1}{m-1})_+ \ + \ (1 - p_k)(f_k + \frac{1}{m-1})_+ \\ &\quad - \ (1 - p_k)(1 + \frac{1}{m-1}) \ . \end{aligned}$$

We first consider the case $f_k \geq -1/(m-1)$. Here,

$$\Delta(f) = (1 - p_k)(f_k - 1) \ + \ \sum_{j \neq k} (1 - p_j)(f_j + \frac{1}{m-1})_+ \ .$$

The zero-sum constraint $\sum_{j=1}^m f_j = 0$ simply implies $f_k - 1 = -\sum_{j\neq k}(f_j + \frac{1}{m-1})$. Divide the sum into the sets $J^+(k)$ and $J^-(k)$ to obtain

$$\Delta(f) = \sum_{j\in J^+(k)} (p_k - p_j)\ (f_j + \frac{1}{m-1}) \ +\ (1-p_k) \sum_{j\in J^-(k)} |f_j + \frac{1}{m-1}| \ .$$

For the case $f_k < -1/(m-1)$, observe that

$$\frac{m}{m-1} = \sum_{j\neq k}(f_j + \frac{1}{m-1}) + f_k + \frac{1}{m-1} < \sum_{j\neq k}(f_j + \frac{1}{m-1})$$

to obtain

$$\begin{aligned}
\Delta(f) &= (1-p_k)\ (-\frac{m}{m-1}) + \sum_{j\neq k}(1-p_j)(f_j + \frac{1}{m-1})_+ \\
&> (p_k - 1)\sum_{j\neq k}(f_j + \frac{1}{m-1}) + \sum_{j\neq k}(1-p_j)(f_j + \frac{1}{m-1})_+ \\
&= \sum_{j\in J^+(k)} (p_k - p_j)\ (f_j + \frac{1}{m-1}) \ +\ (1-p_k) \sum_{j\in J^-(k)} |f_j + \frac{1}{m-1}| \ .
\end{aligned}$$

In both cases clearly $L(f) - L(f^*)$ is always non-negative since $p_k - p_j$ is non-negative for all $j \neq k$. It follows that

$$R(f) - R(f^*) = \sum_{k=1}^m \int (L(f) - L(f^*))\ \mathbb{1}(p_k = \max_{j=1,\ldots,m} p_j)\ dQ$$

is always non-negative, with Q the unknown marginal distribution of X. ■

One can show that the smallest multi-hinge risk is $m/(m-1)$ times the smallest prediction error; therefore the trivial upper bound of the excess prediction error is $R(f) - \frac{m-1}{m}R(f^*)$. Although there is no such explicit relation like (1.17) as in the binary SVM, Tewari and Bartlett (2005) and Zhang (2004a,b) show that the convergence to zero (in probability) of the excess multi-hinge risk $R(f) - R(f^*)$ implies the convergence to zero with the same rate (in probability) of the excess prediction error $P(g(f(X)) \neq Y) - P(g(f^*(X)) \neq Y)$.

Since the result of Lee et al. (2004), recent research investigates what conditions are needed to formulate natural extensions of binary large margin-based losses that preserve the property above. Zhang (2004a,b) calls it *infinite-sample*

consistency, Tewari and Bartlett (2005) call it *classification calibration* and Zou et al. (2006) call it *admissible loss*. Basically, the idea is (see Zhang, 2004b) to find a vector function f as in (2.13) which minimizes an empirical loss of the form

$$\frac{1}{n}\sum_{i=1}^{n} l_{Y_i}(f(X_i)) \ ,$$

where $l_Y(\cdot) : \mathbb{R}^m \to \mathbb{R}$ is a nonnegative real-valued function indexed by the category $Y \in \{1, \ldots, m\}$ that takes an m-component vector as its parameter. Using the prediction rule (2.14), the multicategory loss function $l_Y(f(X))$ should be chosen such that it favors a vector classifier $f(X)$ whose $f_Y(X)$ is larger than the alternatives $f_j(X)$, $j \neq Y$. That is, it encourages the correct prediction by implicitly maximizing the difference of f_Y and the remaining components. In the set-up of Lee et al. (2004), the multicategory loss is chosen as

$$l_Y(f(X)) := \sum_{j=1, j\neq Y}^{m} (f_j(X) + \frac{1}{m-1})_+ \ .$$

Clearly, minimizing the loss above implies maximizing f_Y because of the zero-sum constraint $\sum_{i=1}^{m} f_j = 0$.

Zhang (2004b), Tewari and Bartlett (2005) and Zou et al. (2006) show that the MSVM-type losses proposed by Weston and Watkins (1999) and Crammer and Singer (2001) are not Bayes consistent. Zhang (2004a,b) also investigate formulation of $l_Y(f(X))$ that can be used to estimate the conditional probabilities p_j.

Chapter 3

A probability bound for ℓ_1-penalized SVMs

This chapter appeared as the article "Classifiers of support vector machine type with
ell_1 complexity regularization" (Tarigan and van de Geer, 2006). Therefore, notation is slightly different and some concepts are introduced for the second time. In addition, we shorten the main title and the title of Section 2 of the article.

3.1 Introduction

Let (X, Y) be random variables, with $X \in \mathcal{X}$ a *feature* and $Y \in \{-1, +1\}$ a binary *label*. The problem is to predict Y given X. A classifier is a function $f : \mathcal{X} \to \mathbf{R}$. Using the classifier f, we predict the label $+1$ when $f(X) \geq 0$, and the label -1 when $f(X) < 0$. Thus, a classification error occurs when $Yf(X) \leq 0$.

Let P be the distribution of the pair (X, Y), and denote the marginal distribution of X by Q. Moreover, write the regression of Y on X as

$$\eta(x) := P(Y = 1|X = x), \ x \in \mathcal{X} \ .$$

Our aim is to find a classifier which makes the correct classification with high probability. The probability of misclassification by the classifier f, or *prediction error* of f, is

$$P(Yf(X) \leq 0) \ .$$

Bayes' (decision) rule is

$$f^* := \begin{cases} +1 & \text{if } \eta \geq 1/2 \\ -1 & \text{if } \eta < 1/2 \ . \end{cases}$$

It is easy to see that the prediction error is the smallest when using Bayes' rule. The function η is however not known. To estimate Bayes' rule, we take a sample from P. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be observed i.i.d. copies of (X, Y). These observations are called the *training set*. The sample size is assumed to be large enough to permit a nonparametric approach to the estimation problem. We assume $n \geq 8$ to avoid nonsense expressions later on.

Let $\mathcal{F}$ be a collection of classifiers. In empirical risk minimization, one chooses the classifier in $\mathcal{F}$ that has the smallest number of misclassifications in the sample (see Vapnik, 2000, 1998). However, if $\mathcal{F}$ is a rich set, this classifier will generally be hard to compute. We will indeed consider a very high-dimensional class $\mathcal{F}$ in this paper. By replacing the number of misclassifications (i.e., 0/1 loss) by *hinge* loss one can overcome computational problems. The *support vector machine* adds an ℓ_2 penalty (or quadratic penalty) to the hinge loss function. We propose instead to employ an ℓ_1 penalty. This yields a computationally feasible complexity regularization method and we show that the procedure can yield estimators that adjust to favorable distributions P.

The empirical hinge loss function is

$$R_n(f) := \frac{1}{n} \sum_{i=1}^{n} l(Y_i f(X_i)) \ ,$$

where $l(z) = (1 - z)_+$, with z_+ denoting the positive part of $z \in \mathbf{R}$. The function l is called the *hinge* function. Define the theoretical hinge loss

$$R(f) := \mathbb{E} R_n(f) = E(l(Yf(X))) \ .$$

Hinge loss is consistent, in the sense that Bayes' decision rule f^* minimizes the theoretical hinge loss

$$f^* = \arg\min_{\text{all } f} R(f)$$

(see Lin, 2002).

For the collection of classifiers $\mathcal{F}$, we choose a subset of a high-dimensional linear space. Consider a given system $\{\psi_k : \ k = 1, \ldots, m\}$ of functions on $\mathcal{X}$. We call $\{\psi_k\}$ the collection of base functions. We assume throughout that $C_Q^2 < \infty$ where

$$C_Q^2 := \max_{1 \le k \le m} \int \psi_k^2 dQ \tag{3.1}$$

is the largest squared $L_2(Q)$ norm of the base functions ψ_k. However, we do not require C_Q to be known.

For $\alpha \in \mathbf{R}^m$ define

$$f_\alpha(x) := \sum_{k=1}^m \alpha_k \psi_k(x), \ x \in \mathcal{X} \ .$$

We then take $\mathcal{F} \subset \{f_\alpha : \ \alpha \in \mathbf{R}^m\}$. The number of base functions ψ_k is allowed to be very large, up to

$$m \le n^D \ , \tag{3.2}$$

for some $D \ge 1$. The support vector machine (SVM) minimizes the empirical hinge loss with, to avoid over-fitting, an ℓ_2 penalty on the coefficients α, that is, a quadratic penalty on the classifier f_α proportional to $\sum \alpha_k^2$ (or a weighted version thereof). In fact, classical SVM's take $\mathcal{F}$ not exactly as (a subset of a) finite-dimensional space, but rather as a reproducing kernel Hilbert space (see also Subsection 3.2.3 for more details). SVM's have been introduced by Boser et al. (1992), and have been applied extensively since then. The book by Schölkopf and Smola (2002) contains a good overview of SVM's and related learning theory.

As variant of the SVM procedure, we propose to add an ℓ_1 complexity penalty, instead of an ℓ_2 complexity penalty, to the empirical hinge loss. The ℓ_1 penalty is proportional to the ℓ_1 norm $\sum_{k=1}^m |\alpha_k|$ of the coefficients. The ℓ_1 penalized minimum hinge loss estimator $\hat{f}_n$ is then defined as

$$\hat{f}_n := \arg \min_{f_\alpha \in \mathcal{F}} \left\{ R_n(f_\alpha) + \hat{\lambda}_n \sum_{k=1}^m |\alpha_k| \right\}, \tag{3.3}$$

where $\hat{\lambda}_n$ is a regularization parameter.

Under Conditions A, B and C below, an appropriate choice for the regularization parameter is

$$\hat{\lambda}_n := c \ \max(\hat{C}_n, 4) \ D\mathbf{K}^2 \sqrt{\frac{\log n}{n}} \ .$$

Here, c is required to be larger than some given universal constant, but is otherwise arbitrary. The quantity $\hat{C}_n$ is the largest empirical L_2 norm of the base functions $\{\psi_k\}$ (see (3.8)), i.e., an estimate of C_Q defined in (3.1). The constant D is from (3.2). Finally, $\mathbf{K}$ is either an assumed given bound K_0 on the the sup norm of the functions in $\mathcal{F}$, or, under some other assumptions, $\mathbf{K} = 1$. More precisely, in Theorem 3.2.1, we require, for technical reasons, that

$$\|f\|_\infty \leq K_0 \ \forall \ f \in \mathcal{F} \ , \tag{3.4}$$

for some constant $K_0 \geq 1$, where

$$\|f\|_\infty := \sup_{x \in \mathcal{X}} |f(x)|$$

denotes the sup norm of the function f. The dependency on K_0 of our results will be given explicitly. On the other hand, in Theorem 3.2.2, we let $\mathbf{K} = 1$. We assume there that for some constant K_n

$$\|f - \tilde{f}\|_\infty \leq \max(K_n \|f - \tilde{f}\|_{1,\nu}, 2), \ \forall \ f, \tilde{f} \in \mathcal{F} \ . \tag{3.5}$$

Here, $\|\cdot\|_{1,\nu}$ denotes the $L_1(\nu)$ norm, with ν some measure on $\mathcal{X}$, depending on the other conditions (Conditions A and B). The constant K_n is allowed to be large depending in fact on the rate an "oracle" would have (see Theorem 3.2.2). The assumption (3.5) is more attractive than (3.4). Nevertheless, we first present the result under condition (3.4), because as it turns out, our proof of Theorem 3.2.2 heavily relies on Theorem 3.2.1.

The ℓ_1 penalty generally leads to sparse representations, i.e., it tends to result in an estimator $\hat{f}_n = \sum_{k=1}^m \hat{\alpha}_k \psi_k$ with only a few non-zero coefficients $\hat{\alpha}_k$ (see Zhu et al., 2004). It is related to soft thresholding (see Donoho, 1995), and is referred to as the LASSO in Tibshirani (1996) and Hastie et al. (2001, Chapter 10.12).

The difference $R(f) - R(f^*)$ is called the hinge *excess risk* at f. We show in Theorem 3.2.1 and Theorem 3.2.2 that the hinge excess risk at $\hat{f}_n$ depends on the smoothness of the boundary of $\{f^* = 1\}$, as well as on the *margin* behavior. The latter quantifies the identifiability of Bayes' decision rule. For definiteness,

we assume it can be summarized in a *margin parameter* (or *noise level*) κ, the definition of the parameter κ being presented in Condition A below.

Whether or not Theorem 3.2.1 and Theorem 3.2.2 produce good rates for the prediction error depends very much on the choice of the system of base functions $\{\psi_k\}$. A more detailed discussion of the problem can be found below, after the statement of the conditions and theorems.

In the literature, the term adaptivity generally refers to attaining (up to log terms) the minimax rate. Adaptivity and minimax rates have been established in Tsybakov (2004), Tsybakov and van de Geer (2005). These papers use empirical risk minimization, which make the methods proposed there difficult to implement. In other work, e.g., Koltchinskii (2001), Koltchinskii and Panchenko (2002), and Koltchinskii (2006), Lugosi and Wegkamp (2004), for example Rademacher complexities are applied. Audibert (2004) establishes adaptivity to the margin for a Gibbs classifier. Scott and Nowak (2006) develop a computationally attractive tree method that adaptively attains minimax rates no faster than $n^{-1/2}$. However, their method only applies to low dimensional input spaces. In this paper, we establish rates in Section 3.2, that depend on margin and complexity, but we will not show that this is in fact (near) adaptation to minimax rates. To show the latter, one has to carefully define the class of probability measures over which one studies the minimax bounds (see also Remark 3.4.1).

This paper is organized as follows. In the next section, we present the conditions (Conditions A, B and C) and main theorems. The results are followed by a discussion of their impact and the relation with averaging classifiers and with kernel SVM's. Here we also address the problem that the hinge excess risk may not be a good approximation of the prediction error.

Section 3.3 takes a closer look at the conditions. The margin condition (Condition A) is shown to follow from assumptions on the amount of mass located near $\eta = 1/2$, and possibly also on the behavior near the boundaries $\eta = 0$ and $\eta = 1$. Also an extension is considered, as well as an extension of Condition B. It can be observed that the choice of the smoothing parameter $\hat{\lambda}_n$ does not depend on the constants appearing in Conditions A and B, and that the procedure adjusts to favorable values of these constants. Condition C is a technical condition on the base functions.

In Section 3.4, we consider an example. The rates obtained there depend on the roughness of the boundary of Bayes' decision rule and on the margin. We consider the case where Bayes' decision rule is a boundary fragment. We

apply Haar wavelets, which provide piecewise constant approximations of the boundary, and are closely related with the binary decision trees studied in Scott and Nowak (2006).

The proofs of the main theorems are given in Section 3.5. Here, we use the toolbox provided by empirical process theory, such as concentration and contraction inequalities. The proofs of the results in Section 3.4 can be found in Section 3.6.

3.2 A probability inequality

3.2.1 Conditions and main theorems

Let ν be some measure on $\mathcal{X}$, and let $\|\cdot\|_{p,\nu}$ be the $L_p(\nu)$ norm ($1 \leq p < \infty$). We assume Conditions A and B below to hold for the same (unknown) measure ν.

Condition A is an identifiability condition, which we refer to as *margin* condition.

Condition A. *There exist constants $\sigma > 0$ and $\kappa \geq 1$, such that for all $f \in \mathcal{F}$,*

$$R(f) - R(f^*) \geq \|f - f^*\|_{1,\nu}^{\kappa} / \sigma^{\kappa} \ . \tag{3.6}$$

The parameter κ is called the *margin* parameter. Its value is generally not known.

Next we impose conditions on the system $\{\psi_k\}$. We use the notation $\psi = (\psi_1, \ldots, \psi_m)^T$ and define $\Sigma_\nu := \int \psi\psi^T d\nu$ (assumed to exist). The smallest eigenvalue of Σ_ν is denoted by ρ_ν^2.

Condition B. *The smallest eigenvalue ρ_ν^2 of Σ_ν is non-zero.*

The value of ρ_ν^2 is generally also unknown.

The last condition puts a normalization on the system of functions $\{\psi_k\}$.

Condition C. *We assume*

$$\max_{1\leq k\leq m} \|\psi_k\|_\infty \leq \sqrt{\frac{n}{\log n}} \ . \tag{3.7}$$

Recall also the requirement that $C_Q < \infty$*, and* $m \leq n^D$ *for some* $D \geq 1$*.*

We now introduce the concepts approximation error, and estimation error. Let $N(\alpha)$ be the number of non-zero coefficients in the vector α, i.e.,

$$N(\alpha) := \#\{\alpha_k \neq 0\}, \ \alpha \in \mathbf{R}^m \ .$$

Given $N \in \{1, \ldots, m\}$, the approximation error is

$$\inf \{R(f_\alpha) - R(f^*) : \ f_\alpha \in \mathcal{F}, \ N(\alpha) = N\} .$$

Let

$$\hat{C}_n^2 := \max_{1\leq k\leq m} \frac{1}{n} \sum_{i=1}^n \psi_k^2(x_i) \tag{3.8}$$

be the empirical version of the constant C_Q^2 defined in (3.1), and let the smoothing parameter be

$$\hat{\lambda}_n := c(\hat{C}_n \vee 4) D\mathbf{K}^2 \sqrt{\frac{\log n}{n}} \ . \tag{3.9}$$

Here, $c \geq c_0$, with c_0 a universal constant. (From Section 3.5, a suitable choice is $c_0 = 864$.) Moreover, here and throughout we use, for $a, b \in \mathbf{R}$, the notation $a \vee b := \max\{a, b\}$. Likewise, $a \wedge b := \min\{a, b\}$. The value of $\mathbf{K}$ will be specified in Theorem 3.2.1 and Theorem 3.2.2. We let λ_n be the theoretical version of $\hat{\lambda}_n$, i.e.

$$\lambda_n := c(C_Q \vee 4) D\mathbf{K}^2 \sqrt{\frac{\log n}{n}} \ . \tag{3.10}$$

As function of n, the value of the theoretical smoothing parameter behaves like $\sqrt{\log n/n}$. This is as in hard- and soft-thresholding (see e.g., Donoho, 1995).

Define (a bound for) the "estimation error" as

$$V_n(N) := 2\delta^{-\frac{1}{2\kappa-1}} \left(18\sigma\lambda_n^2 N D\mathbf{K}/\rho_\nu^2\right)^{\frac{\kappa}{2\kappa-1}} \ , \tag{3.11}$$

where $0 < \delta \leq 1/2$ is fixed, but otherwise arbitrary. Theorem 3.2.1 tells us that the estimation error and approximation error are traded off over all $f_\alpha \in \mathcal{F}$. The trade off is reflected in the quantity

$$\epsilon_n := (1+4\delta) \inf\Big\{ R(f_\alpha) - R(f^*) + V_n(N(\alpha)) + \\ +2\lambda_n \mathbf{K} \sqrt{\frac{\log n}{n}} : \ f_\alpha \in \mathcal{F} \Big\} . \tag{3.12}$$

By the trade off, the ℓ_1 penalized minimum hinge loss estimator adjusts to certain properties of the unknown distribution P. Thus, it has the potential to produce fast rates for the excess risk $R(\hat{f}_n) - R(f^*)$.

Theorem 3.2.1. *Let $\hat{f}_n$ be the ℓ_1 penalized minimum hinge loss estimator defined in (3.3), with regularization parameter $\hat{\lambda}_n$ given in (3.9), where $c \geq c_0$, with c_0 an appropriate universal constant. Suppose that Conditions A, B, and C hold. Assume also that $\mathcal{F} \subset \{f_\alpha : \ \alpha \in \mathbf{R}^m\}$ and*

$$\|f\|_\infty \leq K_0 \ \forall \ f \in \mathcal{F}$$

where $K_0 \geq 1$. Let $\mathbf{K} = K_0$ in the definition (3.9) ((3.10)) of $\hat{\lambda}_n$ (λ_n). Then, for a universal constant c_1,

$$\mathbb{P}\left(R(\hat{f}_n) - R(f^*) > \epsilon_n \right) \leq \frac{c_1}{n^2} . \tag{3.13}$$

The estimation error $V_n(N)$ is a bound for the error due to sampling, when a priori the estimator were required to have only a given set, of cardinality N, of non-zero coefficients. An oracle would choose the optimal set of non-zero coefficients by balancing sparseness and approximation error. In this sense, the theorem shows that the estimator mimics an oracle. Note that the balance is based on the ℓ_0 penalty which counts the number of non-zero coefficients. The similarity of ℓ_0 and ℓ_1 penalties is well known, and in fact goes through for under-determined systems (see Donoho, 2004a,b).

It is of interest to examine the behavior of the estimator for large n. Suppose the constants C_Q, D, K_0, σ and κ, and ρ_ν are fixed (i.e. not depending on the sample size n). Let us call this the *standard situation*. In the standard situation, $\lambda_n^2 = O(\log n/n)$, and the estimation error is of order

$$V_n(N) = O\left(N \log n / n \right)^{\frac{\kappa}{2\kappa - 1}} .$$

For example (for a given N), the worst case corresponds to $\kappa = \infty$, giving $V_n(N) = O(\sqrt{N \log n/n})$. The rates for the estimation error are as in Tsybakov and van de Geer (2005). However, the latter paper deals with empirical risk minimization and prediction error excess risk (see below for the definition of the latter), which means that the rates established there may be quite different from those following from Theorem 3.2.1.

