Journal of Machine Learning Research 6 (2005) 2027–2047 Submitted 3/05; Revised 9/05; Published 12/05

Asymptotics in Empirical Risk Minimization

Leila Mohammadi MOHAMMADI@EURANDOM.TUE.NL
EURANDOM
Post Office Box 513
5600 MB Eindhoven
The Netherlands

Sara van de Geer GEER@STAT.MATH.ETHZ.CH
Seminar für Statistik
ETH-Zentrum, LEO D11, 8092 Zürich
Switzerland

Editor: John Shawe-Taylor

Abstract

In this paper, we study a two-category classification problem. We indicate the categories by labels $Y = 1$ and $Y = -1$. We observe a covariate, or feature, $X \in \mathcal{X} \subset \mathbb{R}^d$. Consider a collection $\{h_a\}$ of classifiers indexed by a finite-dimensional parameter a, and the classifier h_{a^*} that minimizes the prediction error over this class. The parameter a^* is estimated by the empirical risk minimizer $\hat{a}_n$ over the class, where the empirical risk is calculated on a training sample of size n. We apply the Kim Pollard Theorem to show that under certain differentiability assumptions, $\hat{a}_n$ converges to a^* with rate $n^{-1/3}$, and also present the asymptotic distribution of the renormalized estimator.

For example, let V_0 denote the set of x on which, given $X = x$, the label $Y = 1$ is more likely (than the label $Y = -1$). If X is one-dimensional, the set V_0 is the union of disjoint intervals. The problem is then to estimate the thresholds of the intervals. We obtain the asymptotic distribution of the empirical risk minimizer when the classifiers have K thresholds, where K is fixed. We furthermore consider an extension to higher-dimensional X, assuming basically that V_0 has a smooth boundary in some given parametric class.

We also discuss various rates of convergence when the differentiability conditions are possibly violated. Here, we again restrict ourselves to one-dimensional X. We show that the rate is n^{-1} in certain cases, and then also obtain the asymptotic distribution for the empirical prediction error.

Keywords: asymptotic distribution, classification theory, estimation error, nonparametric models, threshold-based classifiers

1. Introduction

In the theory of classification, the problem is to predict the unknown nature of a feature. The topic plays a basic role in several fields, such as data mining, artificial intelligence and neural networks. In this paper we discuss the classification problem from a parametric-statistical point of view.

Let the training set $(X_1,Y_1),\cdots,(X_n,Y_n)$ consist of n independent copies of the couple (X,Y) with distribution P, where $X \in \mathcal{X} \subset \mathbb{R}^d$ is called a feature and $Y \in \{-1,1\}$ is the label of X. A classifier h is a function $h : \mathcal{X} \to \{-1,1\}$, attaching the label $h(X)$ to the feature X. The error, or risk, of a classifier h is defined as $P(h(X) \neq Y)$. Following Vapnik (2000) and Vapnik (1998), we

consider the empirical counterpart of the risk which is the number of misclassified examples, i.e.,

$$P_n(h(X) \neq Y) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(h(X_i) \neq Y_i).$$

Here, and throughout, $\mathbb{1}(A)$ denotes the indicator function of a set A. We will study empirical risk minimization over a model class $\mathcal{H}$ of classifiers h. We take $\mathcal{H}$ to be parametric, in the sense that

$$\mathcal{H} = \{h_a : \ a \in \mathcal{A}\},$$

with $\mathcal{A}$ a subset of finite-dimensional Euclidean space.

Let

$$F_0(x) := P(Y = 1|X = x) \tag{1}$$

be the conditional probability of the label $Y = 1$ if the feature X has value x. Given a new feature $x \in \mathcal{X}$, we want to guess whether the label is $Y = 1$ or $Y = -1$. A natural solution is to predict $Y = 1$ when the label $Y = 1$ is more likely than the label $Y = -1$ (Bayes rule). Thus the set

$$V_0 := \{x \in \mathcal{X} : F_0(x) > 1/2\}, \tag{2}$$

plays a key role in classification. Bayes classifier is

$$h_0 = 2\mathbb{1}\{V_0\} - 1.$$

The collection $\mathcal{H}$ of classifiers is viewed as model class for h_0. However, we will not require that $h_0 \in \mathcal{H}$. If $h_0 \notin \mathcal{H}$, the model is misspecified.