The prediction error excess risk is $P(Yf(X) \leq 0) - P(Yf^*(X) \leq 0)$. Rates of convergence for the hinge excess risk imply the same rates for the prediction error excess risk, as Zhang (2004c) has shown that

$$P(Yf(X) \leq 0) - P(Yf^*(X) \leq 0) \leq R(f) - R(f^*) \tag{3.14}$$

(see also Bartlett et al., 2006). It is easy to see however, that this inequality cannot be reversed. In particular, for $0 < \epsilon < 1$, the classifier $f := \epsilon f^*$ has zero prediction error excess risk, but, in view of Remark 3.3.1 below, hinge excess risk equal to the constant $(1-\epsilon) \int |1-2\eta| dQ$. In the trade off given by Theorem 3.2.1 however, it is the hinge excess risk that enters as approximation error. In other words, this trade off may not reflect the trade off between non-sparseness and prediction error excess risk. For that, we need a margin condition of the form of Condition A, with Bayes' rule f^* replaced by the minimizer of the hinge loss over all $f \in \mathcal{F}$, and in addition an extended version of (3.14). In Section 3.4, we discuss an example where the two types of excess risk will be of the same order of magnitude.

We now consider a variant of Theorem 3.2.1, with essentially weaker conditions.

Theorem 3.2.2. *Let $\hat{f}_n$ be the ℓ_1 penalized minimum hinge loss estimator defined in (3.3), with regularization parameter $\hat{\lambda}_n$ given in (3.9), where $c \geq c_0$, with c_0 an appropriate universal constant. Suppose that Conditions A, B, and C hold. Assume also that $\mathcal{F} \subset \{f_\alpha : \alpha \in \mathbf{R}^m\}$ is a convex set, and that for some constant K_n,*

$$\|f - \tilde{f}\|_\infty \leq (K_n \|f - \tilde{f}\|_{1,\nu}) \vee 2 \quad \forall\ f, \tilde{f} \in \mathcal{F}. \tag{3.15}$$

Then, take $\mathbf{K} = 1$ in (3.9) and (3.10). If

$$2\sigma \epsilon_n^{1/\kappa} K_n \leq 1\ , \tag{3.16}$$

we have, for a universal constant c_1,

$$\mathbb{P}\left(R(\hat{f}_n) - R(f^*) > \epsilon_n\right) \leq \frac{c_1}{n^2}\ . \tag{3.17}$$

Theorem 3.2.2 illustrates that the condition on the sup norm of Theorem 3.2.1 can be weakened. The constant K_n will generally grow with n. For certain systems $\{\psi_k\}_{k=1}^m$, condition (3.16) is met when the number of base functions is small enough, yet large enough to allow to balance approximation error and estimation error. In other cases, inequality (3.15) is a restriction on the allowed linear combinations. When there are no known bounds on κ and on the complexity of the problem, it is actually not possible to verify (3.16). This is however in line with Conditions A and B, which can also not be verified.

In practice, we recommend that the ℓ_1 penalized estimator is computed over **all** f_α, $\alpha \in \mathbf{R}^m$, and that the choice of the smoothing parameter $\hat{\lambda}_n$ is decided upon by applying cross validation.

Remark 3.2.1. Our proof of the two theorems relies on the fact that one has the Lipschitz property

$$|l(y, f(x)) - l(y, \tilde{f}(x))| \leq |f(x) - \tilde{f}(x)|, \text{ for all } f, \tilde{f} \in \mathcal{F} ,$$

where $l(y, f(x)) = (1 - yf(x))_+$ is the hinge loss function. Theorem 3.2.2 moreover uses the convexity of this loss function. The results can be extended to hold for any convex loss function $l(y, f_\alpha(x))$ with this Lipschitz property. The extension can for example be used to derive similar results as in Theorems 3.2.1 and 3.2.2 for robust regression. Loubes and van de Geer (2002) and van de Geer (2003) are essentially along this line, but there fixed design instead of random design is studied.

Results for averages of classifiers, and kernel estimators, using ℓ_1 penalties, call for a different mathematical theory. We will briefly explain why in the next two subsections.

3.2.2 Averaging classifiers

When averaging classifiers, one introduces a collection of base classifiers $\{\psi_k\}$ and forms weighted averages $f_\alpha := \sum_k \alpha_k \psi_k$, where the weights α_k are assumed to be positive and sum up to one. More generally, one may consider arbitrary linear combinations. One often supposes that the base classifiers $\{\psi_k\}$ form a VC class of fixed dimension V (for example stumps). This setup is different from ours in several respects. Firstly, the class of base classifiers may be infinite. However, one may usually replace it by a finite set, virtually without changing the situation. A more severe problem is that Σ_ν generally will have very small

eigenvalues, as the base classifiers are highly correlated. And finally, Bayes' decision rule is generally not well approximated by such averages (unless it is itself one of the base classifiers). This means that generally, the hinge excess risk for such averages is always large. It is not clear however whether the same will be true for the prediction error excess risk. We conclude that Theorem 3.2.1 or Theorem 3.2.2 is not intended for the situation of averaging.

The picture is clearer when one alternatively considers estimating the regression function η, for example using exponential, quadratic or logistic loss. For these loss functions, Blanchard, Lugosi, and Vayatis (2003) have obtained rates of convergence for averaged classifiers. They also consider ℓ_1 penalties, but different loss functions and their results are not in the framework of sparseness. Their rates of convergence follow from the VC dimension of the set of base classifiers.

3.2.3 Kernel representations

It is customary to minimize the hinge loss over a reproducing kernel Hilbert space, with kernel $\mathcal{K}$ (say) on $\mathcal{X} \times \mathcal{X}$. Suppose that $\mathcal{K}$ has eigenexpansion

$$\mathcal{K}(x, \tilde{x}) = \sum_{k=1}^{\infty} \beta_k \phi_k(x) \phi_k(\tilde{x}), \ (x, \tilde{x}) \in \mathcal{X} \times \mathcal{X} \ .$$

Here, $\{\beta_k\}$ are the (non-zero) eigenvalues of $\mathcal{K}$, and $\{\phi_k\}$ are the eigenfunctions. Suppose we use the representation $f_\alpha = \sum_{k=1}^{\infty} \alpha_k \psi_k$, with $\psi_k = \phi_k$. Then in our setup, we employ the penalty

$$\text{pen}(f_\alpha) = \hat{\lambda}_n \sum_k |\alpha_k| \ . \tag{3.18}$$

This penalty is meaningful if Bayes' rule f^* can be well approximated by a sparse representation in terms of the eigenfunctions of the kernel $\mathcal{K}$. The more usual penalty is

$$\text{pen}(f_\alpha) = \lambda \|f_\alpha\|_{\mathcal{K}}^2 \ , \tag{3.19}$$

where λ is a regularization parameter, and where $\|f_\alpha\|_{\mathcal{K}}^2 := \sum_k |\alpha_k|^2 / \beta_k^2$. (see for example Schölkopf and Smola (2002, Chapter 1.5)). The eigenvalues $\{\beta_k\}$ of the kernel typically decrease to zero as $k \to \infty$ (for example, for Gaussian kernels the decay is exponentially fast), so that the penalty in (3.18) is substantially different from the more standard choice (3.19).

We conjecture that for a choice of λ depending only on $\{X_i\}$ (and not on $\{Y_i\}$), the penalty in (3.19) cannot be adaptive to κ in the sense we put forward in Theorem 3.2.1. The reason why we believe this to be true, is that with the quadratic penalty (3.19), a good choice for λ will be such that the estimation error, which depends on κ, is overruled. We do expect that λ in (3.19) can be chosen (rate-)adaptively using cross validation.

We remark that the kernel usually is allowed to depend on a second regularization parameter, called the *width*. For example, for $\mathcal{X}$ a compact in $\mathbf{R}^d$, one may apply the Gaussian kernel

$$\mathcal{K}(x, \tilde{x}) := \exp(-|x - \tilde{x}|^2/h^2) \ ,$$

with width (proportional to) h^d. Both λ and h are often chosen data-dependent. Rates for the general kernels and the penalty (3.19), but with the restriction $\kappa = 1$, are given in Blanchard, Bousquet, and Massart (2004). The situation with Gaussian kernels, penalty (3.19), and known κ is examined in Steinwart and Scovel (2005b).

3.3 On Conditions A, B and C

Conditions A and B together take care that the result follows easily from a probability inequality for the empirical process (see Lemmas 3.5.1, 3.5.2 and 3.5.3). Condition C makes sure that indeed the probability inequality holds (see Lemmas 3.5.4, 3.5.5, 3.5.6 and 3.5.7).

3.3.1 On Condition A

Condition A is a lower bound for the hinge excess risk in terms of the $L_1(\nu)$ norm $\|\cdot\|_{1,\nu}$. We have restricted ourselves to this particular form for ease of exposition. A more general assumption is

$$R(f) - R(f^*) \geq G\left(\|f - f^*\|_{1,\nu}^{1/2}\right), \ \forall \ f \in \mathcal{F} \ ,$$

with $G(\cdot) := \int_0^{\cdot} g(z)dz$ and g a continuous, strictly increasing function on $[0, \infty)$ satisfying $g(0) = 0$. The estimation error will then be

$$V_n(N) = 2\delta H\left(\frac{3\lambda_n\sqrt{2ND\mathbf{K}}}{\delta\rho_\nu}\right) \ ,$$

where

$$H(\cdot) := \int_0^{\cdot} g^{-1}(z)dz \ .$$

This follows from the proof of Lemma 3.5.1, replacing there Lemma 3.5.3 by Young's inequality (for the latter, see e.g. Hardy, Littlewood, and Pólya (1988, Chapter 8.3)).

We restricted ourselves in Condition A to $G(z) = z^{2\kappa}/\sigma^{\kappa}$. It appears in similar form (for prediction error instead of hinge loss) in e.g., Mammen and Tsybakov (1999), Audibert (2004), Tsybakov (2004), Bartlett et al. (2006) and Scott and Nowak (2006). It follows essentially from conditions on the behavior of η near $\{\eta = 1/2\}$ and is therefore often called the *margin* condition, or condition on the *noise level*, see Condition AA below, which was first formulated by Tsybakov (2004).

Condition AA. *There exist constants $C \geq 1$ and $\gamma \geq 0$ such that for all $z > 0$*

$$Q(\{|1 - 2\eta| \leq z\}) \leq (Cz)^{1/\gamma} \ . \tag{3.20}$$

(Here, we use the convention $(Cz)^{1/\gamma} = 1\{z \geq 1/C\}$ for $\gamma = 0$.)

The case where $\gamma = 0$ corresponds to the situation where the function η stays away from $\frac{1}{2}$. It is the situation studied in Blanchard et al. (2004). The larger the value of γ the weaker (3.20) becomes, and for $\gamma = \infty$ it is satisfied for all distributions. If η only takes values very near to $\frac{1}{2}$, Bayes' decision rule is not much better than flipping a fair coin and (3.20) can only hold for large values of γ.

We will see that Condition A is closely intertwined with assumptions on the behavior of η near $\{\eta = 0\}$ and $\{\eta = 1\}$ as well. In principle, values of η near 0 or 1 are favorable as they make the learning problem easier. However, these values make it harder to identify Bayes rule in (say) $\|\cdot\|_{Q,1}$ norm. We show that Condition A holds with $d\nu = \eta(1-\eta)dQ$ and $\kappa = 1 + \gamma$. This is a slight modification of Tsybakov (2004).

Lemma 3.3.1. *Suppose Condition AA is met. Then for all f with $\|f - f^*\|_{\infty} \leq K$,*

$$R(f) - R(f^*) \geq \sigma_K^{-1} \|f - f^*\|_{1,\nu}^{1+\gamma} \ , \tag{3.21}$$

with $d\nu = \eta(1-\eta)dQ$ and with

$$\sigma_K = C \left(\frac{K}{4}(1/\gamma + 1) \right)^{\gamma} (1 + \gamma) \ . \tag{3.22}$$

Thus, Condition A holds with $\sigma = \sigma_K^{1/\kappa}$ *and* $\kappa = 1 + \gamma$.

Proof. By straightforward manipulation, we obtain

$$R(f) - R(f^*) = \int_{-1 \le f \le 1} |(f - f^*)(1 - 2\eta)| dQ$$

$$+ \int_{\substack{f<-1 \\ \eta \le 1/2}} |f - f^*| \eta dQ + \int_{\substack{f<-1 \\ \eta > 1/2}} |(f - f^*)(1 - 2\eta)| dQ$$

$$+ \int_{\substack{f<-1 \\ \eta > 1/2}} (|f| - 1)(1 - \eta) dQ + \int_{\substack{f>1 \\ \eta \le 1/2}} |(f - f^*)(1 - 2\eta)| dQ$$

$$+ \int_{\substack{f>1 \\ \eta \le 1/2}} (|f| - 1)\eta dQ + \int_{\substack{f>1 \\ \eta > 1/2}} |f - f^*|(1 - \eta) dQ \ .$$

This implies the inequality

$$R(f) - R(f^*) \ge \int |f - f^*||1 - 2\eta| d\nu \ .$$

with $d\nu = \eta(1 - \eta)dQ$. Hence, for any $z > 0$,

$$\begin{aligned} R(f) - R(f^*) &\ge \int_{|1-2\eta|>z} |f - f^*||1 - 2\eta| d\nu \\ &\ge z \int_{|1-2\eta|>z} |f - f^*| d\nu \\ &= z\|f - f^*\|_{1,\nu} - z \int_{|1-2\eta|\le z} |f - f^*| d\nu \ . \end{aligned}$$

But, since $\|f - f^*\|_\infty \le K$ and $\eta(1 - \eta) \le 1/4$,

$$\int_{|1-2\eta|\le z} |f - f^*| d\nu \le (K/4)Q(\{|1 - 2\eta| \le z\}) \le (K/4)(Cz)^{1/\gamma} \ ,$$

where we invoked Condition AA. Thus, for all $z > 0$,

$$R(f) - R(f^*) \ge z\|f - f^*\|_{1,\nu} - (K/4)(Cz)^{1/\gamma} z \ .$$

When $\gamma > 0$, we take

$$z = \left(\frac{4\|f - f^*\|_{1,\nu}}{C^{1/\gamma} K(1/\gamma + 1)} \right)^{\gamma} ,$$

and for $\gamma = 0$, we take $z \uparrow 1/C$. We thus arrive at the result of the lemma. ■

Remark 3.3.1. An intermediate result of the proof of Lemma 3.3.1 is that

$$R(f) - R(f^*) \geq \int_{-1 \leq f \leq 1} |(f - f^*)(1 - 2\eta)| dQ ,$$

with equality if $\|f\|_\infty \leq 1$. For an f taking only the values ± 1, the hinge excess risk is therefore equal to twice the prediction error excess risk. (We will use this in Section 3.4.) The proof of Lemma 3.3.1 thus leads also to Zhang (2004c) inequality (see (3.14)).

Remark 3.3.2. The choice $d\nu = \eta(1 - \eta)dQ$ is in our view quite natural, as the variance of $Yf(X)$ is equal to

$$\mathrm{var}(Yf(X)) = 4 \int f^2 \eta(1 - \eta) dQ .$$

This variance generally plays a role in considerations on the empirical process indexed by $\{(x, y) \mapsto yf(x)\}$. There are however also other reasonable candidates for ν. For example, let us define

$$\tau := \min\{\eta, 1 - \eta, |1 - 2\eta|\} ,$$

and suppose that instead of Condition AA, one has for some set S, some $C \geq 1$ and some $\gamma \geq 0$,

$$Q(\{\tau \leq z\} \cap S) \leq (Cz)^{1/\gamma} \ \forall \ z > 0 .$$

Then, from the same arguments as used in the proof of Lemma 3.3.1, one sees that Condition A holds for all $\|f - f^*\|_\infty \leq K$, with $d\nu = 1_S dQ$, $\kappa = 1 + \gamma$ and $\sigma_K = C(K(1/\gamma + 1))^\gamma (1 + \gamma)$. For the set S one may want to take $S = \mathcal{X}$ or $S = \{\eta \notin \{0, 1\}\}$. Recall that ν also plays a role in Condition B, which means we would like to take the set S as large as possible.

Of course, if η stays away from 0 and 1, say $t \leq \eta \leq 1 - t$ for some $0 < t < 1/2$, then the above discussed choices $d\nu = \eta(1 - \eta)dQ$ and $v = Q$ are, up to constants, the same. In Blanchard et al. (2003), it is noted that one may force oneself into such a situation by adding extra noise, namely, by

replacing Y_i by $Y_i' = \omega_i Y_i$ ($Y' = \omega_0 Y$) where $\{\omega_i\}$ is a sequence of independent random variables, with $\mathbb{P}(\omega_i = 1) = 1 - \mathbb{P}(\omega_i = -1) = 3/4$, independent of $\{(X, Y), (X_i, Y_i)\}$. For any classifier f, the prediction error excess risk for predicting Y is equal to twice the prediction error excess risk for predicting its noisy variant Y'. Such a simple relation is generally not true for the hinge excess risk.

In Blanchard et al. (2004), the condition that η stays away from 0 and 1 is imposed as well, in order to enable a precise formulation of a good penalty in that context.

Remark 3.3.3. From Lemma 3.3.1, one sees that the condition that for some K, $\|f - f^*\|_\infty \leq K$ may be needed for Condition A to hold. This is not a priori assumed in Theorem 3.2.2. Therefore, let us mention the following weaker version of Condition A. Suppose for simplicity that the infimum in the definition of ϵ_n is attained, say in f_{α^*}. So

$$f_{\alpha^*} = \arg\min\{R(f_\alpha) - R(f^*) + V_n(N(\alpha)) : \ f_\alpha \in \mathcal{F}\} \ .$$

Then, in Theorem 3.2.2, it suffices to assume (3.6) for those $f \in \mathcal{F}$ with $\|f - f_{\alpha^*}\|_\infty \leq 2$. Thus, if $\|f_{\alpha^*} - f^*\|_\infty \leq K_0$ for some K_0, it suffices to assume (3.6) for those $f \in \mathcal{F}$ with $\|f - f^*\|_\infty \leq K_0 + 2$.

3.3.2 On condition B

The role of the $\|\cdot\|_{1,\nu}$ norm

Recall first that f^* is a renormalized indicator function. In Condition A, the occurrence of the $\|\cdot\|_{1,\nu}$norm is closely related with the fact that for indicator functions, the $L_2(\nu)$ norm $\|\cdot\|_{2,\nu}$ is equal to $\|\cdot\|_{1,\nu}^{1/2}$. In our proof, the $L_2(\nu)$ norm appears as intermediate in the inequality

$$\left(\sum_k |\alpha_k|\right)^2 \leq N(\alpha) K \|f_\alpha\|_{1,\nu}/\rho_\nu^2 \ ,$$

where it is assumed that $\|f_\alpha\|_\infty \leq K$ (see Lemma 3.5.2). In Tarigan and van de Geer (2004), it is shown that when ν is the Lebesgue measure on $[0,1]^d$, one has in fact the following improved inequality for standard compactly supported wavelet systems $\{\psi_k\}$ on $[0,1]^d$,

$$\left(\sum_k |\alpha_k|\right)^2 \leq \ \text{const.} N(\alpha) \|f_\alpha\|_{1,\nu}^2$$

provided that $\{k: \ \alpha_k \neq 0\}$ is the set of all wavelets up to a given resolution level. In this paper, we do not employ this improved variant to avoid digressions. Moreover, as pointed out in Section 3.4, wavelets may not lead to sparse approximations of Bayes' decision rule.

Improving the estimation error bound

The smallest eigenvalue ρ_ν^2 appears in our definition (3.11) of the estimation error. If ρ_ν tends to zero as n tends to infinity, this will slow down the rates. Therefore, it is desirable to have ρ_ν stay away from zero. However, it is as yet unclear to what extend one can find systems $\{\psi_k\}$ with this property and with at the same time good approximating properties.

We now propose a possible improvement of the bound for the estimation error. We replace Condition B by

Condition BB. *For each index set $\mathcal{J} \subset \{1, \ldots, m\}$ there exist a non-negative $\mathcal{N}_{\mathcal{J},\nu}$, such that for all α with $\|f_\alpha\|_\infty \leq K$, one has*

$$\left(\sum_{k \in \mathcal{J}} |\alpha_k| \right)^2 \leq \mathcal{N}_{\mathcal{J},\nu} K \|f_\alpha\|_{1,\nu} \ .$$

With this condition we also have an extension of Lemma 3.5.2.

Theorem 3.2.1 and Theorem 3.2.2 hold with Condition BB instead of B, if we make the following adjustments. Define for $\mathcal{J}(\alpha) := \{k: \ \alpha_k \neq 0\}$,

$$\mathcal{N}_\nu(\alpha) := \mathcal{N}_{\mathcal{J}(\alpha),\nu} \ .$$

Next, let $V_n(\alpha)$ be definition (3.11) with N/ρ_ν^2 replaced by $\mathcal{N}_\nu(\alpha)$:

$$V_n(\alpha) = 2\delta^{-\frac{1}{2\kappa-1}} \left(18\sigma\lambda_n^2 \mathcal{N}_\nu(\alpha) D\mathbf{K} \right)^{\frac{\kappa}{2\kappa-1}} \ .$$

Then, Theorems 3.2.1 and 3.2.2 remain true if in (3.12), we replace $V_n(N(\alpha))$ by $V_n(\alpha)$.

We give an elementary lemma to verify Condition B. It shows that for e.g. systems orthogonal in $L_2(\nu)$, only the eigenvalues of the system chosen by the oracle matter.

Lemma 3.3.2. *Suppose that for some strictly positive weights $\{w_k\}_{k=1}^m$, the smallest eigenvalue of $W\Sigma_\nu W$, with $W = \mathrm{diag}(w_1, \ldots, w_m)$, is equal to one. Then Condition BB holds with*

$$\mathcal{N}_{\mathcal{J},\nu} = \sum_{k\in\mathcal{J}} w_k^2 \ .$$

Proof. Let $\|v\|^2 = v^T v$ denote the squared length of a vector $v \in \mathbf{R}^m$. We know that for all $v \in \mathbf{R}^m$,

$$\|v\|^2 \leq v^T W\Sigma_\nu W v \ , \tag{3.23}$$

as $W\Sigma_\nu W$ has smallest eigenvalue equal to one. So

$$\Big(\sum_{k\in\mathcal{J}} |\alpha_k|\Big)^2 \leq \Big(\sum_{k\in\mathcal{J}} w_k^2\Big)\Big(\sum_{k\in\mathcal{J}} \frac{\alpha_k^2}{w_k^2}\Big) \ ,$$

and, by using (3.23) with $v = W^{-1}\alpha$,

$$\sum_{k\in\mathcal{J}} \alpha_k^2/w_k^2 \ = \ \alpha^T W^{-2}\alpha \ \leq \ \alpha^T\Sigma_\nu\alpha \ = \ \|f_\alpha\|_{2,\nu}^2 \ .$$

Finally, we invoke that $\|f_\alpha\|_{2,\nu}^2 \leq K\|f_\alpha\|_{1,\nu}$ for $\|f_\alpha\|_\infty \leq K$. ■

3.3.3 On Condition C

We need a bound on both the $L_2(Q)$ norm $\|\psi_k\|_{2,Q}$ as well as the sup norm $\|\psi_k\|_\infty$, which holds for all k. This allows us there to apply Bernstein's inequality (see Lemma 3.5.5 and Lemma 3.5.7). The uniform bound $\|\psi_k\|_\infty \leq \sqrt{n/\log n}$ holds for example for most compactly supported wavelet systems on $[0,1]^d$, with per dimension no more than about $(\log_2(n/\log n))/d$ resolution levels. This means that up to constants, the number of functions (wavelets) m in $\{\psi_k\}$ is also no more than about $n/\log n$. This bound on the resolution can mean that the rate of approximation is limited beforehand (see the example of Section 3.4).

In general, we assume polynomial growth $m \leq n^D$. This is quite standard in model selection problems, and rates are generally logarithmic in the a priori number of parameters m.

3.4 An example: boundary fragments

The choice of the base functions $\{\psi_k\}$ plays a crucial role in considerations on their approximating properties. Recall that Bayes' decision rule f^* takes only the values ± 1. Approximating such a function by for example an orthogonal series is not always very natural, as a good approximation might require very many non-zero coefficients. *Wedgelets* (Donoho, 1999) and *curvelets* (Candès and Donoho, 2004) are good alternatives to wavelets. Because these are over-complete systems, our Condition B does not hold, so that the results in Section 3.2 are not applicable. This is the reason why we have chosen to nevertheless study wavelet approximations, in particular Haar functions. As Haar functions consider successive splits of intervals, this approach is related to classification by dyadic trees. Scott and Nowak (2006) derive rates for dyadic decision trees in a context similar to our example.

The main purpose of this section is to illustrate that Theorem 3.2.1 (or Theorem 3.2.2) can produce rates that adjust to roughness of the boundary of Bayes' decision rule, as well as to the margin. We will make some simplifying assumptions (in particular, Assumptions 1-4 below) to avoid digressions.

We consider the case $\mathcal{X} = [0,1]^2$. Moreover, we suppose that f^* is a boundary fragment, i.e., for some function $g^* : [0,1] \to [0,1]$,

$$f^*(x) = \begin{cases} +1 & \text{if } x \in \{(u,v) \in [0,1]^2 : \ g^*(u) \geq v\} \\ -1 & \text{else .} \end{cases} \tag{3.24}$$

We also suppose that the boundary g^* is exactly the set where the regression function η is equal to $1/2$, i.e.,

$$\eta(u,v) \begin{cases} > 1/2 & \text{if } g^*(u) > v \ , \\ = 1/2 & \text{if } g^*(u) = v \ , \\ < 1/2 & \text{if } g^*(u) < v \ . \end{cases} \tag{3.25}$$

Let μ be Lebesgue measure on $[0,1]^2$. For a function g on $[0,1]$, we use the notation

$$\|g\|_{p,\mu} := \|\bar{g}\|_{p,\mu}, \ 1 \leq p < \infty \ ,$$

where $\bar{g}(u,v) = g(u)$, $(u,v) \in [0,1]^2$.

To bound the excess risk, we make the following assumptions.

Assumption 1. The distribution Q of X has density $q = dQ/d\mu$, satisfying for some constant $0 < c_q < \infty$,

$$1/c_q \leq q \leq c_q \ .$$

Assumption 2. For some constant $0 < s \leq 1/2$, $s \leq g^*(u) \leq 1 - s$ for all u.

Assumption 3. For some constants $0 < c_\eta < \infty$ and $0 < \gamma < \infty$,

$$|v - g^*(u)|^\gamma / c_\eta \leq |2\eta(u, v) - 1| \leq c_\eta |v - g^*(u)|^\gamma, \ \forall |\ (u, v) \in [0, 1]^2.$$

Thus, we require in Assumption 3 that for each u, $2\eta(u, v)$ is Hölder continuous with exponent γ at $v = g^*(u)$, and also a Hölder type lower bound on its increments.

Lemma 3.4.1. *Let Assumptions 1-3 hold. Then Condition AA holds with $C = c_\eta(2c_q)^\gamma \vee 1/s$. Moreover, we have for each boundary fragment f_g with boundary g, i.e.,*

$$f_g(x) = \begin{cases} +1 & \text{if } x \in \{(u, v) \in [0, 1]^2 : \ g(u) \geq v\} \\ -1 & \text{else} \ , \end{cases} \tag{3.26}$$

the upper bound

$$R(f_g) - R(f^*) \leq 2c_\eta c_q \|g - g^*\|_{\kappa,\mu}^\kappa \ ,$$

where $\kappa = 1 + \gamma$.

For $r \geq 1$, we define the class of Hölder continuous functions with exponent $1/r$,

$$\mathcal{G}_r(\text{Hölder})$$

$$:= \{g : [0, 1] \to [0, 1] : \ |g(u) - g(\tilde{u})| \leq |u - \tilde{u}|^{1/r}, \ \forall \ u, \tilde{u}\} \ .$$

We call r the *roughness* parameter. We let $\mathcal{G}_0$ be the class of all constant functions on $[0, 1]$.

The next lemma studies the approximation of functions in $\mathcal{G}_r(\text{Hölder})$. Later, we will see that Condition C results in a bound on the resolution level,

and hence on the one-dimensional precision level of our measurements. This precision level, say δ, is defined as the smallest value such that our approximations are piecewise constant on the grid Δ^2, where $\Delta = \{k\delta : \ k = 0, 1, \ldots\}$. In our situation, we will have $\frac{1}{2}\sqrt{\log n/n} \leq \delta \leq \sqrt{\log n/n}$.

For $a > 0$ define $\lfloor a \rfloor$ as the largest integer less than or equal to a. Likewise $\lceil a \rceil$ is the smallest integer bigger than or equal to a.

Lemma 3.4.2. *Suppose $g^* \in \mathcal{G}_r$(Hölder) for some $r \geq 1$. Then for all $\epsilon \geq \delta$, there is a function g^*_ϵ which is constant on the intervals $(u_{j-1}, u_j]$, $u_j = j\epsilon^r$, $j = 1, 2, \ldots, \lceil \epsilon^{-r} \rceil$, and with values in Δ, such that*

$$\|g^* - g^*_\epsilon\|_\infty \leq \epsilon + \delta \ .$$

Let $\{h_{j,l}\}$ be the orthonormal Haar basis of $L_2([0,1], \text{Lebesgue measure})$. So

$$h_{1,1} := 1 \ , \quad h_{1,2} := \mathbb{1}_{[0,1/2)} - \mathbb{1}_{[1/2,1)} \ ,$$

and generally,

$$\begin{aligned} & 2^{-(l-2)/2} h_{j,l} \\ & := \mathbb{1}_{[2(j-1)2^{-l+1}, (2j-1)2^{-l+1})} - \mathbb{1}_{[(2j-1)2^{-l+1}, (2j)2^{-l+1})} \ , \end{aligned} \tag{3.27}$$

for $j = 1, \ldots, 2^{l-2}$ and $l = 2, 3, \ldots$. We use the expansion

$$f_\alpha = \sum_{k=1}^{L} \sum_{l=1}^{L} \sum_i \sum_j \alpha_{i,j,k,l} h_{i,k} h_{j,l} \ ,$$

where $\{h_{j,l}\}$ is the one-dimensional Haar system. We take one-dimensional resolution levels L, with L the largest integer such that $2^{2(L-2)} \leq n/\log n$. This means we have one-dimensional measurement precision $\delta = 2^{-(L-1)} \leq \sqrt{\log n/n}$. As a consequence of Lemma 3.4.2, we thus obtain the following lemma.

Lemma 3.4.3. *Suppose Assumptions 1-3 are met. Let $\mathcal{F} = \{f_\alpha : \ \|f_\alpha\|_\infty \leq K_0\}$, where $K_0 \geq 1$, and let $\delta_n = \sqrt{\log n/n}$. Consider integers N, with $N = JL^2$ and where $2 \leq J \leq \delta_n^{-r}$ is an integer. If $g^* \in \mathcal{G}_r$(Hölder), we have*

$$\inf_{f_\alpha \in \mathcal{F}, N(\alpha) \leq N} R(f_\alpha) - R(f^*) \leq 2^{\kappa+1} c_q c_\eta \times \left(\left(\frac{2L^2}{N} \right)^{\kappa/r} + \delta_n^\kappa \right) \ .$$

Furthermore, if $g^* \in \mathcal{G}_0$,

$$\inf_{f_\alpha \in \mathcal{F}, N(\alpha) = L} R(f_\alpha) - R(f^*) \leq 2 c_q c_\eta \delta_n^\kappa \ .$$

Finally we assume

Assumption 4. For some constant $0 < t < 1/2$, it holds that $t \leq \eta \leq 1 - t$.

See Remark 3.3.2 for a discussion of this assumption.

Theorem 3.4.1. *Suppose that Assumptions 1-4 hold. Let $\mathcal{F} = \{f_\alpha : \|f_\alpha\|_\infty \leq K_0\}$. Consider the standard situation, i.e., the case where the constants K_0, c_q, c_η, s, γ, r and t do not depend on n. Let $\kappa = 1 + \gamma$. Then*

$$\mathbb{E}R(\hat{f}_n) - R(f^*) = O\left(\frac{\log^2 n}{n}\right)^{\beta},$$

with $\beta = \kappa/(2\kappa - 1 + r)$. If $g^* \in \mathcal{G}_0$ we have

$$\mathbb{E}R(\hat{f}_n) - R(f^*) = O\left(\frac{\log^{3/2} n}{n}\right)^{\beta},$$

with

$$\beta = \begin{cases} \kappa/(2\kappa - 1) & \text{if } \kappa \geq 3/2 \\ \kappa/2 & \text{if } \kappa \leq 3/2 \end{cases}.$$

Note that the coefficient β is decreasing in r, so the rates are faster when r is smaller. Moreover, when $r \geq 1$, the coefficient β is increasing in κ, i.e., the rates are then faster for larger values of κ. We conclude that the ℓ_1 penalized minimum hinge loss estimator adjusts to the values of both roughness r and margin parameter κ.

For the case $g^* \in \mathcal{G}_0$, we do not obtain the rate $(\log^{3/2} n/n)^{\kappa/(2\kappa-1)}$ for all values of κ due to limited precision level. When it were known a priori that g^* is constant, one would only have to consider the one-dimensional problem, and a precision level of order $n/\log n$ could be taken. In that case, the rate would be of order $(\log^{3/2} n/n)^{\kappa/(2\kappa-1)}$ for all values of κ.

Remark 3.4.1. Note that for roughness $r \geq 1$, the rates in Theorem 3.4.1 become better as κ increases, which makes the definition of minimax rates a subtle matter. The situation is as in Scott and Nowak (2006). They present a formulation of minimax rates, but their concept of roughness is different from ours.

3.5 Proof of Theorems 3.2.1 and 3.2.2

3.5.1 Proof of Theorem 3.2.1

Let us write $\hat{f}_n = f_{\hat{\alpha}_n}$. Moreover, let

$$I(\alpha) := \sum_{k=1}^{m} |\alpha_k|, \ \alpha \in \mathbf{R}^m \ ,$$

and

$$\nu_n(f) := \sqrt{n}(R_n(f) - R(f)), \ f \in \mathcal{F} \ .$$

Up to and including Lemma 3.5.6, we fix an arbitrary $\alpha^* \in \mathbf{R}^m$, with $f_{\alpha^*} \in \mathcal{F}$. The result of Theorem 3.2.1 then follows from taking the infimum over all such f_{α^*}, of $\epsilon_n(f_{\alpha^*})$ defined below in (3.29). This is done at the very end of this section.

Set $K := 2K_0$. Let Ω^* be the set

$$\Omega^* := \left\{ \sup_{f \in \mathcal{F}} \frac{|\nu_n(f) - \nu_n(f_{\alpha^*})|}{I(\alpha - \alpha^*) + K\sqrt{\frac{\log n}{n}}} \leq \sqrt{n} \ \frac{\lambda_n}{2} \right\} \bigcap \left\{ \frac{\lambda_n^2}{4} \leq \hat{\lambda}_n^2 \leq 4D\lambda_n^2 \right\} \ . \tag{3.28}$$

Recall that

$$\lambda_n = c'(C_Q \vee 4)DK^2\sqrt{\frac{\log n}{n}} \ , \quad \hat{\lambda}_n = c'(\hat{C}_n \vee 4)DK^2\sqrt{\frac{\log n}{n}} \ ,$$

where $4c' = c \geq 4c_0'$. Moreover, we take $c_0' = 216$ ($c_0 = 4c_0'$). We show in Lemmas 3.5.4, 3.5.5 , 3.5.6 and 3.5.7, that under Condition C, the set $\{\omega \notin \Omega^*\}$ has probability at most

$$\bar{c}_1 \exp(-K^2 \log n/2) + 2\exp(-2\log n) \ .$$

Here, $\bar{c}_1$ is an appropriate universal constant. Lemma 3.5.1 below tells us that Conditions A and B yield, on Ω^*, the bound

$$\epsilon_n(f_{\alpha^*}) = (1+4\delta)\left\{ R(f_{\alpha^*}) - R(f^*) + V_n(N(\alpha^*)) + \lambda_n K\sqrt{\frac{\log n}{n}} \right\} \tag{3.29}$$

for the excess risk $R(\hat{f}_n) - R(f^*)$. Lemma 3.5.2 and Lemma 3.5.3 are tools used in Lemma 3.5.1.

Lemma 3.5.1. *Assume Conditions A and B. Then on Ω^*,*

$$R(\hat{f}_n) - R(f^*) \leq \epsilon_n(f_{\alpha^*}) \ , \tag{3.30}$$

where $\epsilon_n(f_{\alpha^*})$ is given in (3.29).

Proof. We use similar arguments as in Loubes and van de Geer (2002), van de Geer (2003) and Tsybakov and van de Geer (2005). Define $N^* = N(\alpha^*)$, and for each $\alpha \in \mathbf{R}^m$,

$$I_1(\alpha) := \sum_{k:\ \alpha_k^* \neq 0} |\alpha_k| \ , \quad I_2(\alpha) := I(\alpha) - I_1(\alpha) = \sum_{k:\ \alpha_k^* = 0} |\alpha_k| \ .$$

Then

$$\begin{aligned}
& R(\hat{f}_n) - R(f_{\alpha^*}) \\
& = -\left(\nu_n(\hat{f}_n) - \nu_n(f_{\alpha^*})\right)/\sqrt{n} + +\hat{\lambda}_n(I(\alpha^*) - I(\hat{\alpha}_n)) \\
& \quad + [R_n(\hat{f}_n) + \hat{\lambda}_n I(\hat{\alpha}_n)] - [R_n(f_{\alpha^*}) + \hat{\lambda}_n I(\alpha^*)] \\
& \leq -\left(\nu_n(\hat{f}_n) - \nu_n(f_{\alpha^*})\right)/\sqrt{n} + \hat{\lambda}_n \left(I(\alpha^*) - I(\hat{\alpha}_n)\right) \ .
\end{aligned}$$

The latter inequality is true because

$$R_n(\hat{f}_n) + \hat{\lambda}_n I(\hat{\alpha}_n) \leq R_n(f_{\alpha^*}) + \hat{\lambda}_n I(\alpha^*) \ ,$$

as $\hat{f}_n$ is the minimizer of the penalized empirical hinge loss over $\mathcal{F}$, and $f_{\alpha^*} \in \mathcal{F}$. Thus, on Ω^*,

$$\begin{aligned}
R(\hat{f}_n) - R(f_{\alpha^*}) \ &\leq \ \frac{\lambda_n}{2}\left(I(\hat{\alpha}_n - \alpha^*) + K\sqrt{\frac{\log n}{n}}\right) + \\
& \qquad + \hat{\lambda}_n \left(I(\alpha^*) - I(\hat{\alpha}_n)\right) \\
& = \ \frac{\lambda_n}{2}\left(I_1(\hat{\alpha}_n - \alpha^*) + I_2(\hat{\alpha}_n) + K\sqrt{\frac{\log n}{n}}\right) + \\
& \qquad + \hat{\lambda}_n \left(I_1(\alpha^*) - I_1(\hat{\alpha}_n) - I_2(\hat{\alpha}_n)\right) \ ,
\end{aligned}$$

where we use that $I_2(\hat{\alpha}_n - \alpha^*) = I_2(\hat{\alpha}_n)$ and $I_2(\alpha^*) = 0$. Since $\lambda_n/2 \leq \hat{\lambda}_n$ on Ω^*, we find on that set that

$$\begin{aligned} R(\hat{f}_n) - R(f_{\alpha^*}) &\leq \frac{\lambda_n}{2}\left(I_1(\hat{\alpha}_n - \alpha^*) + K\sqrt{\frac{\log n}{n}} \right) + \\ &\quad + \hat{\lambda}_n \left(I_1(\alpha^*) - I_1(\hat{\alpha}_n) \right) \ . \end{aligned}$$

Now use that $I_1(\alpha^*) - I_1(\hat{\alpha}_n) \leq I_1(\hat{\alpha}_n - \alpha^*)$, and that on Ω^*, $\hat{\lambda}_n \leq 2\sqrt{D}\lambda_n$. Invoking the bounds $1/2 \leq 1 \leq \sqrt{D}$, we obtain that on Ω^*,

$$R(\hat{f}_n) - R(f_{\alpha^*}) \leq 3\lambda_n \sqrt{D} I_1(\hat{\alpha}_n - \alpha^*) + \lambda_n K \sqrt{\log n/n} \ .$$

We now use Lemma 3.5.2, and the triangle inequality, to arrive at

$$\begin{aligned} &R(\hat{f}_n) - R(f_{\alpha^*}) \leq 3\lambda_n \sqrt{N^* DK} \|\hat{f}_n - f^*\|_{1,\nu}^{1/2}/\rho_\nu + \\ &+3\lambda_n \sqrt{N^* DK} \|f_{\alpha^*} - f^*\|_{1,\nu}^{1/2}/\rho_\nu + \lambda_n K \sqrt{(\log n)/n} \ . \end{aligned} \tag{3.31}$$

Let us use the short hand notation

$$\hat{d} := R(\hat{f}_n) - R(f^*), \ d^* := R(f_{\alpha^*}) - R(f^*) \ .$$

Then the application of Lemma 3.5.3 below (with $v = 3\lambda_n\sqrt{N^*DK}/\rho_\nu$ and respectively $t = \|\hat{f}_n - f^*\|_{1,\nu}^{1/2}$ and $t = \|f_{\alpha^*} - f^*\|_{1,\nu}^{1/2}$), and Condition A, to the first two terms in the right hand side of (3.31) yields

$$\hat{d} \leq \delta(\hat{d} + d^*) + 2\delta \left(\frac{3\lambda_n}{\rho_\nu} \frac{\sqrt{\sigma N^* DK}}{\delta} \right)^{\frac{2\kappa}{2\kappa-1}} + \lambda_n K \sqrt{\frac{\log n}{n}} \ .$$

Since for $\delta \leq 1/2$, the inequality $(1+\delta)/(1-\delta) \leq 1 + 4\delta$ holds, we now have shown that

$$\begin{aligned} \hat{d} &\leq (1+4\delta) \left\{ d^* + 2\delta \left(\frac{3\lambda_n}{\rho_\nu} \frac{\sqrt{\sigma N^* DK}}{\delta} \right)^{\frac{2\kappa}{2\kappa-1}} + \lambda_n K \sqrt{\frac{\log n}{n}} \right\} \\ &= (1+4\delta) \left\{ d^* + 2\delta^{-\frac{1}{2\kappa-1}} \left[\frac{9\sigma\lambda_n^2 N^* DK}{\rho_\nu^2} \right]^{\frac{\kappa}{2\kappa-1}} + \lambda_n K \sqrt{\frac{\log n}{n}} \right\} \ . \end{aligned}$$

■

Lemma 3.5.2. *Assume Condition B. Let $\mathcal{J} \subset \{1, \ldots, m\}$ be some index set, with cardinality $N = |\mathcal{J}|$. Then for $\|f_\alpha\|_\infty \leq K$,*

$$\left(\sum_{k \in \mathcal{J}} |\alpha_k| \right)^2 \leq NK \|f_\alpha\|_{1,\nu} / \rho_\nu^2 \ .$$

Proof. Clearly

$$\left(\sum_{k \in \mathcal{J}} |\alpha_k| \right)^2 \leq N \sum_k \alpha_k^2 \ .$$

But

$$\sum_k \alpha_k^2 = \alpha^T \alpha \leq \alpha^T \Sigma_\nu \alpha / \rho_\nu^2 = \|f_\alpha\|_{2,\nu}^2 / \rho_\nu^2 \leq K \|f_\alpha\|_{1,\nu} / \rho_\nu^2 \ .$$

■

Lemma 3.5.1 applies Lemma 3.5.3 below. Such inequalities are standardly used in the recent classification literature (see e.g., Tsybakov and van de Geer, 2005). Lemma 3.5.3 is an immediate consequence of Young's inequality (see e.g., Hardy et al., 1988, Chapter 8.3), using some straightforward bounds to simplify the expressions.