In the statistical theory of classification, rates of convergence of empirical classifiers have been studied by a number of researchers, see for example Lugosi and Vayatis (2004), Lugosi and Nobel (1999), Lugosi and Wegkamp (2004), Koltchinskii and Panchenko (2002), Boucheron et al. (2005), Koltchinskii (2003a), Koltchinskii (2003b), Mohammadi (2004) and Tsybakov and van de Geer (2005). These papers generally consider a high-dimensional model class and use regularization to tackle the curse of dimensionality. Rates of convergence for the regularized estimators are obtained, and also non-asymptotic bounds. In this paper, we consider a low-dimensional model class. This means that we place the subject in the context of classical parametric statistics. Under regularity assumptions, one can establish rates, as well as the asymptotic distributions. Indeed, our main aim is to show that one can apply certain statistical results to the classification problem with parametric model class. In practice, one may not be willing to assume a simple parametric model class, as the complexity of the problem is not known a priori. In this sense, our study is primarily a theoretical one.

In Section 2, we generalize the problem considered in Mohammadi and van de Geer (2003). It gives an application of the cube root asymptotics derived by Kim and Pollard (1990). We briefly explain the main idea of the Kim Pollard Theorem. Its exact conditions are given in Section 4. We study in Subsection 2.1 the case where $\mathcal{X}$ is one-dimensional. The set $V_0 \subset \mathbb{R}$ is then a union of disjoint intervals, and our aim is to estimate the boundaries of the intervals. These boundaries will be called thresholds. The situation that V_0 is the union of intervals has also been considered in Breiman et al. (1984). They explain how to use the training set to split the feature space $\mathcal{X}$ and construct trees. See also Kearns et al. (1997) for a comparison of various algorithms in this case.

A simple case, with just one threshold, has been presented in Mohammadi and van de Geer (2003). We will establish the asymptotic behavior of estimators of the thresholds, using the set of classifiers with K thresholds as model class. Here K is fixed, and not bigger than, but not necessarily equal to, the number of thresholds of Bayes classifier. We moreover assume that F_0 is differentiable. In Subsection 2.2, we extend the situation to higher-dimensional feature space, $\mathcal{X} := \mathbb{R}^d$, $d \geq 1$. The problem there is related to assuming a single index model for the regression of Y on X, i.e.,

$$F_0(x) = \eta_0(x^T a^*),$$

where a^* is an unknown vector parameter, and η_0 is an unknown (monotone) function. We let $X = (U,V)$, with $U \in \mathbb{R}^{d-1}$ and $V \in \mathbb{R}$ and minimize the empirical classification error over the classifiers

$$h_a(u,v) := 2\mathbb{1}\{k_a(u) \geq v\} - 1,$$

where a is an r-dimensional parameter and $k_a : \mathbb{R}^{d-1} \to \mathbb{R}$ is some given smooth function of a. Under differentiability conditions, this will again lead to cube root asymptotics.

In Section 3, we study various other rates, and also the asymptotic distribution in the case of a $(1/n)$-rate. We consider here only one-dimensional $\mathcal{X}$. The Kim Pollard Theorem and the proofs of the results in Section 2 are given in Section 4.

We note here that we will mainly concentrate on the estimation of the parameter a^* that minimizes the prediction error over the class $\mathcal{H}$. One may argue that the most interesting and useful subject is perhaps not the convergence of the estimator $\hat{a}_n$ to a^*, but rather the convergence of the prediction error of (the classifier $h_{\hat{a}_n}$ corresponding to) $\hat{a}_n$. We remark however that our approach to study the former is via the latter. For example, in Corollary 2 the asymptotic distribution of the prediction error follows as a corollary.

The conclusion is that by considering some assumptions on the distribution of the data, we can prove rates of convergence and asymptotic distributions. In computer learning theory, usually no or minimal distributional assumptions are made. The results of the present paper give more insight in the dependency of the asymptotic behavior on the underlying distribution.

We consider asymptotics as $n \to \infty$, regarding the sample $(X_1,Y_1),\ldots,(X_n,Y_n)$ as the first n of an infinite sequence of i.i.d. copies of (X,Y). The distribution of the infinite sequence $(X_1,Y_1),(X_2,Y_2),\ldots$ is denoted by $\mathbf{P}$. The marginal distribution function of X is denoted by G. In case that the density of the distribution G of X with respect to Lebesgue measure exists, it is denoted by g. The Euclidean norm is denoted by $\|\cdot\|$.

2. Cube Root Asymptotics

We first examine in Subsection 2.1 the case where the feature space $\mathcal{X}$ is the unit interval in $\mathbb{R}$ so that Bayes rule is the union of some subintervals in $[0,1]$. As model class, we take the union of a, possibly smaller, number of subintervals. Next, we consider in Subsection 2.2 the situation where $\mathcal{X} = \mathbb{R}^d$ with $d > 1$. Our model class is then the class of graphs of smooth parametric functions. In both situations, the class of classifiers $\mathcal{H}$ is parametric, i.e. it is of the form

$$\mathcal{H} = \{h_a : \ a \in \mathcal{A}\},$$

with $\mathcal{A}$ a subset of $\mathbb{R}^r$, where the dimension r is fixed (not depending on n).

Define the empirical risk

$$L_n(a) := P_n(h_a(X) \neq Y), \tag{3}$$

and the theoretical risk

$$L(a) := P(h_a(X) \neq Y). \tag{4}$$

Moreover, let

$$\hat{a}_n = \arg\min_{a \in \mathcal{A}} L_n(a)$$

be the empirical risk minimizer, and let

$$a^* = \arg\min_{a \in \mathcal{A}} L(a)$$

be its theoretical counterpart. We assume that a^* exists and is unique. We also assume that the estimator $\hat{a}_n$ exists, but it need not be unique. In fact, in the situations that we consider, there will be many solutions for $\hat{a}_n$. Our results will hold for any choice of $\hat{a}_n$.

We will derive cube root asymptotics. Let us first sketch where the $n^{-1/3}$-rate of convergence comes from. One may write down the equality

$$L(\hat{a}_n) - L(a^*) = -\left[\nu_n(\hat{a}_n) - \nu_n(a^*)\right]/\sqrt{n} + \left[L_n(\hat{a}_n) - L_n(a^*)\right], \tag{5}$$

with

$$\nu_n(a) = \sqrt{n}\left[L_n(a) - L(a)\right],\ a \in \mathcal{A},$$

being the empirical process indexed by $\mathcal{A}$. Since $L_n(\hat{a}_n) - L_n(a^*) \leq 0$, this equality implies

$$L(\hat{a}_n) - L(a^*) \leq -\left[\nu_n(\hat{a}_n) - \nu_n(a^*)\right]/\sqrt{n}. \tag{6}$$

Under regularity conditions $L(a) - L(a^*)$ behaves like the squared distance $\|a - a^*\|^2$. Moreover, again under regularity conditions, the right hand side of (6) behaves in probability like $\sigma(\hat{a}_n)/\sqrt{n}$, where $\sigma(a)$ is the standard deviation of $[\nu_n(a) - \nu_n(a^*)]$. Due to the fact that we are dealing with indicator functions, the standard deviation of $[\nu_n(a) - \nu_n(a^*)]$ behaves like the *square root* $\|a - a^*\|^{1/2}$ of the distance between a and a^*. Inserting this in (6) yields that $\|\hat{a}_n - a^*\|^2$ is bounded by a term behaving in probability like $\|\hat{a}_n - a^*\|^{1/2}/\sqrt{n}$. But this implies $\|\hat{a}_n - a^*\|$ is of order $n^{-1/3}$ in probability.

Let us continue with a rough sketch of the arguments used for establishing the asymptotic distribution. We may write

$$\hat{a}_n = \arg\min_a \left[n^{\frac{1}{6}} \left[\nu_n(a) - \nu_n(a^*)\right] + n^{\frac{2}{3}} \left[L(a) - L(a^*)\right] \right].$$

When we already have the $n^{-1/3}$-rate, it is convenient to renormalize to

$$n^{\frac{1}{3}}(\hat{a}_n - a^*) = \arg\min_t \left[n^{\frac{1}{6}} \left[\nu_n(a^* + n^{-\frac{1}{3}}t) - \nu_n(a^*)\right] + n^{\frac{2}{3}} \left[L(a^* + n^{-\frac{1}{3}}t) - L(a^*)\right] \right].$$

Now, under differentiability assumptions,

$$n^{\frac{2}{3}} \left[L(a^* + n^{-\frac{1}{3}}t) - L(a^*)\right] \approx t^T \mathcal{V} t/2,$$

where $\mathcal{V}$ is the matrix of second derivatives of L at a^*. Moreover, the process $\{n^{1/6}[\nu_n(a^*+n^{-\frac{1}{3}}t)-\nu_n(a^*)]:\ t\in\mathbb{R}^r\}$ converges in distribution to some zero mean Gaussian process, say W. We then apply the "Argmax" Theorem ("Argmin" Theorem in our case), see e.g., van der Vaart and Wellner (1996). The result is that $n^{1/3}(\hat{a}_n-a^*)$ converges in distribution to the location of the minimum of $\{W(t)+t^T\mathcal{V}t/2:\ t\in\mathbb{R}^r\}$.