Lemma 3.5.3. *For all $\kappa \geq 1$, and all positive v, t and δ, it holds that*

$$vt \leq \delta t^{2\kappa} / \sigma^\kappa + \delta^{-\frac{1}{2\kappa - 1}} (\sigma v^2)^{\frac{\kappa}{2\kappa - 1}} \ .$$

We now will show that the set Ω^* has probability close to one. To this end, a concentration inequality will be applied Theorem 3.5.1 is from Massart (2000) who improves the constants from Ledoux (1996). These authors actually assume certain measurability conditions. To avoid digressions, we will skip all measurability issues.

Theorem 3.5.1. *Let $Z_1, \ldots, Z_n$ be i.i.d. copies of a random variable $Z \in \mathcal{Z}$. Let Γ be a class of real-valued functions on $\mathcal{Z}$ satisfying $\sup_z |\gamma(z)| \leq K$ for all $\gamma \in \Gamma$. Define*

$$\mathbf{Z} := \sup_{\gamma \in \Gamma} \left| \frac{1}{n} \sum_{i=1}^n \{\gamma(Z_i) - \mathbb{E}\gamma(Z_i)\} \right| \tag{3.32}$$

and

$$\tau^2 := \sup_{\gamma\in\Gamma} \mathrm{Var}(\gamma(Z)) \ . \tag{3.33}$$

Then for any positive z,

$$\mathbb{P}\left(\mathbf{Z} \geq 2\mathbb{E}\mathbf{Z} + \tau\sqrt{8z/n} + 69Kz/(2n)\right) \leq \exp(-z) \ . \tag{3.34}$$

Lemma 3.5.4. *Define $\mathcal{F}_M := \{f_\alpha \in \mathcal{F} : I(\alpha - \alpha^*) \leq M,\ \|f_\alpha - f_{\alpha^*}\|_\infty \leq K\}$, and*

$$\mathbf{Z}_M := \sup_{f_\alpha\in\mathcal{F}_M} |\nu_n(f_\alpha) - \nu_n(f_{\alpha^*})| / \sqrt{n} \ .$$

Then for all M satisfying $C_Q M \geq K\sqrt{(\log n)/n}$ (where C_Q is given in (3.1)), we have

$$\mathbb{P}\left(\mathbf{Z}_M \geq 2\mathbb{E}\mathbf{Z}_M + 36K^2 C_Q M\sqrt{\frac{\log n}{n}}\right)$$

$$\leq \exp(-(C_Q^2 M^2 \vee K^2)\log n) \ .$$

Proof. In Theorem 5.1, we take

$$\Gamma = \{\gamma_\alpha : \ f_\alpha \in \mathcal{F}_M\} \ ,$$

where

$$\gamma_\alpha(x, y) = l(yf_\alpha(x)) - l(yf_{\alpha*}(x)) \ ,$$

and where $l(z) = (1 - z)_+$ is the hinge function. Since l is Lipschitz, we have

$$|\gamma_\alpha(x, y)| \leq |f_\alpha(x) - f_{\alpha^*}(x)| \ .$$

Note first that this implies $\mathbf{Z}_M \leq K$, so that it suffices to consider values M with for $C_Q M \leq K\sqrt{n/\log n}$. The Lipschitz property also implies that $\tau^2 \leq \sup_{f_\alpha\in\mathcal{F}_M} \|f_\alpha - f_{\alpha^*}\|_{2,Q}^2$. So $\tau \leq C_Q M \wedge K := \tau_1$. We now take $z = (C_Q^2 M^2 \vee K^2)\ \log n$.

Then, for $K \leq C_Q M \leq K\sqrt{n/\log n}$,

$$\tau_1\sqrt{8z/n} + 69Kz/(2n)$$

$$
\begin{aligned}
&= KC_QM\sqrt{8\log n/n} + 69KC_Q^2M^2\log n/(2n) \\
&\leq 3KC_QM\sqrt{\log n/n} + (69/2)K^2C_QM\sqrt{\log n/n} \\
&\leq 36K^2C_QM\sqrt{\log n/n}\ ,
\end{aligned}
$$

where we used that $K \geq 2$. Moreover, for $K\sqrt{\log n/n} \leq C_QM \leq K$,

$$
\begin{aligned}
&\tau_1\sqrt{8z/n} + 69Kz/(2n) \\
&= KC_QM\sqrt{8\log n/n} + 69KK^2\log n/(2n) \\
&\leq 3KC_QM\sqrt{\log n/n} + (69/2)K^2C_QM\sqrt{\log n/n} \\
&\leq 36K^2C_QM\sqrt{\log n/n}\ .
\end{aligned}
$$

The result thus follows from Theorem 5.1. ■

Lemma 3.5.5. *Suppose Condition C is met. For $\mathbf{Z}_M$ defined in Lemma 3.5.4, it holds that*

$$
\mathbb{E}\mathbf{Z}_M \leq 36(C_Q \vee 4)DM\sqrt{\log n/n}\ . \tag{3.35}
$$

Proof. This follows from similar arguments as in van de Geer (2003), using the fact that the function $z \mapsto l(z) = (1-z)_+$, $z \in \mathbf{R}$ is Lipschitz. Let us briefly summarize these arguments. Let $\epsilon_1, \ldots, \epsilon_n$ be a Rademacher sequence independent of $(X_1, Y_1), \ldots, (X_n, Y_n)$. By symmetrization and the contraction inequality (see Ledoux and Talagrand, 1991), we find

$$
\begin{aligned}
\mathbb{E}\mathbf{Z}_M &\leq 4\mathbb{E}\left(\sup_{f_\alpha \in \mathcal{F}_M}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(f_\alpha(X_i) - f_{\alpha^*}(X_i))\right|\right) \\
&\leq 4M\mathbb{E}\left(\max_{k=1,\ldots,m}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\psi_k(X_i)\right|\right)\ .
\end{aligned}
$$

By Bernstein's inequality (see e.g. Shorack and Wellner, 1986, page 855), we know that for any $z > 0$,

$$
\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^n \epsilon_i\psi_k(X_i)\right| \geq z\right) \leq 2\exp\left(-\frac{nz^2}{2z\|\psi_k\|_\infty/3 + 2\|\psi_k\|_{2,Q}^2}\right)\ .
$$

Use Condition C, which says that for all k, $\|\psi_k\|_\infty \leq \sqrt{\frac{n}{\log n}}$. Moreover, $\|\psi_k\|_{2,Q} \leq C_Q$ for all k. We find for all $z \geq 1$, using $m \leq n^D$, and $D \geq 1$,

$$\mathbb{P}\left(\max_{k=1,\ldots,m}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \psi_k(X_i)\right| \geq 3(C_Q \vee 4)Dz\sqrt{\frac{\log n}{n}}\right)$$

$$\leq 2m\exp(-\frac{9(C_Q \vee 4)^2 D^2 z \log n}{2(C_Q \vee 4)D + 2C_Q^2})$$

$$\leq 2m\exp(-9Dz\log n/4) \ \leq \ 2\exp(-5Dz\log n/4) \ .$$

Now, for any positive random variable U and any positive t,

$$\mathbb{E}(U) = \int_0^\infty \mathbb{P}(U \geq z)dz \leq t\left(1 + \int_1^\infty \mathbb{P}(U \geq tz)dz\right) \ .$$

Apply this with

$$U = \max_{k=1,\ldots,m}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \psi_k(X_i)\right| \ .$$

It follows that

$$\begin{aligned}\mathbb{E}\mathbf{Z}_M &\leq 3(1+2)4M(C_Q \vee 4)D\sqrt{\log n/n} \\ &= 36M(C_Q \vee 4)D\sqrt{\log n/n} \ ,\end{aligned}$$

where we used the bound

$$\exp(-5D\log n/4)/(5D\log n/4) \leq 1 \ ,$$

because $D \geq 1$ and $n \geq 8$. ■

Next, we show that for $\lambda_n \geq 216(C_Q \vee 4)DK^2\sqrt{\log n/n}$, the set

$$\left\{\left|\nu_n(\hat{f}_n) - \nu_n(f_{\alpha^*})\right| \leq \sqrt{n}\ \frac{\lambda_n}{2}\left(I(\hat{\alpha}_n - \alpha^*) + K\sqrt{\log n/n}\right)\right\}$$

has probability at least $1 - \bar{c}_1\exp(-K^2\log n/2)$.

To simplify the exposition, we write

$$\Delta_n(f_\alpha) := \frac{\nu_n(f_\alpha) - \nu_n(f_{\alpha^*})}{I(\alpha - \alpha^*) + K\sqrt{\log n/n}} \ ,$$

$$B_n := 108(C_Q \vee 4)DK^2\sqrt{\log n} \ .$$

Lemma 3.5.6. *Suppose Condition C is met. We have for a universal constant* $\bar{c}_1$

$$\mathbb{P}\Big(\sup_{\substack{f_\alpha \in \mathcal{F} \\ \|f_\alpha - f_{\alpha^*}\|_\infty \leq K}} |\Delta_n(f_\alpha)| > B_n \Big) \leq \bar{c}_1 \exp(-\frac{K^2 \log n}{2}) \ . \tag{3.36}$$

Proof. This follows from the peeling device, which is designed to establish bounds for weighted empirical process, as discussed in van de Geer (2000, Chapter 5.3). We split up $\mathbf{R}^m$ into the sets

$$S_1 = \{\alpha : \ C_Q I(\alpha - \alpha^*) \leq K\sqrt{\log n/n}\} \ ,$$

$$S_2 = \{\alpha : \ K\sqrt{\log n/n} < C_Q I(\alpha - \alpha^*) \leq K\}$$

$$\subseteq \bigcup_{j=0}^{j_0} \{\alpha : \ 2^{-(j+1)} K < C_Q I(\alpha - \alpha^*) \leq 2^{-j} K\}$$

with $2^{-j_0} < \sqrt{\log n/n}$, and

$$S_3 = \{\alpha : \ C_Q I(\alpha - \alpha^*) \geq K\}$$

$$= \bigcup_{j=1}^{\infty} \{\alpha : \ 2^{j-1} K < C_Q I(\alpha - \alpha^*) \leq 2^j K\} \ .$$

The combination of Lemma 3.5.4 and Lemma 3.5.5, and invoking $K \geq 2$, yields that for $C_Q M \geq K\sqrt{\log n/n}$,

$$\mathbb{P}\left(\mathbf{Z}_M \geq 54(C_Q \vee 4) D M K^2 \sqrt{\frac{\log n}{n}} \right)$$
$$\leq \exp(-(C_Q^2 M^2 \vee K^2) \log n) \ . \tag{3.37}$$

We find on the set S_1,

$$\mathbb{P}\left(\sup_{f_\alpha \in \mathcal{F}, \alpha \in S_1} |\Delta_n(f_\alpha)| \geq B_n \right)$$

$$\leq \mathbb{P}\left(\sup_{f_\alpha \in \mathcal{F}, \alpha \in S_1} |\nu_n(f_\alpha) - \nu_n(f_{\alpha^*})| \geq K\sqrt{\log n/n} \ B_n \right)$$

$$\leq \exp(-K^2 \log n) \ .$$

Next we consider the set S_2. Take j_0 as the smallest integer such that $2^{-j_0} < \sqrt{\log n/n}$. Then from (3.37),

$$\mathbb{P}\left(\sup_{f_\alpha \in \mathcal{F}, \alpha \in S_2} |\Delta_n(f_\alpha)| \geq B_n \right)$$

$$\leq \sum_{j=0}^{j_0} \mathbb{P}\Big(\sup_{\substack{f_\alpha \in \mathcal{F} \\ C_Q I(\alpha - \alpha^*) \leq 2^{-j} K}} |\nu_n(f_\alpha) - \nu_n(f_{\alpha^*})| \ \geq \ (2^{-j} K) \ B_n/2 \Big)$$

$$\leq \log n \ \exp(-K^2 \log n) \ ,$$

as for $n \geq 8$, $j_0 + 1 \leq \log n$.

Finally, we consider the set S_3. We find

$$\mathbb{P}\left(\sup_{f_\alpha \in \mathcal{F}, \alpha \in S_3} |\Delta_n(f_\alpha)| \geq B_n \right)$$

$$\leq \sum_{j=1}^{\infty} \mathbb{P}\Big(\sup_{\substack{f_\alpha \in \mathcal{F} \\ C_Q I(\alpha - \alpha^*) \leq 2^{j} K}} |\nu_n(f_\alpha) - \nu_n(f_{\alpha^*})| \ \geq \ (2^{j} K) B_n/2 \Big)$$

$$\leq \sum_{j=1}^{\infty} \exp(-2^{2j} K^2 \log n) \ .$$

We conclude that

$$\mathbb{P}\left(\sup_{f_\alpha \in \mathcal{F}} |\Delta_n(f_\alpha)| \geq B_n \right)$$

$$\leq \exp(-K^2 \log n) + \log n \exp(-K^2 \log n) + \sum_{j=1}^{\infty} \exp(-2^{2j} K^2 \log n)$$

$$\leq \bar{c}_1 \exp\left(-\frac{K^2 \log n}{2} \right) \ ,$$

for a universal constant $\bar{c}_1$. ■

Lemma 3.5.7. *Suppose Condition C holds. Then*

$$\mathbb{P}\left(\frac{\lambda_n^2}{4} \leq \hat{\lambda}_n^2 \leq 4D\lambda_n^2\right) \geq 1 - 2\exp(-2\log n) \ . \tag{3.38}$$

Proof. Recall the definitions

$$\lambda_n = c' \ (C_Q \vee 4)DK^2\sqrt{\frac{\log n}{n}}, \ \hat{\lambda}_n = c' \ (\hat{C}_n \vee 4)DK^2\sqrt{\frac{\log n}{n}} \ ,$$

and

$$C_Q^2 = \max_{1\leq k\leq m} \ \|\psi_k\|_{2,Q}^2, \ \hat{C}_n^2 = \max_{1\leq k\leq m} \ \frac{1}{n}\sum_{i=1}^n \psi_k^2(X_i) \ .$$

We first bound the probability of the set $\{\hat{\lambda}_n^2 < \lambda_n^2/4\}$. We consider two cases: $C_Q \leq 4$ and $C_Q > 4$. If $C_Q \leq 4$, we have

$$\hat{C}_n \vee 4 = \begin{cases} \hat{C}_n \geq C_Q \vee 4 & \text{if } \hat{C}_n > 4 \\ 4 = C_Q \vee 4 & \text{if } \hat{C}_n \leq 4 \ . \end{cases}$$

In other words, if $C_Q \leq 4$, we have $\hat{\lambda}_n \geq \lambda_n$, and so the set $\{\hat{\lambda}_n^2 < \lambda_n^2/4\}$ has probability zero.

Now, let $\psi_{\max}$ a base function for which the maximum $L_2(Q)$ norm is attained. Then clearly,

$$\hat{C}_n^2 \geq \|\psi_{\max}\|_{2,Q_n}^2 \ .$$

From Bernstein's inequality (see e.g. Shorack and Wellner, 1986, page 855), we now establish that

$$\begin{aligned} \mathbb{P}\left(\hat{C}_n^2 < C_Q^2/4\right) &\leq \mathbb{P}\left(\|\psi_{\max}\|_{2,Q_n}^2 - \|\psi_{\max}\|_{2,Q}^2 < -3C_Q^2/4\right) \\ &\leq \exp\left[-\frac{n9C_Q^4/16}{C_Q^2\|\psi_{\max}^2\|_\infty/2 + 2C_Q^2\|\psi_{\max}^2\|_{2,Q}^2}\right] \\ &\leq \exp[-9C_Q^2\log n/40] \leq \exp[-C_Q^2\log n/8] \ , \end{aligned}$$

since, $\|\psi_{\max}^2\|_{2,Q} = C_Q^2$, and by Condition C, $\|\psi_{\max}^2\|_\infty \leq n/\log n$. With $C_Q > 4$, this gives

$$\mathbb{P}\left(\hat{C}_n^2 < C_Q^2/4\right) \leq \exp(-2\log n) \ .$$

Next, we bound the probability of $\{\hat{\lambda}_n^2 > 4D\lambda_n^2\}$. We consider the cases $\hat{C}_n \leq 4$ and $\hat{C}_n > 4$. For $\hat{C}_n \leq 4$, one has

$$(\hat{C}_n \vee 4)^2 = 4^2 \leq 4D(C_Q \vee 4)^2 \ ,$$

so then $\{\hat{\lambda}_n^2 \leq 4D\lambda_n^2\}$ holds trivially.

Now the case $\hat{C}_n > 4$. Note first that, again by Bernstein's inequality

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\leq k\leq m} \|\psi_k\|_{2,Q_n}^2 - \|\psi_k\|_{2,Q}^2 > 3D(C_Q \vee 4)^2\right) \\
&\leq m \ \exp\left[-\frac{9D^2(C_Q \vee 4)^4 \log n}{2D(C_Q \vee 4)^2 + 2C_Q^2}\right] \\
&\leq m \exp[-9D(C_Q \vee 4)^2 \log n/4] \\
&\leq \exp[-5D(C_Q \vee 4)^2 \log n/4] \leq \exp[-2\log n] \ .
\end{aligned}$$

But clearly, if

$$\max_{1\leq k\leq m} \|\psi_k\|_{2,Q_n}^2 - \|\psi_k\|_{2,Q}^2 \leq 3D(C_Q \vee 4)^2 \ ,$$

one has for $\hat{C}_n > 4$

$$(\hat{C}_n \vee 4)^2 = \hat{C}_n^2 \leq 3D(C_Q \vee 4)^2 + C_Q^2 \leq 4D(C_Q \vee 4)^2.$$

■

To conclude the proof of Theorem 3.2.1, we observe that for any $z \geq 0$

$$\mathbb{P}\left(R(\hat{f}_n) - R(f^*) > z\right) = \lim_{z_t \downarrow z} \mathbb{P}\left(R(\hat{f}_n) - R(f^*) > z_t\right) \ ,$$

because a distribution function is right-continuous. Let ϵ_n be defined as in (3.12), i.e.,

$$\epsilon_n = \inf\{\epsilon_n(f_{\alpha^*}) : \ \alpha^* \in \mathbf{R}^m, \ f_{\alpha^*} \in \mathcal{F}\} .$$

We may then write

$$\epsilon_n = \lim_{t\to\infty} \epsilon_{n,t} \ ,$$

for a sequence $\{\epsilon_{n,t}\}_{t=1}^{\infty}$, with

$$\epsilon_{n,t} = \epsilon_n(f_{\alpha_t^*}) \ ,$$

for some $\alpha_t^* \in \mathbf{R}^m$, $f_{\alpha_t^*} \in \mathcal{F}$, $t = 1, 2, \ldots$. Therefore, by Lemmas 3.5.1–3.5.7,

$$\begin{aligned} \mathbb{P}\left(R(\hat{f}_n) - R(f^*) > \epsilon_n\right) &= \lim_{t\to\infty} \mathbb{P}\left(R(\hat{f}_n) - R(f^*) > \epsilon_{n,t}\right) \\ &\leq c_1 \exp(-2\log n) \ , \end{aligned}$$

where $c_1 = \bar{c}_1 + 2$.