Kim and Pollard (1990) make these rough arguments precise. See Section 4 for the exact conditions.

2.1 One-Dimensional Feature Space

With a one-dimensional feature space, $\mathcal{X}=[0,1]$, Bayes rule is described by the number, say K_0, and the locations, say $a^0=(a_1^0,\ldots a_{K_0}^0)^T$, where $2F_0-1$ changes sign. We call the locations of the sign changes *thresholds*. With a sign change we mean that the function has strictly opposite sign in sufficiently small intervals to the left and right side of each threshold. The boundary points $a_0^0=0$ and $a_{K_0+1}^0=1$ are thus not considered as locations of a sign change.

Let $K\in\mathbb{N}$ and U_K be the parameter space

$$U_K:=\{a=(a_1,\ldots,a_K)\in[0,1]^K:a_1<\ldots<a_K\}. \tag{7}$$

Let for $a\in U_K$

$$h_a(x):=\sum_{k=1}^{K+1}b_k\mathbb{1}\{a_{k-1}\le x<a_k\},$$

where $a_0=0$, $a_{K+1}=1$ and $b_1=-1,\ b_{k+1}=-b_k,\ k=2,\ldots,K$. Let $\mathcal{H}$ be the collection of classifiers

$$\mathcal{H}=\{h_a:a\in U_K\}. \tag{8}$$

Let

$$L(a):=P(h_a(X)\neq Y),\ \ L_n(a):=P_n(h_a(X)\neq Y). \tag{9}$$

The empirical risk minimizer is

$$\hat{a}_n:=\arg\min_{a\in U_K}L_n(a). \tag{10}$$

We emphasize that we take the number of thresholds K in our model class fixed. Ideally, one would like to choose K equal to K_0, but the latter may be unknown. Kearns et al. (1997), investigate an algorithm which calculates $\hat{a}_n$ for all values of K, and a comparison of various regularization algorithms for estimating K_0. With a consistent estimator $\hat{K}$ in our model class, the asymptotics presented in this paper generally still go through. However, Kearns et al. (1997) and also later papers, e.g. Bartlett et al. (2002) show that the choice of K is very important in practice. Non-asymptotic bounds for a related problem are in Birgé (1987).

The following theorem states that $\hat{a}_n$ converges to the minimizer a^* of $L(a)$ with rate $n^{-1/3}$ and also provides its asymptotic distribution after renormalization. We assume in this theorem that $K\le K_0$. If $K=K_0$, one can show that when the minimizer a^* is unique, it is equal to a_0, i.e., then h_{a^*} is Bayes classifier. The case $K<K_0$ is illustrated at the end of this subsection.

We use the notation $\mathbb{1}(u,v>0)$ for $\mathbb{1}(u>0)\mathbb{1}(v>0)$, for scalars u and v. Likewise, we write $\mathbb{1}(u,v<0)$ for $\mathbb{1}(u<0)\mathbb{1}(v<0)$.