3.5.2 Proof of Theorem 3.2.2

For simplicity, let us assume that the infimum in

$$\inf\{R(f_\alpha) - R(f^*) + V_n(N(\alpha)) : \ f_\alpha \in \mathcal{F}\} \ ,$$

is attained for some $f_{\alpha^*} \in \mathcal{F}$:

$$f_{\alpha^*} := \arg\min\{R(f_\alpha) - R(f^*) + V_n(N(\alpha)) : \ f_\alpha \in \mathcal{F}\} \ .$$

Define

$$t_n := \frac{K_n \|\hat{f}_n - f_{\alpha^*}\|_{1,\nu}}{2 + K_n \|\hat{f}_n - f_{\alpha^*}\|_{1,\nu}} \ .$$

Let

$$\tilde{f}_n := (1 - t_n)\hat{f}_n + t_n f_{\alpha^*} \ .$$

Then $\tilde{f}_n \in \mathcal{F}$ because $\mathcal{F}$ is convex. Moreover,

$$\|\tilde{f}_n - f_{\alpha^*}\|_\infty = \frac{2\|\hat{f}_n - f_{\alpha^*}\|_\infty}{2 + K_n \|\hat{f}_n - f_{\alpha^*}\|_{1,\nu}} \leq \frac{2\|\hat{f}_n - f_{\alpha^*}\|_\infty}{K_n \|\hat{f}_n - f_{\alpha^*}\|_{1,\nu}} \leq 2 \ .$$

Observe also that by the convexity of the hinge loss and of the ℓ_1 norm, for $\tilde{\alpha}_n = (1 - t_n)\hat{\alpha}_n + t_n \alpha^*$,

$$R_n(\tilde{f}_n) - R_n(f_{\alpha^*}) + \hat{\lambda}_n(I(\tilde{\alpha}_n) - I(\alpha^*))$$

$$\leq (1 - t_n)\left[R_n(\hat{f}_n) - R_n(f_{\alpha^*}) + \hat{\lambda}_n(I(\hat{\alpha}_n) - I(\alpha^*))\right] \leq 0 \ .$$

Let Ω^* be defined as in (3.28), but with $\mathcal{F}$ replaced by $\tilde{\mathcal{F}} := \mathcal{F} \cap \{\|f - f_{\alpha^*}\|_\infty \leq 2\}$. Then, by the same arguments as in the proof of Theorem 3.2.1, with $\hat{f}_n$ now replaced by $\tilde{f}_n$, and with $K = 2$, we see that on Ω^*,

$$R(\tilde{f}_n) - R(f^*) \leq \epsilon_n \ .$$

Next, Condition A implies

$$\|\tilde{f}_n - f^*\|_{1,\nu}^\kappa \leq \sigma^\kappa \Big(R(\tilde{f}_n) - R(f^*) \Big) \ .$$

On the other hand, by the triangle inequality and again Condition A,

$$\begin{aligned} \|\tilde{f}_n - f^*\|_{1,\nu} &\geq \|\tilde{f}_n - f_{\alpha^*}\|_{1,\nu} - \|f_\alpha^* - f^*\|_{1,\nu} \\ &\geq (1 - t_n)\|\hat{f}_n - f_{\alpha^*}\|_{1,\nu} - \sigma\epsilon_n^{1/\kappa} \ , \end{aligned}$$

because $R(f_{\alpha^*}) - R(f^*) \leq (1 + 4\delta)(R(f_{\alpha^*}) - R(f^*)) \leq \epsilon_n$. So

$$R(\tilde{f}_n) - R(f^*) \leq \epsilon_n$$

implies

$$(1 - t_n)\|\hat{f}_n - f_{\alpha^*}\|_{1,\nu} \leq 2\sigma\epsilon_n^{1/\kappa} \ ,$$

or

$$\|\hat{f}_n - f_{\alpha^*}\|_{1,\nu} \leq 2\sigma\epsilon_n^{1/\kappa} + K_n\sigma\epsilon_n^{1/\kappa}\|\hat{f}_n - f_{\alpha^*}\|_{1,\nu} \ ,$$

or, since $2K_n\sigma\epsilon_n^{1/\kappa} \leq 1$,

$$\|\hat{f}_n - f_{\alpha^*}\|_{1,\nu} \leq 4\sigma\epsilon_n^{1/\kappa} \ .$$

But then, using once more that $2K_n\sigma\epsilon_n^{1/\kappa} \leq 1$,

$$\|\hat{f}_n - f_{\alpha^*}\|_\infty \leq 2 \ .$$

In other words, on Ω^*, we have that $\hat{f}_n \in \tilde{\mathcal{F}}$.

But this means that on Ω^*, we can apply the arguments of Theorem 3.2.1, to arrive at

$$R(\hat{f}_n) - R(f_{\alpha^*}) \leq \epsilon_n \ .$$

Since $\mathbf{P}(\Omega^*) \geq 1 - c_1/n^2$, this completes the proof. ■

3.6 Proof of the results in Section 3.4

Proof of Lemma 3.4.1. Consider for some $z > 0$ the set $|2\eta - 1| \leq z$. Since $|2\eta(u,v) - 1| \geq |v - g^*(u)|^\gamma / c_\eta$ we have

$$\{|2\eta - 1| \leq z\} \subset \{(u,v) : |v - g^*(u)| \leq (c_\eta z)^{1/\gamma}\} .$$

It follows that when $(c_\eta z)^{1/\gamma} \leq s$,

$$Q(\{|2\eta - 1| \leq z\}) \leq Q(\{(u,v) : |v - g^*(u)| \leq (c_\eta z)^{1/\gamma}\})$$

$$\leq c_q \mu(\{(u,v) : |v - g^*(u)| \leq (c_\eta z)^{1/\gamma}\}) = 2c_q (c_\eta z)^{1/\gamma} .$$

So Condition AA holds with $C = c_\eta (2c_q)^\gamma \vee 1/s$.

By Remark 3.3.1, since f_g takes values in $\{\pm 1\}$, we have

$$R(f_g) - R(f^*) = 2 \int_{f_g \neq f^*} |2\eta - 1| dQ .$$

Hence,

$$\begin{aligned} R(f_g) - R(f^*) &= 2 \int \int_{g(u) \wedge g^*(u)}^{g(u) \vee g^*(u)} |2\eta(u,v) - 1| q(u,v) dv du \\ &\leq 2c_\eta c_q \int \int_{g(u) \wedge g^*(u)}^{g(u) \vee g^*(u)} |v - g^*(u)|^\gamma du \\ &= \frac{2c_\eta c_q}{\kappa} \|g - g^*\|_{\kappa,\mu}^\kappa \leq 2c_\eta c_q \|g - g^*\|_{\kappa,\mu}^\kappa . \end{aligned}$$

■

Proof of Lemma 3.4.2. Let

$$g_\epsilon^*(u) := \lfloor g^*(u_j)/\delta \rfloor \delta, \quad u_{j-1} < u \leq u_j .$$

Then $|g_\epsilon^*(u) - g^*(u)| \leq \delta + |u_j - u|^{1/r} \leq \delta + \epsilon$. ■

Proof of Lemma 3.4.3. Let $g^* \in \mathcal{G}_r$(Hölder). Consider the integer l such that $2^{l-1} \leq J \leq 2^l$. Take $\epsilon = 2^{-(l-1)/r}$. Then $\lceil \epsilon^{-r} \rceil = \epsilon^{-r} \leq J$. Moreover $\epsilon \leq (J/2)^{-1/r}$. The function g_ϵ^* is piecewise constant on at most J intervals. We note that for the one-dimensional expansion in the Haar basis, of the indicator function of a half-interval $1_{[a,1]}$, with $a \in \Delta$, we need no more than L non-zero

coefficients. So we need at most $L^2 J$ non-zero coefficients to expand f_ϵ^*. Here, f_ϵ^* is the boundary fragment with boundary g_ϵ^*. Hence by Lemma 3.4.2

$$\|g_\epsilon^* - g^*\|_\infty \leq (J/2)^{-1/r} + \delta_n \ .$$

Note also that f_ϵ^* takes only the values ± 1, so $f_\epsilon^* \in \mathcal{F}$.

Finally, if $g^* \in \mathcal{G}_0$, we consider the function $g_{\delta_n}^* := \lfloor g^*/\delta_n \rfloor \delta_n$. We clearly need no more then L coefficients to expand the boundary fragment $f_{\delta_n}^*$ corresponding to $g_{\delta_n}^*$.

The proof is completed by applying Lemma 3.4.1, and the inequality $(a+b)^\kappa \leq 2^\kappa (a^\kappa + b^\kappa)$, $a, b > 0$. ■

Proof of Theorem 3.4.1. By Lemma 3.4.1, Condition A holds with $\kappa = 1+\gamma$ and with $d\nu = \eta(1-\eta)dQ$.

Now, the base functions $\{\psi_k\}$ have $L_2(\mu)$ norm equal to one, and are orthogonal in $L_2(\mu)$. So, by Assumption 1 and Assumption 4, we have that $d\nu = \eta(1-\eta)dQ \geq s^2/c_q d\mu$. Therefore, we know that the smallest eigenvalue ρ_ν^2 of Σ_ν satisfies $\rho_\nu^2 \geq s^2/c_q$. Thus, Condition B is met as well.

Condition C is met, since $2^{2(L-2)} \leq \log n/n$ implies

$$\|\psi_k\|_\infty \leq \sqrt{n/\log n}, \ \forall \ k \ .$$

The result now follows from Theorem 3.2.1. To see this, we invoke Lemma 3.4.3.

When $g^* \in \mathcal{G}_r$(Hölder), we let

$$N := \left\lfloor \left(n/\log^2 n\right)^{\frac{r}{2\kappa+r-1}} \right\rfloor L^2 \ . \tag{3.39}$$

Then

$$J := N/L^2 \leq \left(n/\log^2 n\right)^{\frac{r}{2\kappa+r-1}} \leq (n/\log n)^{\frac{r}{2\kappa+r-1}}$$

$$\leq (n/\log n)^{r/2} = \delta_n^{-r} \ .$$

For the estimation error, we now have

$$V_n(N) = O\left(\frac{N\log n}{n}\right)^{\frac{\kappa}{2\kappa-1}} = O\left(\frac{\log^2 n}{n}\right)^{\frac{\kappa}{2\kappa+r-1}} \ .$$

In view of Lemma 3.4.3,

$$\inf_{N(\alpha)=N} R(f_\alpha) - R(f^*) = O\left(L^2/N\right)^{\frac{\kappa}{r}} = O\left(\frac{\log^2 n}{n}\right)^{\frac{\kappa}{2\kappa+r-1}} .$$

When $g^* \in \mathcal{G}_0$, the result immediately follows from Theorem 3.2.1 by taking $N := L$ and applying Lemma 3.4.3. ■

Chapter 4

A moment bound for multi-hinge classifiers

This chapter is based on a research report (Tarigan and van de Geer, 2007) that has been accepted for publication in the Journal of Machine Learning Research, Dec. 2007. Most material of the first section of the article are in Sections 2.7–2.8 of this thesis.

4.1 Introduction

In Sections 2.7–2.8 we have discussed the statistical property of some large margin based losses as proxy losses for the true multicategory 0–1 loss. It has been shown that the multi-hinge loss (2.16) is Bayes consistent and that the excess multi-hinge risk minimization leads to the minimization of the excess prediction error.

In this chapter we first identify the margin condition wrt. multi-hinge loss (2.16), and then we show a moment bound for the excess multi-hinge risk based on two kinds of complexity constraints: entropy with bracketing and empirical entropy. Obtaining such a result based on the latter is harder than finding one based on the former. We obtain fast rates of convergence that adapt to the unknown margin and complexity parameters.

In the multicategory setting, we recall that a classifier is defined as a vector mapping $f = (f_1, \ldots, f_m)$ in $\mathbb{R}^m$ with the zero-sum constraint $\sum f_j = 0$ and that a predictor g induced by f is defined through the relation $g(f) = \arg\max f_j$. We also recall that the smallest risk R^* is the risk of the Bayes classifier f^* as in (2.18) having value 1 in the jth component and $-1/(m-1)$ in the remaining components, when $j = \arg\max_{j=1,\ldots,m} p_j$, with $p_j(x) = \mathbb{P}(Y = j | X = x)$, $x \in \mathcal{X}$. Let $\mathcal{F}$ be a model class of candidate classifiers. For $j = 1, \ldots, m$, we assume that each f_j is a member of the same class $\mathcal{F}_o = \{h : \mathbb{R}^d \to \mathbb{R}, h \in L_2(Q)\}$, with Q the unknown marginal distribution of X. That is,

$$\mathcal{F} = \{f = (f_1, \ldots, f_m) : \ \sum_{j=1}^{m} f_j = 0, \ f_j \in \mathcal{F}_o\} \ . \tag{4.1}$$

Let P_n be the empirical distribution of (X, Y) based on the observations $\{(X_i, Y_i)\}_{i=1}^n$ and Q_n the corresponding empirical distribution of X based on $X_1, \ldots, X_n$. We endow $\mathcal{F}$ with the following squared semi-metrics

$$\begin{aligned} \|f - \tilde{f}\|_{2,Q}^2 &:= \sum_{j=1}^{m} \int |f_j - \tilde{f}_j|^2 \, dQ \ , \quad \text{and} \\ \|f - \tilde{f}\|_{2,Q_n}^2 &:= \sum_{j=1}^{m} \frac{1}{n} \sum_{i=1}^{n} |f_j(X_i) - \tilde{f}_j(X_i)|^2 \ , \end{aligned}$$

for all $f, \tilde{f} \in \mathcal{F}$. Let $H_B(\epsilon, \mathcal{F}_o, L_2(Q))$ and $H(\epsilon, \mathcal{F}_o, L_2(Q_n))$ denote the ϵ-entropy with bracketing and the empirical ϵ-entropy of the class $\mathcal{F}_o$, respectively (see Appendix for the definitions of the entropies). The complexity of a model class can be summarized in a complexity parameter $\rho \in (0, 1)$. Let A be some positive constant. We consider classes $\mathcal{F}_o$ satisfying one of the following complexity constraints:

$$\begin{aligned} H_B(\epsilon, \mathcal{F}_o, L_2(Q)) &\leq A\epsilon^{-2\rho} \ , \text{ for all } \epsilon > 0 \ , \quad \text{or} \\ H(\epsilon, \mathcal{F}_o, L_2(Q_n)) &\leq A\epsilon^{-2\rho} \ , \text{ for all } \epsilon > 0 \ , \text{ a.s. for all } n \geq 1 \ . \end{aligned}$$

It is straightforward to show that for all $\epsilon > 0$:

$$\begin{aligned} H_B(\epsilon, \mathcal{F}, \|\cdot\|_{2,Q}) &\leq (m-1) \, H_B(\epsilon(m-1)^{-1/2}, \ \mathcal{F}_o, \ L_2(Q)) \ , \\ H(\epsilon, \mathcal{F}, \|\cdot\|_{2,Q_n}) &\leq (m-1) \, H(\epsilon(2m-2)^{-1/2}, \ \mathcal{F}_o, \ L_2(Q_n)) \ . \end{aligned}$$

We define the minimizer of the empirical multi-hinge loss

$$\hat{f}_n := \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1, j \neq Y_i}^{m} (f_j(X_i) + \frac{1}{m-1})_+ \ , \tag{4.2}$$

where the model class $\mathcal{F}$ as defined in (4.1) satisfies either a constraint on entropy with bracketing or an empirical entropy constraint as described above.

Besides the model class complexity, the rate of convergence also depends on the so-called margin condition (see Condition A below) that quantifies the identifiability of the Bayes rule and is summarized in a margin parameter (or noise level) $\kappa \geq 1$. In Tarigan and van de Geer (2006), a probability inequality has been obtained for l_1-penalized excess hinge risk in the binary case that adapts to the unknown parameters. In this chapter, we show a moment inequality for the excess multi-hinge risk of $\hat{f}_n$ over the model class $\mathcal{F}$ with rate of convergence $n^{-\kappa/(2\kappa-1+\rho)}$, which is faster than $n^{-1/2}$.

In Section 4.2 we present our main result based on the margin and complexity conditions. We give the proof of the main result in Section 4.3, together with our supporting lemmas.

4.2 A moment bound

We first state the margin and the complexity conditions.

Condition A *(Margin condition).* There exist constants $\sigma > 0$ and $\kappa \geq 1$ such that for all $f \in \mathcal{F}$,

$$R(f) - R^* \ \geq \ \frac{1}{\sigma^\kappa} \Big(\sum_{j=1}^{m} \int |f_j - f_j^*| \ dQ \Big)^\kappa \ .$$

Condition B1 *(Complexity constraint under ϵ-entropy with bracketing).* Let $0 < \rho < 1$ and let A be a positive constant. The ϵ-entropy with bracketing satisfies the inequality

$$H_B(\epsilon, \mathcal{F}_o, L_2(Q)) \leq A\epsilon^{-2\rho} \ , \quad \text{for all } \epsilon > 0 \ .$$

Condition B2 *(Complexity constraint under empirical ϵ-entropy).* Let $0 < \rho < 1$ and let A be a positive constant. The empirical ϵ-entropy, almost surely

for all $n \geq 1$, satisfies the inequality

$$H(\epsilon, \mathcal{F}_o, L_2(Q_n)) \leq A\epsilon^{-2\rho} \ , \quad \text{for all } \epsilon > 0 \ .$$

Now we come to the main result.

Theorem 4.2.1. *Assume Condition A is met and that $|f_j - f_j^*| \leq M$ for all $j = 1, \ldots, m$, and all $f = (f_1, \ldots, f_m) \in \mathcal{F}$. Let $\hat{f}_n$ be the multi-hinge loss minimizer defined in (4.2). Suppose that either Condition B1 or Condition B2 holds. Then for small values of $\delta > 0$,*

$$\mathbb{E}[R(\hat{f}_n) - R^*] \leq \frac{1+\delta}{1-\delta} \ \inf\left\{ R(f) - R^* + C_0 \ n^{-\frac{\kappa}{2\kappa-1+\rho}} \ : \ f \in \mathcal{F} \right\}$$

with C_0 some constant depending only on m, M, κ, σ, A and ρ.

Condition A follows from the condition on the behaviour of the conditional probabilities p_j. We formulate this in Condition AA below. We require that, for a fixed $x \in \mathbb{R}^d$, there is no pair of categories having the same conditional probabilities each of which stays away from 1. Originally the terminology "margin condition" comes from the binary case of the prediction error considered in the work of Mammen and Tsybakov (1999) and Tsybakov (2004), where the behaviour of p_1, the conditional probability of category 1, is restricted near $\{x : p_1(x) = 1/2\}$. The "margin" set $\{x : p_1(x) = 1/2\}$ identifies the Bayes predictor which assigns a new x to class 1 if $p_1(x) > 1/2$ and class 2 otherwise. The margin condition is also called the *condition on the noise level*, and it is summarized in a margin parameter κ. Boucheron, Bousquet, and Lugosi (2005, Section 5.2) discuss the noise condition and its equivalent variants, corresponding to the fast rates of convergence, in the binary case. Thus, Condition AA is a natural extension for the multicategory case wrt. hinge loss. Lemma 4.2.1 below gives the connection between Condition A and Condition AA. Define, for $x \in \mathcal{X}$,

$$\tau(x) := \min_{j \neq k} \{ |p_j(x) - p_k(x)|, 1 - p_j(x) \} \ , \tag{4.3}$$

where j and k take values in $\{1, 2, \ldots, m\}$.

Condition AA. Let τ as defined in (4.3). There exist constants $C \geq 1$ and $\gamma \geq 0$ such that $\forall z > 0$,

$$Q(\{\tau \leq z\}) \leq (Cz)^{1/\gamma} \ .$$

[Here we use the convention $(Cz)^{1/\gamma} = \mathbb{1}\{z \geq 1/C\}$ for $\gamma = 0$.]

Lemma 4.2.1. *Suppose Condition AA is met. Then for all $f \in \mathcal{F}$ with $|f_j - f_j^*| \leq M$ for all $j = 1, \ldots, m$,*

$$R(f) - R^* \geq \frac{1}{\sigma_M} \Big(\sum_{j=1}^{m} \int |f_j - f_j^*| \, dQ \Big)^{1+\gamma} ,$$

where $\sigma_M = C \left(mM(1/\gamma + 1) \right)^{\gamma} (1 + \gamma)$. That is, Condition A holds with $\sigma = (\sigma_M)^{1/\kappa}$ and $\kappa = 1 + \gamma$.

Proof. Let τ be defined as in (4.3). We write $L(f) = \mathbb{E}_{Y|X}[l(Y, f(X))|X = x]$ and recall that $p_j(x) = P(Y = j|X = x)$, for all $j = 1, \ldots, m$. We fix an arbitrary $x \in \mathbb{R}^d$. From the proof of Lemma 2.8.1, clearly

$$\begin{aligned} (L(f) - L(f^*)) \, \mathbb{1}(p_k = \max_{j=1,\ldots,m} p_j) &\geq \tau \sum_{j \neq k} |f_j - f_j^*| \\ &\geq \frac{\tau}{2} \sum_{j=1}^{m} |f_j - f_j^*| , \end{aligned}$$

where the second inequality is obtained from the fact that $|f_k - f_k^*| \leq \sum_{j \neq k} |f_j - f_j^*|$. That is, the excess risk is lower bounded by

$$\frac{1}{2} \sum_{j=1}^{m} \int \tau |f_j - f_j^*| dQ \, .$$

It implies that, for all $z > 0$,

$$R(f) - R^* \geq \frac{z}{2} \sum_{j=1}^{m} \Big[\int |f_j - f_j^*| dQ - \int_{\tau \leq z} |f_j - f_j^*| dQ \Big] .$$

Since $|f_j - f_j^*| \leq M$ for all j, and by Condition AA, the second integral in the inequality above can be upper bounded by $M(Cz)^{1/\gamma}$. Thus, for all $z > 0$,

$$R(f) - R^* \geq \frac{z}{2} \sum_{j=1}^{m} \int |f_j - f_j^*| dQ \; - \; \frac{z}{2} mM(Cz)^{1/\gamma} \, .$$

We take $z = \Big(\sum_{j=1}^{m} \int |f_j - f_j^*| dQ \Big)^{\gamma} / \Big(mMC^{1/\gamma}(1 + \gamma^{-1}) \Big)^{\gamma}$ when $\gamma > 0$, and $z \uparrow 1/C$ when $\gamma = 0$. ∎

The complexity constraints B1 and B2 cover some interesting classes, including Vapnik-Chervonenkis (VC) subgraph classes and VC convex hull classes. See, for example, van der Vaart and Wellner (1996, Section 2.7), van de Geer (2000, Sections 2.4, 3.7, 7.4, 10.1 and 10.3) and Song and Wellner (2002). In the situation when the approximation error $\inf_{f\in\mathcal{F}} R(f) - R^*$ is zero (the model class $\mathcal{F}$ contains the Bayes classifier), Steinwart and Scovel (2005a) obtain the same rate of convergence for the excess hinge risk under the margin condition A and the complexity condition B2. They consider the RKHS-regularization setting (2.15) for the binary case instead. We do not explore the behaviour of the approximation error $\inf_{f\in\mathcal{F}} R(f) - R^*$. This problem is still open and very hard to solve even in the binary case.