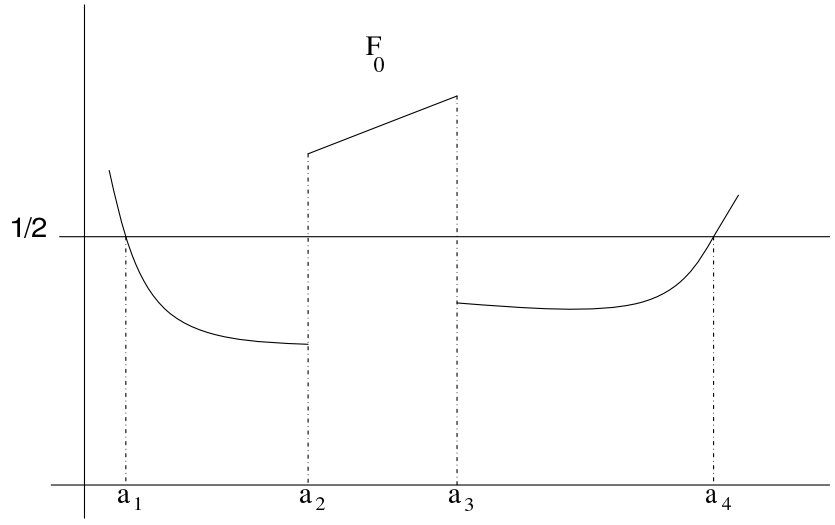

Figure 1: F_0 and the points at which $2F_0 - 1$ changes sign.

Theorem 1 *Suppose $F_0(0) < 1/2$, that*

$$a^* = (a_1^*, a_2^*, \ldots, a_K^*) := \arg\min_{a \in U_K} L(a), \tag{11}$$

is the unique minimizer of $L(a)$, that a^* is in the interior of U_K, and that $L(a)$ is a continuous function of a. Suppose that F_0 has non-zero derivative f_0 in a neighborhood of a_k^*, $k = 1, \ldots, K$. Let $g(a_k^*) > 0$, for all $k = 1, \ldots, K$, where g, the density of G, is continuous in a neighborhood of a^*. Then the process

$$\{n^{2/3}\left[L_n(a^* + tn^{-1/3}) - L_n(a^*)\right] : \ t \in \mathbb{R}^K\}$$

(where we define $L_n(a) = 0$ for $a \notin U_K$), converges in distribution to a Gaussian process $\{Z(t) : \ t \in \mathbb{R}^K\}$ with continuous sample paths, and expected value $\mathbb{E}Z(t) = t^T \mathcal{V} t/2$, where

$$\mathcal{V} = \begin{bmatrix} 2f_0(a_1^*)g(a_1^*) & 0 & \ldots & 0 \\ 0 & -2f_0(a_2^*)g(a_2^*) & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & (-1)^{K-1}2f_0(a_K^*)g(a_K^*) \end{bmatrix},$$

and covariance kernel $H = [H(s,t)]$, where

$$H(s,t) = \sum_{k=1}^{K} g(a_k^*)\left[\min(s_k, t_k)\mathbb{1}(s_k, t_k > 0) - \max(s_k, t_k)\mathbb{1}(s_k, t_k < 0)\right].$$

Moreover,

$$n^{\frac{1}{3}}(\hat{a}_n - a^*) \to^{\mathcal{L}} \arg\min Z(t).$$

The proof can be found in Section 4, where it is also noted that the diagonal elements of the matrix $\mathcal{V}$ are all positive.

Under the assumptions of Theorem 1

$$L(\hat{a}_n) - L(a^*) \approx (\hat{a}_n - a^*)^T \mathcal{V} (\hat{a}_n - a^*)/2$$

for large n. The theorem therefore also provides us the rate $n^{-2/3}$ for the convergence of the prediction error $L(\hat{a}_n)$ of the classifier $h_{\hat{a}_n}$, to the prediction error of h_{a^*}, and the asymptotic distribution of the prediction error $L(\hat{a}_n)$ after renormalization. We present this asymptotic distribution in a corollary.

Corollary 2 *Suppose the conditions of Theorem 1 are met. Then*

$$n^{\frac{2}{3}}[L(\hat{a}_n) - L(a^*)] \to^{\mathcal{L}} U^T \mathcal{V} U/2,$$

where $U = \arg\min_t Z(t)$*, and* Z *is defined in Theorem 1.*

Recall that one of the conditions in the above theorem is that L has a unique minimizer in the interior of U_K. This implies that K should not be larger than K_0. Let us consider the situation $K = 1, K_0 = 2$ and discuss when there is a unique minimizer.

Suppose $K = 1$ and

$$F_0(x) \begin{cases} < 1/2 & x \notin [a_1^0, a_2^0], \\ > 1/2 & x \in (a_1^0, a_2^0), \end{cases} \tag{12}$$

where a_1^0 and a_2^0 are unknown and $0 < a_1^0 < a_2^0 < 1$. Note that

$$\begin{aligned} L(a) &= P(Y = 1,\ h_a(X) = -1) + P(Y = -1,\ h_a(X) = 1) \\ &= \int_0^a F_0 dG + \int_a^1 (1 - F_0) dG \\ &= \int_0^a (2F_0 - 1) dG + \int_0^1 (1 - F_0) dG. \end{aligned}$$

If $\int_{a_1^0}^1 (2F_0 - 1) dG > 0$, then $a^* = a_1^0$ is the unique minimizer of L. If $\int_{a_1^0}^1 (2F_0 - 1) dG < 0$, then L has a unique minimum at 1. The minimizer is not in the open interval $(0,1)$, and Theorem 1 indeed fails. In this case, the convergence result is the same as Theorem 5 below (under its assumptions). If $\int_{a_1^0}^1 (2F_0 - 1) dG = 0$, then L has two minima at 1 and a_1^0.

2.2 Higher-Dimensional Feature Space

In this subsection, $\mathcal{X} \subset \mathbb{R}^d$ with $d > 1$, and we write for $X \in \mathcal{X}$,

$$X = (U, V),\ U \in \mathbb{R}^{d-1},\ V \in \mathbb{R}.$$

Consider given functions

$$k_a : \mathbb{R}^{d-1} \to \mathbb{R},\ \ a \in \mathcal{A},$$

and classifiers

$$h_a = 2\mathbb{1}\{C_a\} - 1,\ a \in \mathcal{A},$$

where

$$C_a := \{(u,v) : v \leq k_a(u)\}, \ \ a \in \mathcal{A}.$$

This kind of classifiers has been frequently considered and discussed in classification theory. We study the case where the parameter space is finite-dimensional, say $\mathcal{A} = \mathbb{R}^r$. A famous example is when k_a is linear in a, see for instance Hastie et al. (2001). Tsybakov and van de Geer (2005) consider this case for large r, depending on n. In contrast, we assume throughout that r is fixed.

Let again

$$a^* = \arg\min_a L(a),$$

be the minimizer of the theoretical risk $L(a)$, and

$$\hat{a}_n = \arg\min_a L_n(a)$$

be the empirical risk minimizer. We would like to know the asymptotic distribution of $\hat{a}_n$.

In this subsection, we suppose that the class $\{C_a : a \in \mathbb{R}^r\}$ is VC, i.e., that $\{k_a(u) : \ \ a \in \mathbb{R}^r\}$ is VC-subgraph. We also suppose that k_a is a regular function of the parameter $a \in \mathbb{R}^r$, i.e., the gradient

$$\frac{\partial}{\partial a} k_a(u) = k_a'(u) \tag{13}$$

of $k_a(u)$ exists for all u, and also its Hessian

$$\frac{\partial^2}{\partial a \partial a^T} k_a(u) = k_a''(u). \tag{14}$$

We will need to exchange the order of differentiation and integration of certain functions. To be able to do so, we require locally dominated integrability, which is defined as follows.

Definition 3 *Let $\{f_a : \ a \in \mathcal{A}\}$, $\mathcal{A} \subset \mathbb{R}^r$, be a collection of functions on some measurable space $(\mathcal{U}, \mu)$. It is called locally dominated integrable with respect to the measure μ and variable a if for each a there is a neighborhood I of a and a nonnegative μ-integrable function g_1 such that for all $u \in \mathcal{U}$ and $b \in I$,*

$$|f_b(u)| \leq g_1(u).$$

The probability of misclassification using the classifier h_a is

$$L(a) = P(h_a(X) \neq Y) = \int_{C_a} (1 - F_0) dG + \int_{C_a^c} F_0 dG$$

$$= \int_{C_a} (1 - 2F_0) dG + P(Y = 1).$$

Suppose that the density g of G, with respect to Lebesgue measure, exists. We use the notation

$$m(x) := (1 - 2F_0(x)) g(x). \tag{15}$$

Assumption A: Assume existence of the derivatives (13) and (14) and also of

$$m'(u,v) := \frac{\partial}{\partial v} m(u,v).$$

Assume furthermore that the functions $m(u,k_a(u))k_a'(u)$ and $\frac{\partial}{\partial a^T}[m(u,k_a(u))k_a'(u)]$ are locally dominated integrable with respect to Lebesgue measure and variable a. Also, assume that the function $\int k_a'(u)g(u,k_a(u))du$ is uniformly bounded for a in a neighborhood of a^*, and that for each u, $m'(u,k_a(u))$ and $k_a''(u)$ are continuous in a neighborhood of a^*.