4.3 Proof of Theorem 4.2.1

As shorthand notation we write for the loss $l_f(X,Y) := l(Y, f(X))$. Let $f^o := \arg\min_{f\in\mathcal{F}} R(f)$, the minimizer of the theoretical risk in the model class $\mathcal{F}$. We also write $\nu_n(l_f) := \sqrt{n}\ (R_n(f) - R(f))$. Since $R_n(\hat{f}_n) - R_n(f) \le 0$ for all $f \in \mathcal{F}$, we have the following basic inequality (see (1.11)):

$$R(\hat{f}_n) - R^* \le |\nu_n(l_{\hat{f}_n}) - \nu_n(l_{f^o})|/\sqrt{n} + R(f^o) - R^* \ .$$

This upper bound enables us to work with the increments of the empirical process $\{\nu_n(l_f) - \nu_n(l_{f^o}) \ : \ l_f \in \mathcal{L}\}$ indexed by the multi-hinge loss $l_f \in \mathcal{L}$, where $\mathcal{L} = \{l_f : f \in \mathcal{F}\}$.

The procedure of the proof is based on the proof of Lemma 2.1 in del Barrio, Deheuvels, and van de Geer (2007), page 206. We write

$$Z_n(l_f) := \frac{|\nu_n(l_f) - \nu_n(l_{f^o})|}{\left(\|l_f - l_{f^o}\|_{2,P} \vee n^{-\frac{1}{2+2\rho}}\right)^{1-\rho}} \ , \ l_f \in \mathcal{L} \ ,$$

where $(a \vee b) := \max\{a, b\}$, $\|l_f\|_{2,P}^2 := \int l_f^2(x,y)\ dP(x,y)$ and ρ is from either Condition B1 or B2. For short hand of notation, we also write $Z_n = Z_n(l_{\hat{f}_n})$. Then

$$\begin{aligned} & R(\hat{f}_n) - R^* \\ & \le (Z_n/\sqrt{n})\ \left(\ \|l_{\hat{f}_n} - l_{f^o}\|_{2,P}^{1-\rho} \vee n^{-\frac{1-\rho}{2+2\rho}}\right)\ +\ R(f^o) - R^* \ . \end{aligned} \tag{4.4}$$

Applying the triangular inequality and Lemma 4.3.1 below gives

$$\|l_{\hat{f}_n} - l_{f^o}\|_{2,P}^{1-\rho} \le (m-1)^{(1-\rho)/2}\ \left(\|\hat{f}_n - f^*\|_{2,Q}^{1-\rho} + \|f^o - f^*\|_{2,Q}^{1-\rho}\right) \ .$$

Observe that for any $f \in \mathcal{F}$ with $|f_j - f_j^*| \leq M$, and for all j, Condition A gives $\|f - f^*\|_{2,Q}^2 \leq M\sigma\ (R(f) - R^*)^{1/\kappa}$. Thus,

$$\begin{aligned}&\|l_{\hat{f}_n} - l_{f^o}\|_{2,P}^{1-\rho}\\ &\leq C_1 \left\{[R(\hat{f}_n) - R^*]^{(1-\rho)/2\kappa} + [R(f^o) - R^*]^{(1-\rho)/2\kappa}\right\},\end{aligned}$$

with $C_1 = ((m-1)M\sigma)^{(1-\rho)/2}$. Denote by $\mathcal{R}$ the right hand side of the above inequality. Hence, from (4.4) we have

$$R(\hat{f}_n) - R^* \leq (Z_n/\sqrt{n})\ (\mathcal{R} \vee n^{-\frac{1-\rho}{2+2\rho}}) + R(f^o) - R^*\ .$$

We consider first the case $(\mathcal{R} \vee n^{-\frac{1-\rho}{2+2\rho}}) = \mathcal{R}$. That is,

$$\begin{aligned}R(\hat{f}_n) - R^* \ \leq\ & \frac{Z_n}{\sqrt{n}}\ C_1 \left\{[R(\hat{f}_n) - R^*]^{(1-\rho)/2\kappa}\ +\right.\\ &\left.+[R(f^o) - R(f^*)]^{(1-\rho)/2\kappa}\right\} +\ R(f^o) - R^*\ .\end{aligned}$$

Two applications of Lemma 4.3.2 below yield for all $0 < \delta < 1$,

$$\begin{aligned}R(\hat{f}_n) - R^* \ \leq\ & \delta(R(\hat{f}_n) - R^*) + (1+\delta)(R(f^o) - R^*) +\\ &+2\big(C_1\ Z_n/\sqrt{n}\big)^{\frac{2\kappa}{2\kappa-1+\rho}}\ \delta^{-\frac{1-\rho}{2\kappa-1+\rho}}\\ \leq\ & \delta(R(\hat{f}_n) - R^*)\ +\\ &+(1+\delta)\Big(R(f^o) - R^* + C_2\ Z_n^r\ n^{-\frac{\kappa}{2\kappa-1+\rho}}\Big)\ ,\end{aligned}$$

with $C_2 = 2\ C_1^r\ \delta^{-\frac{1-\rho}{2\kappa-1+\rho}}$ and $r = 2\kappa/(2\kappa - 1 + \rho)$. Now it is left to show that $\mathbb{E}[Z_n^r]$ is bounded, say by some constant C_3. Then, $C_0 = C_2C_3$ in Theorem 4.2.1.

To show that $\mathbb{E}[Z_n^r]$ is bounded, we use an exponential tail probability of the supremum of the weighted empirical process

$$\{Z_n(l_f)\ :\ l_f \in \mathcal{L}\}\ . \tag{4.5}$$

We recall that $H_B(\epsilon, \mathcal{F}, \|\cdot\|_{2,Q}) \leq (m-1)H_B(\epsilon(m-1)^{-1/2}, \mathcal{F}_o, L_2(Q))$. A key observation is that

$$H_B(\epsilon, \mathcal{L}, L_2(P)) \leq (m-1)\ H_B(\epsilon(m-1)^{-1/2}, \mathcal{F}, \|\cdot\|_{2,Q})\ ,$$

by Lemma 4.3.1. It gives an upper bound for the ϵ-entropy with bracketing of the model class $\mathcal{L}$: $H_B(\epsilon, \mathcal{L}, L_2(P)) \leq A_o \epsilon^{-2\rho}$, for all $\epsilon > 0$, with $A_o = A(m-1)^{2+2\rho}$. Under Condition B1, an application of Lemma 5.13 in van de Geer (2000), presented below in Lemma 4.3.3, gives the desired exponential tail probability. Hence, for some positive constant c,

$$\begin{aligned}\mathbb{E}[Z_n^r] &= \int_0^c \mathbb{P}(Z_n \geq t^{1/r})\, dt \;+\; \int_c^\infty \mathbb{P}(Z_n \geq t^{1/r})\, dt \\ &\leq c \;+\; \int_0^\infty c\, \exp(-\frac{t^{1/r}}{c^2})\, dt \;= c + rc^{2r+1}\Gamma(r)\ .\end{aligned}$$

For the case $\mathcal{R} \leq n^{-(1-\rho)/(2+2\rho)}$, we have

$$R(\hat{f}_n) - R^* \leq Z_n\ n^{-1/(1+\rho)} \;+\; R(f^o) - R^*\ .$$

We conclude by noting that $n^{-1/(1+\rho)} \leq n^{-\kappa/(2\kappa-1+\rho)}$, where $\kappa \geq 1$ and $0 < \rho < 1$.

Now we consider the case where Condition B2 holds instead of B1. By virtue of the proof above, we need only to verify an exponential probability of the supremum of the process (4.5) under Condition B2 instead of B1. This is done by employing Lemmas 4.3.4–4.3.7 below. Again, a key observation is that Lemma 4.3.1 and Condition B2 give us $H(\epsilon, \mathcal{L}, L_2(P_n)) \leq A(m-1)^{2+2\rho}\epsilon^{-2\rho}$.■

Lemma 4.3.1 gives an upper bound of the squared $L_2(P)$-metric of the excess loss in terms of $\|\cdot\|_{2,Q}$-metric.

Lemma 4.3.1. $\mathbb{E}[(l_f(X,Y) - l_{f^*}(X,Y))^2] \leq (m-1)\sum_{j=1}^m \int |f_j - f_j^*|^2\ dQ$.

Proof. We write $\Delta(f, f^*) = \mathbb{E}_{Y|X}[(l_f(X,Y) - l_{f^*}(X,Y))^2 | X = x]$ and recall that $p_j(x) = P(Y = j|X = x)$, for all $j = 1, \ldots, m$. We fix an arbitrary $x \in \mathbb{R}^d$. Definition of the loss gives

$$\begin{aligned}\Delta(f, f^*) &= \sum_{j=1}^m p_j \Big(\sum_{i \neq j} (f_i + \frac{1}{m-1})_+ - (f_i^* + \frac{1}{m-1})_+ \Big)^2 \\ &= \sum_{j=1}^m p_j \Big(\sum_{i \in I^+(j)} (f_i - f_i^*) + \sum_{i \in I^-(j)} (-\frac{1}{m-1} - f_i^*) \Big)^2\ ,\end{aligned}$$

where $I^+(j) = \{i \neq j\ :\ f_i \geq -1/(m-1),\ i = 1, \ldots, m\}$ and $I^-(j) = \{i \neq j\ :\ f_i < -1/(m-1),\ i = 1, \ldots, m\}$. Use the facts that $(\sum_{i=1}^n a_i)^2 \leq n \sum_{i=1}^n a_i^2$

for all $n \in \mathbb{N}$ and $a_i \in \mathbb{R}$, and that $\max\{|I^+(j)|, |I^-(j)|\} \leq m-1$, to obtain

$$\Delta(f, f^*)$$

$$\leq (m-1)\sum_{j=1}^{m} p_j \Big(\sum_{i \in I^+(j)} (f_i - f_i^*)^2 \; + \; \sum_{i \in I^-(j)} (-\frac{1}{m-1} - f_i^*)^2 \Big) \; .$$

Clearly, $|-1/(m-1) - f_i^*| \leq |f_i - f_i^*|$ for all $i \in I^-(j)$. Hence,

$$\Delta(f, f^*) \leq (m-1)\sum_{j=1}^{m} p_j \; (\sum_{i \neq j} |f_i - f_i^*|^2)$$

$$= (m-1)\sum_{j=1}^{m} (1-p_j)|f_j - f_j^*|^2 \; ,$$

where the last equality is obtained using $\sum_{j=1}^{m} p_j = 1$. We conclude the proof by bounding $1 - p_j$ with 1 for all j and integrating over all $x \in \mathbb{R}^d$ wrt. the marginal distribution Q. ■

The technical lemma below is an immediate consequence of Young's inequality (see, for example, Hardy et al., 1988, Sec. 8.3), using some straightforward bounds to simplify the expressions.

Lemma 4.3.2 (Technical Lemma)**.** *For all positive ν, t, δ and $\kappa > \beta$:*

$$\nu t^{\beta/\kappa} \; \leq \; \delta t \; + \; \nu^{\frac{\kappa}{\kappa-\beta}} \; \delta^{\frac{-\beta}{\kappa-\beta}} \; .$$

To ease the exposition, throughout Lemma 4.3.3 and Lemma 4.3.4 we write

$$\|\cdot\| = \|\cdot\|_{2,Q} \text{ and } \|\cdot\|_n = \|\cdot\|_{2,Q_n}$$

for the $L_2(Q)$-norm and the $L_2(Q_n)$-norm, respectively.

Lemma 4.3.3. *(Lemma 5.13 van de Geer, 2000) For a probability measure Q, let $\mathcal{H}$ be a class of uniformly bounded functions h in $L_2(Q)$, say $\sup_{h \in \mathcal{H}} |h - h^o|_\infty < 1$, where h^o is a fixed but arbitrary function in $\mathcal{H}$. Suppose that*

$$H_B(\epsilon, \mathcal{H}, L_2(Q)) \leq A_o \epsilon^{-2\rho} \; , \; \textit{for all } \epsilon > 0 \; ,$$

with $0 < \rho < 1$ and $A_o > 0$. Then for some positive constants c and n_o depending only on ρ and A_o,

$$\mathbb{P}\Big(\sup_{h\in\mathcal{H}} \frac{|\nu_n(h) - \nu_n(h^o)|}{\Big(\|h-h^o\| \vee n^{-\frac{1}{2+2\rho}}\Big)^{1-\rho}} \geq t \Big) \leq c\ \exp(-t/c^2)\ ,$$

for all $t > c$ and $n > n_o$.

Lemma 4.3.4. *For a probability measure Q on $(\mathcal{Z}, \mathcal{A})$, let $\mathcal{H}$ be a class of uniformly bounded functions h in $L_2(Q)$, say $\sup_{h\in\mathcal{H}} |h - h^o|_\infty < 1$, where h^o is a fixed but arbitrary element in $\mathcal{H}$. Suppose that*

$$H(\epsilon, \mathcal{H}, L_2(Q_n)) \leq A_o \epsilon^{-2\rho}\ ,\ \textit{for all } \epsilon > 0\ ,$$

with $0 < \rho < 1$ and $A_o > 0$. Then for some positive constants c and n_o depending on ρ and A_o,

$$\mathbb{P}\Big(\sup_{h\in\mathcal{H}} \frac{|\nu_n(h) - \nu_n(h^o)|}{\Big(\|h-h^o\| \vee n^{-\frac{1}{2+2\rho}}\Big)^{1-\rho}} \geq t \Big) \leq c\ \exp(-t/c^2)\ ,$$

for all $t > c$ and $n > n_o$.

Proof. For $n \geq (t^2/8)^{1+\rho/(1-\rho)}$, Chebyshev's inequality and a symmetrization technique (see, for example, van de Geer, 2000, page 32) give

$$\begin{aligned}
&\mathbb{P}\left(\sup_{h\in\mathcal{H}} \frac{|\nu_n(h) - \nu_n(h^o)|}{\Big(\|h-h^o\| \vee n^{-1/(2+2\rho)}\Big)^{1-\rho}} \geq t \right) \\
&\leq\ 4\mathbb{P}\left(\sup_{h\in\mathcal{H}} \frac{|\nu_n^\varepsilon(h) - \nu_n^\varepsilon(h^o)|}{\Big(\|h-h^o\|_n \vee n^{-1/(2+2\rho)}\Big)^{1-\rho}} \geq \sqrt{t}/4 \right) \qquad (4.6) \\
&\quad +\ 4\mathbb{P}\left(\sup_{h\in\mathcal{H}} \frac{\|h-h^o\|_n^{1-\rho}}{\Big(\|h-h^o\| \vee n^{-1/(2+2\rho)}\Big)^{1-\rho}} \geq \sqrt{t}/4 \right)\ , \qquad (4.7)
\end{aligned}$$

where $\nu_n^\varepsilon(h)$ is the symmetrized version of the $\nu_n(h)$. That is, $\nu_n^\varepsilon(h) = (1/\sqrt{n}) \sum_{i=1}^n \varepsilon_i h(Z_i)$, where $\{\varepsilon_i\}_{i=1}^n$ are independent random variables, independent of $\{Z_i\}_{i=1}^n$, with $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 1/2$ for all $i = 1, \ldots, n$.

To handle (4.6), we divide the class $\mathcal{H}$ into two disjoint classes where the empirical distance $\|h - h^o\|_n$ is smaller or larger than $n^{-1/(2+2\rho)}$. Write $\mathcal{H}_n =$

$\{h \in \mathcal{H} : \ \|h - h^o\|_n \le n^{-1/(2+2\rho)}\}$. By Lemma 5.1 in van de Geer (2000), stated below in Lemma 4.3.5, for some positive constant c_1,

$$\mathbb{P}\Big(\sup_{h\in\mathcal{H}_n} \frac{|\nu_n^\varepsilon(h) - \nu_n^\varepsilon(h^o)|}{n^{-(1-\rho)/(2+2\rho)}} \ge \sqrt{t}/4\Big) \le c_1 \exp\Big(- \frac{t\ n^{1/(1+\rho)}}{64\ c_1^2}\Big) \ .$$

Let $J = \min\{j > 1 : 2^{-j} < n^{-1/(2+2\rho)}\}$. We apply the peeling device in the set $\{h \in \mathcal{H} : 2^{-j} \le \|h - h^o\|_n \le 2^{-j+1},\ j = 1, \ldots, J\}$ to obtain that, for all $t > 1$,

$$\mathbb{P}\Big(\sup_{h\in\mathcal{H}_n^c} \frac{|\nu_n^\varepsilon(h) - \nu_n^\varepsilon(h^o)|}{\|h - h^o\|_n^{1-\rho}} \ge \sqrt{t}/4 \ \Big|\ Z_1, \ldots, Z_n\Big)$$

$$\le \sum_{j=1}^{J} \mathbb{P}\Big(\sup_{\substack{h\in\mathcal{H} \\ \|h-h^o\|_n \le 2^{-j+1}}} |\nu_n^\varepsilon(h) - \nu_n^\varepsilon(h^o)| \ge \frac{\sqrt{t}}{4}\ 2^{-j(1-\rho)} \ | \ Z_1, \ldots, Z_n\Big)$$

$$\le \sum_{j=1}^{J} c_2 \exp(-\frac{t\ 2^{2\rho j}}{216\ c_2^2}) \ \le c\ \exp(-t/c^2) \ .$$

To handle (4.7), we use a modification of Lemma 5.6 in van de Geer (2000), stated below in Lemma 4.3.6, where we take t such that $(\sqrt{t}/4)^{1/(1-\rho)} \ge 14u$. ■

Lemma 4.3.5. *(van de Geer, 2000, Lemma 5.1) Let $Z_1, \ldots, Z_n, \ldots$ be i.i.d. with distribution Q on $(\mathcal{Z}, \mathcal{A})$. Let $\{\varepsilon_i\}_{i=1}^n$ be independent random variables, independent of $\{Z_i\}_{i=1}^n$, with $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 1/2$ for all $i = 1, \ldots, n$. Let $\mathcal{H} \subset L_2(Q)$ be a class of functions on $\mathcal{Z}$. Write $\nu_n^\varepsilon(h) := \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i h(Z_i)$, with $h \in \mathcal{H}$. Let*

$$\mathcal{H}(\delta) := \{h \in \mathcal{H} : \|h - h^o\|_{2,Q} \le \delta\} \ ,$$

$$\hat{\delta}_n := \sup_{h\in\mathcal{H}(\delta)} \|h - h^o\|_{2,Q_n} \ ,$$

where h^o is a fixed but arbitrary function in $\mathcal{H}$ and Q_n is the corresponding empirical distribution of Z based on $\{Z_i\}_{i=1}^n$. For

$$a \ge 8C\Big(\int_{a/(32\sqrt{n})}^{\hat{\delta}_n} H^{1/2}(u, \mathcal{H}, Q_n)du \ \vee \ \hat{\delta}_n\Big) \ ,$$

where C is some positive constant, we have

$$\mathbb{P}\Big(\sup_{h\in\mathcal{H}(\delta)} |\nu_n^\varepsilon(h) - \nu_n^\varepsilon(h^o)| \ge \frac{a}{4}\Big| Z_1, \ldots, Z_n\Big) \le C \exp(-\frac{a^2}{64C^2\hat{\delta}_n^2}) \ .$$

The following lemma is a modification of Lemma 5.6 in van de Geer (2000).

Lemma 4.3.6. *For a probability measure S on $(\mathcal{Z}, \mathcal{A})$, let $\mathcal{H}$ be a class of uniformly bounded functions independent of n with $\sup_{h\in\mathcal{H}} |h|_\infty \leq 1$. Suppose that almost surely for all $n \geq 1$,*

$$H(\epsilon, \mathcal{H}, L_2(S_n)) \leq A_o \epsilon^{-2\rho}, \text{ for all } \epsilon > 0 \ ,$$

with $0 < \rho < 1$ and $A_o > 0$. Then, for all n,

$$\mathbb{P}\Big(\sup_{h\in\mathcal{H}} \frac{\|h\|_{2,S_n}}{\|h\|_{2,S} \vee n^{-\frac{1}{2+2\rho}}} \geq 14u \Big) \leq 4 \ \exp(-u^2 \ n^{\frac{\rho}{1+\rho}}) \ ,$$

for all $u \geq 1$.

Proof. Let $\{\delta_n\}$ be a sequence with $\delta_n \to 0$, $n\delta_n^2 \to \infty$, $n\delta_n^2 \geq 2A_o H(\delta_n)$ for all n with $H(\delta_n) = \delta_n^{-2\rho}$. We apply the randomization device in Pollard (1984, page 32), as follows. Let $Z_{n+1}, \ldots, Z_{2n}$ be an independent copy of $Z_1, \ldots, Z_n$. Let $\omega_1, \ldots, \omega_n$ be independent random variables, independent of $Z_1, \ldots, Z_{2n}$, with $\mathbb{P}(\omega_i = 1) = \mathbb{P}(\omega_i = 0) = 1/2$ for all $i = 1, \ldots, n$. Set $Z_i{}' = Z_{2i-1+\omega_i}$ and $Z_i{}'' = Z_{2i-\omega_i}$, $i = 1, \ldots, n$, and $S_n{}' = (1/n)\sum_{i=1}^n \delta_{Z_i{}'}$, $S_n{}'' = (1/n)\sum_{i=1}^n \delta_{Z_i{}''}$, and $\bar{S}_{2n} = (S_n{}' + S_n{}'')/2$. Since the class is uniformly bounded by 1, an application of Chebyshev's inequality gives that for each h in $\mathcal{H}$,

$$\mathbb{P}\Big(\frac{\|h\|_{2,S_n}}{\|h\|_{2,S} \vee \delta_n} \leq 2u \Big) \geq 1 - \frac{1}{4u^2} \geq 3/4,$$

for all $u \geq 1$. Use a symmetrization lemma of Pollard (1984, Lemma II.3.8), stated in Lemma 4.3.7 below, to obtain

$$\mathbb{P}\Big(\sup_{h\in\mathcal{H}} \frac{\|h\|_{2,S_n}}{\|h\|_{2,S} \vee \delta_n} \geq 14u \Big) \leq 2\mathbb{P}\Big(\sup_{h\in\mathcal{H}} \frac{|\|h\|_{2,S_n{}'} - \|h\|_{2,S_n{}''}|}{\|h\|_{2,S} \vee \delta_n} \geq 12u \Big) \ .$$

The peeling device on the set

$$\{h \in \mathcal{H} : (2u)^{j-1}\delta_n \leq \|h\|_{2,S} \leq (2u)^j \delta_n \ , j = 1, 2, \ldots\}$$

and the inequality in Pollard (1984, page 33) give

$$\mathbb{P}\Big(\sup_{h\in\mathcal{H}} \frac{|\|h\|_{2,S_n{}'} - \|h\|_{2,S_n{}''}|}{\|h\|_{2,S} \vee \delta_n} \geq 12u \ \Big| \ Z_1, \ldots, Z_n \Big)$$

$$\leq \sum_{j=1}^{\infty} \mathbb{P}\Big(\sup_{\substack{h\in\mathcal{H} \\ \|h\|_{2,S} \leq (2u)^j \delta_n}} |\|h\|_{S_n{}'} - \|h\|_{S_n{}''}| \geq 6(2u)^j \delta_n \mid Z_1, \ldots, Z_n \Big)$$

$$\leq \sum_{j=1}^{\infty} 2\exp\Big(H(\sqrt{2}(2u)^j\delta_n, \mathcal{H}, \bar{S}_{2n}) - 2n(2u)^{2j}\delta_n^2\Big)$$

$$\leq \sum_{j=1}^{\infty} 2\exp\Big(H((2u)^j\delta_n, \mathcal{H}, S_n{}') + H((2u)^j\delta_n, \mathcal{H}, S_n{}'') - 2n(2u)^{2j}\delta_n^2\Big)$$

$$\leq \sum_{j=1}^{\infty} 2\exp\Big(-n(2u)^{2j}\delta_n^2\Big)\ , \tag{4.8}$$

where the last inequality is obtained using that since $n\delta_n^2 \geq 2A_oH(\delta_n)$, also $nt^2 \geq 2A_oH(t)$ for all $t \geq \delta_n$ (here $t = (2u)^j\delta_n$). Observe that, since $(2u)^{2j} \geq (2u)^2 j > u^2 j$ for all $u \geq 1$ and $j \geq 1$, we have

$$\sum_{j=1}^{\infty} \exp(-n(2u)^{2j}\delta_n^2) \leq 2\exp(-u^2 n\delta_n^2)\ , \tag{4.9}$$

whenever $n\delta_n^2 > \log 2$. We finish the proof by combining (4.8) and (4.9), and taking $\delta_n = n^{-\frac{1}{2+2\rho}}$. ■

Lemma 4.3.7. *(Pollard, 1984, Lemma II.3.8) Let $\{Z(t) : t \in T\}$ and $\{Z'(t) : t \in T\}$ be independent stochastic process sharing an index set T. Suppose there exist constants $\beta > 0$ and $\alpha > 0$ such that $\mathbb{P}(|Z(t)| \leq \alpha) \geq \beta$ for every $t \in T$. Then*

$$\mathbb{P}\Big(\sup_t |Z(t)| > \epsilon\Big) \leq \beta^{-1}\mathbb{P}\Big(\sup_t\ |Z(t) - Z'(t)| > \epsilon - \alpha\Big)\ .$$

Chapter 5

Multicategory Reject Option

5.1 Introduction

As mentioned in Chapter 1, the possibility to take no decision in predicting the category of an observation can improve the performance of a prediction method whenever the cost of rejecting is smaller than the cost of misclassifying.

Recent developments in the study of the asymptotic statistical property of the ERM-based classification with reject option focus on the binary case only. Herbei and Wegkamp (2006) study the rate of convergence of the excess risk with respect to the true binary reject-loss under a margin condition and VC-type complexity of the class of candidate predictors. Bartlett and Wegkamp (2006) also study the convergence rates under a margin condition and Bernstein-type complexity, where they employ a binary hinge loss as a surrogate for the true binary reject-loss. Wegkamp (2007) studies the lasso type penalty approach and uses a convex surrogate loss to replace the true binary reject loss.

In this chapter we study the multicategory case. We consider the so-called multicategory reject-loss and study the fast rates of convergence of the excess risk under a margin condition and some complexity constraints for the classes of candidate predictors. In Section 5.2 we give the setup of the model with the

general cost and reject option. In Section 5.3 we discuss the conditions with respect to the reject-loss and the main result on the rates of convergence of the excess risk. Section 5.4 is the proof of the main result.

5.2 General cost and reject option

The 0–1 loss (1.1) can be regarded as an $m \times m$ cost matrix having value 0 on the diagonal and 1 elsewhere, where a correct prediction costs 0 and an incorrect prediction costs 1. It is the simplest case in classification loss, and it gives equal cost of 1 for any incorrect prediction. One can generalize the loss to allow the case that misclassification costs from one class to the others and vice-versa are not equal. For example, the cost of misclassifying a spam email as non-spam is smaller than that of misclassifying a non-spam email as spam. Let $c_{jk} \in [0,1]$ denote the cost of misclassifying from class j to class k, with $c_{jj} = 0$, for $j,k = 1,\ldots,m$. The restriction that $c_{jk} \leq 1$ is not essential, it is needed only for normalization. Based on the general cost matrix

$$C = (c_{jk}) = \begin{cases} c_{jk} \leq 1 & ; \quad j \neq k \text{ and } j,k = 1,\ldots,m \\ 0 & ; \quad j = k \end{cases} \tag{5.1}$$

the general true loss can be defined as

$$\tilde{l}(Y, g(X)) := \sum_{j=1}^{m} \mathbb{1}(g(X) = j)\Big(\sum_{k=1,k\neq j}^{m} c_{kj}\mathbb{1}(Y = k)\Big) \,. \tag{5.2}$$

Clearly, when $c_{jk} = 1$ for $j \neq k$, (5.2) reduces to the 0–1 loss (1.1).

In the previous setup of the m-category classification problem, under a general cost matrix, a predictor $g : \mathbb{R}^d \to \{1,\ldots,m\}$ is designed to always accept an observation and decide its category. We call this the standard setup. One may be interested in a non-standard setup in the sense that we allow a predictor to reject taking a decision on the categories if the observation is too hard to classify. This is called *classification with reject option*. The reject option can improve performance in applications for which the cost of rejecting certain samples, and handling them with different procedures (e.g., manual classification), is not larger than the cost of misclassifying.

We can think of embedding the reject option as adding the *rejection* category, which we shall denote by ®, into the output space. That is, g now takes value in $\{1,\ldots,m,®\}$. Recall that c_{jk} in (5.1) denotes the cost of misclassifying from category j to category k, with $j,k = 1,\ldots,m$. Let $0 < \alpha \leq \max_{j,k} c_{jk}$

denotes the cost of rejection. The value of α is known in advance. Now the reject cost matrix is an $m+1$ by $m+1$ matrix defined as

$$C_{®} = (r_{jk}) = \begin{cases} c_{jk} & ; \quad j \neq k \text{ and } j,k = 1,\ldots,m \\ \alpha & ; \quad j = 1,\ldots,m \text{ and } k = m+1 \\ 0 & ; \quad \text{else} . \end{cases} \tag{5.3}$$

The corresponding true general α-reject loss is

$$\begin{aligned} \tilde{l}_{®}(Y, g(X)) &= \sum_{j=1}^{m} \mathbb{1}(g(X) = j) \left(\sum_{k=1, k\neq j}^{m} c_{kj} \mathbb{1}(Y = k) \right) \\ &\quad + \alpha \ \mathbb{1}(g(X) = ®) . \end{aligned} \tag{5.4}$$

We consider the case that $c_{jk} = 1$ in (5.3). The corresponding true loss based on (5.4) is

$$\tilde{l}(g) = \tilde{l}_{®}(Y, g(X)) = \sum_{j=1}^{m} \mathbb{1}(g(X) = j, Y \neq j) + \alpha \ \mathbb{1}(g(X) = ®) , \tag{5.5}$$

which we call *0–1–α loss*. Recall that $p_j(x)$ is the conditional probability $P(Y = j|X = x)$, $x \in \mathcal{X}$. The risk wrt. the loss (5.5) is

$$\widetilde{R}(g) = \mathbb{E}_X[\ \sum_{j=1}^{m} (1 - p_j(X)) \mathbb{1}(g(X) = j) + \alpha \ \mathbb{1}(g(X) = ®) \ ] .$$

The predictor that minimizes the risk over all possible predictors is the one that chooses category j when $1 - p_j \leq 1 - p_k$ for all $k \neq j$ and $1 - p_j < \alpha$. This verifies that Bayes predictor is

$$g^* = \begin{cases} j & ; \quad p_j = \max(p_1, \ldots, p_m, 1-\alpha) \\ ® & ; \quad 1 - \alpha = \max(p_1, \ldots, p_m, 1-\alpha) . \end{cases}$$

To make sense of the Bayes predictor above, clearly we have to take

$$0 < \alpha \leq \frac{m-1}{m} .$$

When $\alpha > (m-1)/m$ (that is, $1 - \alpha < 1/m$), we would never reject because there would be a conditional probability p_j for some $j = 1, \ldots, m$, that is larger than $1/m$. The minimum $\tilde{l}$-risk is then

$$\widetilde{R}^* := \widetilde{R}(g^*) = \mathbb{E}_X[\max\{p_1(X), \ldots, p_m(X), 1-\alpha\}] .$$

We associate a predictor function g with a predictor set $G = (G_1, \ldots, G_m, G_{®})$ with $\cup_{j=1}^{m} G_j \cup G_{®} = \mathcal{X}$, where

$$g(x) = j \iff x \in G_j \ ,$$

for each $x \in \mathcal{X}$, and $j \in \{1, \ldots, m, ®\}$. In the case there exist $j, k = 1, \ldots, m+1$ with $j < k$ such that $p_j(x) = p_k(x)$, $x \in \mathcal{X}$, we assign x to the set with the smallest index, that is, G_j. Here we set ® as the $m+1$-th class. Thus, a predictor set is a vector of mutually disjoint sets. Bayes predictor vector set is $G^* = (G_1^*, \ldots, G_m^*, G_{®}^*)$ with

$$G_j^* = \{x : \ p_j(x) = \max(p_1(x), \ldots, p_m(x), 1-\alpha)\} \ ,$$

$$G_{®}^* = \{x : \ 1-\alpha = \max(p_1(x), \ldots, p_m(x), 1-\alpha)\} \ ,$$

for $j = 1, \ldots, m$. For each $x \in \mathcal{X}$, it must be in one of the component sets of G^*. We fix an arbitrary $x \in \mathcal{X}$ and use the short-hand notation $\mathbb{1}_G = \mathbb{1}(x \in G)$. Intersecting each component of G with all components of G^*, the excess risk can be written as

$$\begin{aligned}
&\widetilde{R}(G) - \widetilde{R}^* \\
&= \mathbb{E}_X \left[\sum_{j=1}^{m} (1-p_j)(\mathbb{1}_{G_j} - \mathbb{1}_{G_j^*}) + \alpha(\mathbb{1}_{G_{®}} - \mathbb{1}_{G_{®}^*}) \right] \\
&= \mathbb{E}_X \left[\sum_{j=1}^{m} \sum_{k \neq j} |p_j - p_k| \mathbb{1}_{G_j \cap G_k^*} + \sum_{j=1}^{m} |1-\alpha-p_j| \mathbb{1}_{G_{®} \cap G_j^*} + \right. \\
&\qquad \left. + \sum_{j=1}^{m} |1-\alpha-p_j| \mathbb{1}_{G_j \cap G_{®}^*} \right] .
\end{aligned}$$

We consider the ERM-based predictors

$$\hat{G}_n := \arg\min_{G \in \mathcal{G}} \widetilde{R}_n(G) \tag{5.6}$$

where $\widetilde{R}_n(G)$ is the empirical 0–1–α risk

$$\widetilde{R}_n(G) = \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{j=1}^{m} \mathbb{1}(X_i \in G_j, Y_i \neq j) + \alpha \ \mathbb{1}(X_i \in G_{®}) \right]$$

and $\mathcal{G}$ is a given collection of subsets of $\mathbb{R}^d$.

Fast rates of convergence of the excess risk of $\hat{G}_n$ depend on the complexity constraint on $\mathcal{G}$ and the margin condition. In the next section we state the conditions and our main result.

5.3 The conditions and the main result

Let us write

$$\Delta(G, G') := \mathcal{X} / \left(\cup_{j=1}^{m} G_j \cap G'_j \cup G_{\circledR} \cap G'_{\circledR} \right)$$

for any vector sets G and G' in $\mathbb{R}^d$. Let Q be the marginal distribution of X. We now define a pseudo-distance between two vector sets:

$$d_\Delta(G, G') := Q(\Delta(G, G')) \ .$$

Clearly we have the relations $\widetilde{R}(G) - \widetilde{R}^* \leq d_\Delta(G, G^*) \leq 1$, for any vector set G in $\mathbb{R}^d$. We define the empirical analogue of d_Δ as follows:

$$d_{\Delta,n}(G, G') = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \in \Delta(G, G')) \ .$$

Below is the margin condition.

Condition A *(Margin condition).* There exist constants $\sigma > 0$ and $\kappa \geq 1$ such that for all $G \in \mathcal{G}$,

$$\widetilde{R}(G) - \widetilde{R}^* \ \geq \ \left(\frac{d_\Delta(G, G^*)}{\sigma} \right)^{\kappa} \ .$$

Lemma 5.3.1 below explains the origin of Condition A. That is, the condition on the behaviour of the conditional probabilities p_j.

Define, for $j, k = 1, \ldots, m$,

$$\tilde{\tau}(x) := \min_{j \neq k} \{ |p_j(x) - p_k(x)|, |1 - \alpha - p_j(x)| \} \ , \ x \in \mathcal{X} \ . \tag{5.7}$$

Condition AA. Let $\tilde{\tau}$ be defined as in (5.7). There exist constants $C \geq 1$ and $\gamma \geq 0$ such that $\forall z > 0$,

$$Q(\{\tilde{\tau} \leq z\}) \leq (Cz)^{1/\gamma} \ .$$

[Here we use the convention $(Cz)^{1/\gamma} = \mathbb{1}\{z \geq 1/C\}$ for $\gamma = 0$.]

Lemma 5.3.1. *Suppose Condition AA is met and G^* is Bayes vector set predictor. Then for all $G \in \mathcal{G}$ we have*

$$\widetilde{R}(G) - \widetilde{R}^* \geq \frac{1}{\widetilde{\sigma}}\, d_{\Delta}^{1+\gamma}(G, G^*) \ ,$$

with $\widetilde{\sigma} = C\left((1/\gamma + 1)\right)^{\gamma}(1+\gamma)$. That is, Condition A holds with $\sigma = \widetilde{\sigma}^{1/\kappa}$ and $\kappa = 1 + \gamma$.

Proof. Let $\tilde{\tau}$ defined as in (5.7). We have

$$\begin{aligned}
\widetilde{R}(G) - \widetilde{R}^* &\geq \int_{\Delta(G,G^*)} \widetilde{\tau}\, dQ \\
&\geq \int_{\Delta(G,G^*)\cap\{\widetilde{\tau} > z\}} \widetilde{\tau}\, dQ \\
&\geq z\, [Q(\Delta(G,G^*)) - Q(\{\widetilde{\tau} \leq z\})] \\
&\geq z\, [Q(\Delta(G,G^*)) - (Cz)^{1/\gamma}] \ .
\end{aligned}$$

We take $z = Q^{\gamma}(\Delta(G,G^*))(\gamma/(1+\gamma))^{\gamma}C^{-1}$ when $\gamma > 0$, and $z \nearrow 1/C$ when $\gamma = 0$. ■

Now we state the complexity constraints of the model class. Let $\mathcal{G}$ be a class of vector subsets in $\mathbb{R}^d$ with the following complexity constraints.

Condition B1 *(Complexity constraint under ϵ-entropy with bracketing).* Let $0 < \rho < 1$ and let A be a positive constant. The ϵ-entropy with bracketing satisfies the inequality

$$H_B(\epsilon, \mathcal{G}, d_{\Delta}) \leq A\epsilon^{-\rho} \ , \quad \text{for all } \epsilon > 0 \ .$$

Condition B2 *(Complexity constraint under empirical ϵ-entropy).* Let $0 < \rho < 1$ and let A be a positive constant. The empirical ϵ-entropy satisfies the following inequality, almost surely for all $n \geq 1$.

$$H(\epsilon, \mathcal{G}, d_{\Delta,n}) \leq A\epsilon^{-\rho} \ , \quad \text{for all } \epsilon > 0 \ .$$

Now we come to the main result.

Theorem 5.3.1. *Assume Condition A is met. Suppose that either Condition B1 or Condition B2 holds. Let $\hat{G}_n$ be the 0–1–α loss minimizer defined in (5.6).*

Then for small values of $\delta > 0$,

$$\mathbb{E}[\widetilde{R}(\hat{G}_n) - \widetilde{R}^*] \leq \frac{1+\delta}{1-\delta} \ \inf\left\{\widetilde{R}(G) - \widetilde{R}^* + C_0 \ n^{-\frac{\kappa}{2\kappa-1+\rho}} \ : \ G \in \mathcal{G}\right\}$$

with C_0 some constant depending only on κ, σ, A and ρ.

5.4 Proof of Theorem 5.3.1

We follow the set of arguments as in the proof of Theorem 4.2.1 in Section 4.3. Let $G^o := \arg\min_{G\in\mathcal{G}} \widetilde{R}(G)$, the minimizer of the theoretical 0–1–α risk in the model class $\mathcal{G}$. We write $\nu_n(G) := \sqrt{n}\ (\widetilde{R}_n(G) - \widetilde{R}(G))$, and we have the basic inequality

$$\widetilde{R}(\hat{G}_n) - \widetilde{R}^* \leq |\nu_n(\hat{G}_n) - \nu_n(G^o)|/\sqrt{n} + \widetilde{R}(G^o) - \widetilde{R}^* \ .$$

Now, we write

$$Z_n(G) := \frac{|\nu_n(G) - \nu_n(G^o)|}{\left(d_\Delta^{1/2}(G, G^o) \vee n^{-\frac{1}{2+2\rho}}\right)^{1-\rho}} \ , \ G \in \mathcal{G} \ ,$$

where $(a \vee b) := \max\{a, b\}$ and ρ is from either Condition B1 or B2. As short hand notation, we also write $Z_n = Z_n(\hat{G}_n)$. Then

$$\begin{aligned}
&\widetilde{R}(\hat{G}_n) - \widetilde{R}^* \\
&\leq (Z_n/\sqrt{n}) \ \left(\ d_\Delta^{\frac{1-\rho}{2}}(\hat{G}_n, G^o) \vee n^{-\frac{1-\rho}{2+2\rho}}\right) \ + \ \widetilde{R}(G^o) - \widetilde{R}^* \ . \qquad (5.8)
\end{aligned}$$

The triangular inequality and Condition A give

$$d_\Delta^{\frac{1-\rho}{2}}(\hat{G}_n, G^o) \leq \sigma^{\frac{1-\rho}{2}} \ \left\{[\widetilde{R}(\hat{G}_n) - \widetilde{R}^*]^{\frac{1-\rho}{2\kappa}} + [\widetilde{R}(G^o) - \widetilde{R}^*]^{\frac{1-\rho}{2\kappa}}\right\} \ .$$

Denote by $\mathcal{R}$ the right hand side of the above inequality. Hence, from (5.8) we have

$$\widetilde{R}(\hat{G}_n) - \widetilde{R}^* \leq (Z_n/\sqrt{n}) \ (\mathcal{R} \vee n^{-\frac{1-\rho}{2+2\rho}}) + \widetilde{R}(G^o) - \widetilde{R}^* \ .$$

Similarly as in Section 4.3, we consider first the case $(\mathcal{R} \vee n^{-\frac{1-\rho}{2+2\rho}}) = \mathcal{R}$. That is,

$$\begin{aligned}
\widetilde{R}(\hat{G}_n) - \widetilde{R}^* \ \leq \ & \frac{Z_n}{\sqrt{n}} \ \sigma^{\frac{1-\rho}{2}} \ \left\{[\widetilde{R}(\hat{G}_n) - \widetilde{R}^*]^{(1-\rho)/2\kappa} \ + \right. \\
& \left. + [\widetilde{R}(G^o) - \widetilde{R}(G^*)]^{(1-\rho)/2\kappa}\right\} + \ \widetilde{R}(G^o) - \widetilde{R}^* \ .
\end{aligned}$$

We apply Lemma 4.3.2 twice, where we set $t = \widetilde{R}(\hat{G}_n) - \widetilde{R}^*$ and $t = \widetilde{R}(G^o) - \widetilde{R}(G^*)$, with $\beta = 1 - \rho$, $\kappa = 2\kappa$ and $\nu = Z_n \ \sigma^{\frac{1-\rho}{2}}/\sqrt{n}$ for the two choices of t. We obtain, for all $0 < \delta < 1$,

$$\widetilde{R}(\hat{G}_n) - \widetilde{R}^* \leq \frac{1+\delta}{1-\delta} \ \inf\left\{\widetilde{R}(G) - \widetilde{R}^* + C_1 \ Z_n^r \ n^{-\frac{\kappa}{2\kappa-1+\rho}} \ : \ G \in \mathcal{G}\right\} ,$$

with $C_1 = 2 \ \sigma^{\frac{1-\rho}{2}r} \ \delta^{-\frac{1-\rho}{2\kappa-1+\rho}}$ and $r = 2\kappa/(2\kappa - 1 + \rho)$. We now need to show that $\mathbb{E}[Z_n^r]$ is bounded by some constant, say C_2. Then, $C_0 = C_1 C_2$ in Theorem 5.3.1.

To obtain an exponential tail probability of $Z_n(\hat{G}_n)$, under Condition B1, we apply Lemma 4.3.3 where we set $h = h_G(x) = \mathbb{1}(x \in \Delta(G, G^o))$ and $h^o \equiv 0$. Then, $\|h - h^o\|_{2,Q}^2 = d_\Delta(G, G^o)$. We note that $H_B(\epsilon, \mathcal{G}, L_2(Q)) = H_B(\epsilon^2, \mathcal{G}, d_\Delta) \leq 2\epsilon^{-2\rho}$.

For the case $\mathcal{R} \leq n^{-\frac{1-\rho}{2+2\rho}}$, the same arguments as in the proof of Theorem 4.2.1 hold.

For the case where Condition B2 holds instead of B1, we follow the same arguments as in Section 4.3. We use that $\|h - h^o\|_{2,Q_n}^2 = d_{\Delta,n}(G, G^o)$, with $d_{\Delta,n}(G, G^o) = (1/n)\sum_{i=1}^n \mathbb{1}(X_i \in \Delta(G, G^o))$. ■

Appendix

Definition of entropy Let $\mathcal{G}$ be a subset of a metric space (Λ, d). Let

$$H(\epsilon, \mathcal{G}, d) := \log N(\epsilon, \mathcal{G}, d) \ , \ \text{for all } \epsilon > 0 \ ,$$

where $N(\epsilon, \mathcal{G}, d)$ is the smallest value of N for which there exist functions $g_1, \ldots, g_N$ in $\mathcal{G}$, such that for each $g \in \mathcal{G}$, there is a $j = j(g) \in \{1, \ldots, N\}$, such that

$$d(g, g_j) \leq \epsilon \ .$$

Then $N(\epsilon, \mathcal{G}, d)$ is called the *ϵ-covering number* of $\mathcal{G}$ and $H(\epsilon, \mathcal{G}, d)$ is called the *ϵ-entropy* of $\mathcal{G}$ (for the d-metric).

Definition of entropy with bracketing Let $\mathcal{G}$ be a subset of a metric space (Λ, d) of real-valued functions. Let

$$H_B(\epsilon, \mathcal{G}, d) := \log N_B(\epsilon, \mathcal{G}, d) \ , \ \text{for all } \epsilon > 0 \ ,$$

where $N_B(\epsilon, \mathcal{G}, d)$ is the smallest value of N for which there exist pairs of functions $\{[g_1^L, g_1^U], \ldots, [g_N^L, g_N^U]\}$ in $\mathcal{G}^2$, such that $d(g_j^L, g_j^U) \leq \epsilon$ for all $j = 1, \ldots, N$, and such that for each $g \in \mathcal{G}$, there is a $j = j(g) \in \{1, \ldots, N\}$ such that

$$g_j^L \leq g \leq g_j^U \ .$$

Then $N_B(\epsilon, \mathcal{G}, d)$ is called the *ϵ-covering number with bracketing* of $\mathcal{G}$ and $H_B(\epsilon, \mathcal{G}, d)$ is called the *ϵ-entropy with bracketing* of $\mathcal{G}$ (for the d-metric).

Bibliography

Audibert, Jean-Yves. 2004. Classification under polynomial entropy and margin assumptions and randomized estimators. Preprint, Laboratoire de Probabilités et Modèles Aléatoires. URL `\url{www.proba.jussieu.fr/mathdoc/textes/PMA-908.pdf}`.

Bartlett, Peter L., Michael I. Jordan, and Jon D. McAuliffe. 2006. Convexity, classification and risk bounds. *Journal of the American Statistical Association* 101(473):138–156.

Bartlett, Peter L., and Marten H. Wegkamp. 2006. Classification with a reject option using a hinge loss. Submitted.

Blanchard, Gilles, Olivier Bousquet, and Pascal Massart. 2004. Statistical performance of support vector machines. `www.kyb.mpg.de/publication.htm?user=bousquet`. Manuscript.

Blanchard, Gilles, Gábor Lugosi, and Nicolas Vayatis. 2003. On the rate of convergence of regularized boosting classifiers. *Journal Machine Learning Research* 4:861–894.

Boser, Bernhard E., Isabelle Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *COLT*, 142–152. Pittsburgh ACM.

Boucheron, Stéphane, Olivier Bousquet, and Gábor Lugosi. 2005. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics* 9:323–375.

Bradley, Paul, and Olvi Mangasarian. 1998. Feature selection via concave minimization and support vector machines. In *International Conference on Machine Learning*. Morgan Kaufman.

Burges, Christopher J. C. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2(2):121–167.

Candès, Emmanuel J., and David L. Donoho. 2004. New tight frames of curvelets and optimal representations of objects with piecewise c^2 singularities. *Communications on Pure and Applied Mathematics* LVII:219–266.

Chaaban, Ibrahim, and Michael Scheessele. 2007. Human performance on the usps database. Tech. rep., Indiana Universit South Bend. URL `\url{http://www.cs.iusb.edu/technical_reports/TR-20070619-1.pdf}`.

Christianini, Nello, and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

Crammer, Koby, and Yoram Singer. 2000. On the learnability and design of output codes for multiclass problems. In *Proceeding of the 13th Annual Conference on Computational Learning Theory*, 35–46. Morgan Kaufmann.

———. 2001. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research* 2:265–292.

del Barrio, Eustasio, Paul Deheuvels, and Sara A. van de Geer. 2007. *Lectures on Empirical Processes*. EMS Series of Lectures in Mathematics, European Mathematical Society.

Devroye, Luc, László Györfi, and Gábor Lugosi. 1996. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York.

Donoho, David L. 1995. Denoising via soft-thresholding. *IEEE Transactions on Information Theory* 41:613–627.

———. 1999. Wedgelets: nearly minimax estimation of edges. *The Annals of Statistics* 27:859–897.

———. 2004a. For most large underdetermined systems of equations, the minimal l^1-norm near-solution approximates the sparsest near-solution. Tech. rep., Standford University. URL `\url{www-stat.stanford.edu/~donoho/Reports/2004/l1l0approx.pdf}`.

———. 2004b. For most large underdetermined systems of linear equations, the minimal l^1-norm solution is also the sparsest solution. Tech. rep., Standford University. URL `\url{www-stat.stanford.edu/~donoho/Reports/2004/l1l0EquivCorrected.pdf}`.

Duan, Kaibo, and S. Sathiya Keerthi. 2005. Which is the best multiclass svm method? An empirical study. In *Multiple Classifier Systems*, 278–285. No. 3541 in Lecture Notes in Computer Science, Springer Berlin/Heidelberg.

Duan, Kaibo, S. Sathiya Keerthi, and Aun Neow Poo. 2003. Evaluation of simple performace measures for tuning svm hyperparameters. *Neurocomputing* 51:41–59.

Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. 2004. Least angle regression. *The Annals of Statistics* 32(2):407–499.

Guermeur, Yann. 2002. Combining discriminant models with new multiclass svms. *Pattern Analysis & Applications* 5:168–179.

Guyon, Isabelle. 2008. SVM application list. `http://www.clopinet.com/isabelle/Projects/SVM/applist.html`.

Hardy, Godfrey H., John E. Littlewood, and George Pólya. 1988. *Inequalities*. Cambridge University Press, Cambridge, 2nd ed.

Hastie, Trevor, Saharon Rosset, Robert Tibshirani, and Ji Zhu. 2004. The entire regularization path for the support vector machine. *The Journal of Machine Learning Research* 5:1391–1415.

Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. 2001. *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. Springer-Verlag, New York.

Herbei, Radu, and Marten H. Wegkamp. 2006. Classification with reject option. *The Canadian Journal of Statistics* 34:709–721.

Hofmann, Thomas. 2003. Introduction to machine learning. Lecture Notes.

Hofmann, Thomas, Benhard Schölkopf, and Alexander J. Smola. 2007. Kernel methods in machine learning. *The Annals of Statistics* Submitted.

Hsu, Chih-Wei, and Chih-Jen Lin. 2002. A comparison methods for multiclass support vector machines. *IEEE Transactions on Neural Networks* 13(2):415–425.

Koltchinskii, Vladimir I. 2001. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory* 47:1902–1914.

———. 2006. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics* 34:2593–2656.

Koltchinskii, Vladimir I., and Dmitry Panchenko. 2002. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics* 30:1–50.

Ledoux, Michel. 1996. On Talagrand's deviation inequalities for product measures. *ESIAM: Probability and Statistics* 1:63–87.

Ledoux, Michel, and Michel Talagrand. 1991. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York.

Lee, Yoonkyung. 2002. *Multicategory Support Vector Machines, Theory and Application to the Classification of Microarray Data and Satellite Radiance Data*. Ph.D. thesis, University of Wisconsin–Madison, Departement of Statistics.

Lee, Yoonkyung, and Zhenhuan Cui. 2006. Characterizing the solution path of multicategory support vector machines. *Statistica Sinica* 16(2):391–409.

Lee, Yoonkyung, Yi Lin, and Grace Wahba. 2004. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association* 99(465):67–81.

Lin, Yi. 2002. Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery* 6(3):259–275.

———. 2004. A note on margin-based loss functions in classification. *Statistics and Probability Letters* 73–82.

Loubes, Jean-Michel, and Sara A. van de Geer. 2002. Adaptive estimation in regression, using soft thresholding type panalties. *Statistica Neerlandica* .

Lugosi, Gábor, and Marten Wegkamp. 2004. Complexity regularization via localized random penalties. *The Annals of Statistics* 32:1679–1697.

Mammen, Enno, and Alexandre B. Tsybakov. 1999. Smooth discrimination analysis. *The Annals of Statistics* 27:1808–1829.

Massart, Pascal. 2000. About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability* 28:863–884.

Pollard, David. 1984. *Convergence of Stochastic Processes*. Springer-Verlag New York Inc.

Rosset, Saharon, and Ji Zhu. 2007. Piecewise linear regularized solution paths. *The Annals of Statistics* 35(3):1012–1030.

Schölkopf, Benhard, Chris Burges, and Vladimir N. Vapnik. 1995. Extracting support data for a given task. In *First International Conference in Knowledge Discovery and Data Mining*, eds. U. M. Fayyad and R. Uthurusamy. AAAI Press.

Schölkopf, Bernhard, and Alexander J. Smola. 2002. *Learning with Kernels*. MIT Press, Cambridge.

Scott, Clayton, and Robert Nowak. 2006. Minimax-optimal classification with dyadic decision trees. *IEEE Transactions on Information Theory* 52:1335–1353.

Shorack, Galen R., and Jon A. Wellner. 1986. *Empirical Processes with Applications to Statistics*. Wiley, New York.

Song, Shuguang, and Jon A. Wellner. 2002. An upper bound for uniform entropy numbers. URL `\url{www.stat.washington.edu/www/research/reports/\#2002/tr409.ps}`.

Steinwart, Ingo, and Clint Scovel. 2005a. Fast rates for support vector machines. In *COLT*, 279–294.

———. 2005b. Fast rates for support vector machines using Gaussian kernels. Technical Report LA-UR 04-8796, Los Alamos National Laboratory. URL `\url{www.cs.lanl.gov/ml/pubs/2005_fastratesa/paper.pdf}`. Submitted to Ann. Statist.

Tarigan, Bernadetta, and Sara A. van de Geer. 2004. Adaptivity of support vector machines with l_1 penalty. Technical Report MI 2004–14, University of Leiden. URL `\url{www.stat.math.ethz.ch/~geer/reports.html}`.

———. 2006. Classifiers of support machine type with l_1 complexity regularization. *Bernoulli* 12(6):1045–1076.

———. 2007. A moment inequality for multicategory support vector machines. Technical Report 144, Seminar for Statistics, ETH Zurich. Accepted for publication in the Journal of Machine Learning Research, Dec. 2007.

Tewari, Ambuj, and Peter L. Bartlett. 2005. On the consistency of multiclass classification methods. In *COLT*, 143–157.

Tibshirani, Robert. 1996. Regression analysis and selection via the LASSO. *Journal of Royal Statistics Society B* 58:267–288.

Tsybakov, Alexandre B. 2004. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics* 32:135–166.

Tsybakov, Alexandre B., and Sara A. van de Geer. 2005. Square root penalty: adaptation to the margin in classification and in edge estimation. *The Annals of Statistics* 33:1203–1224.

van de Geer, Sara A. 2000. *Empirical Processes in M-estimation*. Cambridge University Press.

———. 2003. Adaptive quantile regression. In *Recent Advances and Trends in Nonparametric Statistics*, eds. M. G. Akritas and D. N. Politis, 235–250. Elsevier.

van der Vaart, Aad W., and Jon A. Wellner. 1996. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

Vapnik, Vladimir N. 1998. *Statistical Learning Theory*. Wiley, New York.

———. 2000. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 2nd ed.

Wahba, Grace. 1990. Spline models for observational data. *Philadelphia: SIAM* .

Wahba, Grace, Yi Lin, and Hao Zhang. 2000. Generalized approximate cross validation for support vector machines. 297–311. Advances in Large Margin Classifiers.

Wang, Li, Ji Zhu, and Hui Zou. 2006. The doubly regularized support vector machine. *Statistica Sinica* 16:589–615.

———. 2007. Hybrid huberized support vector machines for microarray classification. In *Proceedings of the 24-th International Conference on Machine Learning*.

Wang, Lifeng, and Xiaotong Shen. 2007. On l_1-norm multiclass support vector machines: Methodology and theory. *Journal of the American Statistical Association* 102:583–594.

Wegkamp, Marten H. 2007. Lasso type classifiers with a reject option. *Electronic Journal of Statistics* 1:155–168.

Weston, Jason, and Chris Watkins. 1999. Multi-class support vector machines. In *Proceedings of ESANN99*.

Zhang, Tong. 2004a. An infinity-sample theory for multi-category large margin classification. In *Advances in Neural Information Processing Systems 16*, eds. Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf. MIT Press.

———. 2004b. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research* 5:1225–1251.

———. 2004c. Statistical behaviour and consistency of classification methods based on convex risk minimization. *The Annals of Statistics* 32:56–84. With discussion.

Zhu, Ji, Saharon Rosset, Trevor Hastie, and Robert Tibshirani. 2004. 1-norm support vector machines. In *Advances in Neural Information Processing Systems 16*, eds. Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf. MIT Press.

Zou, H., and T. Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society B* 67:301–320.

Zou, Hui, Ji Zhu, and Trevor Hastie. 2006. The margin vector, admissible loss and multi-class margin-based classifiers. Tech. rep., Statistics Departement, Stanford University.

Notation

$\mathcal{X} = \mathbb{R}^d$	domain of the input X
$\mathcal{Y} = \{1, \ldots, m\}$	domain of the category Y, $m \geq 2$
$m \geq 2$	number of categories
P	joint (cumulative) dist. (X, Y)
	also prob. measure wrt. (X, Y)
Q	marginal (cumulative) dist. X
	also prob. measure wrt. X
$p_j(x)$, $j = 1, \ldots, m$	conditional prob. $Y = j$ given $X = x$
$\eta(x)$	$p_1(x)$ in binary case ($m = 2$)
$D_n = \{(X_i, Y_i)\}_{i=1}^n$	iid. sample of (X, Y)
$\mathbb{P}$	prob. measure wrt. $(X, Y)^n$
	also general prob. measure
$\mathbb{E}$	expectation wrt. $(X, Y)^n$
	also general expectation
P_n	empirical measure of (X, Y)
Q_n	empirical measure of X
$g : \mathcal{X} \to \mathcal{Y}$	predictor
$f : \mathcal{X} \to \mathbb{R}^m$	classifier
$\mathcal{G}$	class of predictors or model class
$\mathcal{F}$	class of classifiers or model class
$\tilde{l}(Y, g(X)) = \mathbb{1}(g(X) \neq Y)$	(standard) 0–1 loss function
$l(Y, g(X)) = l_Y(g(X))$	proxy loss function
$P(g(X) \neq Y)$	prediction error
$\tilde{R}(g) = R(\tilde{l}_g) = \mathbb{E}[\tilde{l}(Y, g(X))]$	standard-risk or prediction error
$R(g) = R(l_g) = \mathbb{E}[l(Y, g(X))]$	proxy-risk
ρ	complexity parameter of model class
κ	margin parameter or noise level

Acknowledgements

It has been a long but beautiful road with many views, mathematically and non-mathematically. There have been many great people walking with me on the path, and this page by no means can cover all the names without whom this thesis would be impossible. Nevertheless, I take this chance to express my joy and my gratitude to you.

Sara, thank you for being my *Doktormutter*, providing assistance in numerous ways and bringing me into the realm of learning in statistics. My thanks also go to the co-examiners: Prof. Bühlmann and Dr. Blanchard. Behind the scene were my critical proofreaders: Misja, Rachel and Charles; guys, *bedankt vielmals*!

I thank my colleagues in SfS-Zurich for the (bike-to) work, the OL's, the (glacier) alpine-hikes and the muscle-pumps that strengthen also my scientific vessels. I am grateful to my other colleagues back in ITB-Bandung, CWI-Amsterdam and Leiden, for making it possible for me to carry through and enjoy the research.

In the CWI-phase, the city of *magerebrug* would not be meaningful without Wayan, Sinta, Minnie, Gemma, Stefania, Daniël, Ton, Erich, Sebastian, Simona, Daniela, and Carlota among others, from whom I first learn the west culture and ice-skating. My roommates Rob and Rachel, we went through a lot, yet we rock the Beatles. Leiden was nice to commute to with *vouwfiets* and colorful because of my academic-siblings Jelle, Leila, and Georgia; and the more-friends-than-colleagues Sofia, Federica, Miguel, Szabi, Bart, Luca and Benedetta. I enjoyed the stochastic reading group with Misja, Pieter, Rachel, Stan and Ton. The low-sky country has been *gezellig* also because of the Indonesian scholars Katrin, Pasco, Denny, Sinar, Iwa, Any and Cisca among others, who make the life taste better. My table-tennis bodies: Wim, Misja, Ilja and

Adri, whom I had a great (after-match) time with. My landlady Jenny made my tiny flat comfortable. The M-team has been a great adventurous company with rock concerts, bbq and camping among others; I know how to find you guys.

It was very difficult to start a life in Zurich. Rahel and Mara from SfS have been very accommodating, as well as my landlady Susanne. The Indonesian gangs in Zurich and Bern have provided me food and folksongs. Between business and leisures: Silke, Nicoleta, Patricia, Julian, Yvonne and Peter among others, that makes the Ph.D. life in ETH much nicer. Yv, thanks also for the badminton and the birthday cakes. Peter, thanks also for the *zusammenfassung*. Charles has been not only a very supportive roommate but also a good friend, thank you also for introducing me to bread-baking. Karina has been a sister for me. Sara and Eva have brought me into their Swiss homey-circles in daily life. I thank Bëtte for the beautiful cover and the (snow-shoes) walks.

Back home, to my family who always brings me into pray although they have no idea what my research is about. To my dear fellow country(wo)men, you people never tired of supporting me in the home-sick moments and reminding me in numerous ways the sense of life and to go back home. Some of you visit me in Europe and bring all the exotic-goodies. Mega, you never let me go. *Remuk republik masih tanah air beta, Indonesia sayang.*

In all these years and different places, I went through many difficult situations. I almost gave up but I did it, to finish what I have started. Stefi, my master-lady, not only you have spoiled me with the opera but also you have made the going-home-after-work makes more sense. Racheltje, you are always there when I need you the most. Da Nuyens, *ik ben zo gelukkig* to be part of you. *Smies met skiwi*, so much you give without asking back, I dunno know how to thank you. My *sinyo kunyuk besar*, I owe you. And, all my fellow pilgrims. You keep me faithful to solitude. Wherever we are, there is no boundary for love.

Curriculum Vitae

I was born in Tebing Tinggi, Indonesia, as the sixth and the last child of Albina Ginting and Ngayam Bernadinus Tarigan (†). I stayed in North Sumatra before I moved to West Java to continue my education. I finished high school in 1990 at Regina Pacis, Bogor. Then I moved to Bandung where I obtained a bachelor degree in 1995 and a master degree in 1998, both in Mathematics, at Institut Teknologi Bandung (ITB). Since 1995 I have been recruited as a government employee (*pegawai negeri sipil, PNS*) and worked at ITB as a teaching assistant.

I left to Holland in autumn 2000 where I did research on ruin probability theory at Centrum voor Wiskunde en Informatica (CWI), Amsterdam, with a grant from the Royal Netherlands Academy of Arts and Sciences (KNAW) under the EPAM project in the Scientific Programme Indonesia Netherlands. In summer 2003, I started my Ph.D. study in Leiden University as an AIO (*assistent in opleiding*). I followed my supervisor to Switzerland in autumn 2005, to join Seminar für Statistik (SfS), ETH Zurich, as a *doktorandin*, to continue and finish the study.