Write

$$\mathcal{V}_a := \frac{\partial^2}{\partial a \partial a^T} L(a).$$

Then

$$\mathcal{V}_a = \int \Sigma_a(u) m(u,k_a(u)) du, \tag{16}$$

where

$$\Sigma_a(u) = k_a'(u)k_a'^T(u)\frac{m'(u,k_a(u))}{m(u,k_a(u))} + k_a''(u). \tag{17}$$

In the following theorem, we show that $n^{\frac{1}{3}}(\hat{a}_n - a^*)$ converges to the location of the minimum of some Gaussian process.

Theorem 4 *Suppose that L has a unique minimum at a^* and that it is continuous at a^*. Assume that for all u, the density $g(u,v)$ is continuous as a function of v at $v = k_{a^*}(u)$. Let $\mathcal{V}_a$ be continuous at a^* and $\mathcal{V} := \mathcal{V}_{a^*}$ be positive definite. Under Assumption A, we have*

$$n^{\frac{1}{3}}(\hat{a}_n - a^*) \to^{\mathcal{L}} \arg\min_{t\in\mathbb{R}^r} Z(t)$$

where $\{Z(t) : t \in \mathbb{R}^r\}$ is a Gaussian process with $\mathbb{E}Z(t) = t^T\mathcal{V}t/2$, $t \in \mathbb{R}^r$, and with continuous sample paths and covariance structure

$$Cov(Z(t),Z(s)) = \int g(u,k_{a^*}(u))\alpha^T(u,t,s)k_{a^*}'(u)du, \;\; t,s \in \mathbb{R}^r,$$

with

$$\alpha(u,t,s) = \begin{cases} -s & t^Tk_{a^*}'(u) \le s^Tk_{a^*}'(u) \le 0 \\ -t & s^Tk_{a^*}'(u) \le t^Tk_{a^*}'(u) \le 0 \\ t & 0 \le t^Tk_{a^*}'(u) \le s^Tk_{a^*}'(u) \\ s & 0 \le s^Tk_{a^*}'(u) \le t^Tk_{a^*}'(u) \\ 0 & o.w.. \end{cases} \tag{18}$$

The proof is given in Section 4.

As an example of Theorem 4, suppose $r = d$ and k_a is the linear function

$$k_a(u) := a_1u_1 + \ldots + a_{r-1}u_{r-1} + a_r.$$

It is interesting to compute the matrix $\mathcal{V}$ (see (16) and (17)) in this case. Using our notations, we have

$$k_a'(u) = [u_1 \; u_2 \; \ldots \; u_{r-1} \; 1]^T.$$

Let $f_0(u,v) := \frac{\partial}{\partial v}F_0(u,v)$ and $g'(u,v) := \frac{\partial}{\partial v}g(u,v)$ exist. Then by (15), we have

$$m'(u,v) = -2f_0(u,v)g(u,v) + (1-2F_0(u,v))g'(u,v)$$

and by (16) and (17)

$$\mathcal{V} = \left[\int u_i u_j(-2f_0(u,k_{a_0}(u))g(u,k_{a_0}(u)) + (1-2F_0(u,k_{a_0}(u)))g'(u,k_{a_0}(u)))du_1\ldots du_{r-1}\right],$$

where we define $u_r := 1$.

3. Other Rates of Convergence

In this section, we will investigate the rates that can occur if we do not assume the differentiability conditions needed for the Kim Pollard Theorem. We will restrict ourselves to the case of a one-dimensional feature space, with $\mathcal{X} = [0,1]$.

We first assume $K = 1$, and that $2F_0 - 1$ has at most one sign change (i.e. $K_0 \leq 1$). Then, we briefly discuss what happens for general K_0 and K.

3.1 The Case of One Threshold and at Most One Sign Change

Let $K = 1$ and $K_0 \leq 1$. Now, either $2F_0 - 1$ changes sign at $a^* \in (0,1)$ or there are no sign changes in $(0,1)$, i.e. $K_0 = 0$. In the first case, we assume $F_0(x) < 1/2$ near 0. In the latter case, we assume $F_0(x) < 1/2$ for all $x \in (0,1)$, and let $a^* = 1$, or $F_0(x) > 1/2$ for all $x \in (0,1)$ and let $a^* = 0$. One easily verifies that a^* is the minimizer of $L(a)$ over $a \in [0,1]$. However, if F_0 is not differentiable at a^*, Theorem 1 can not be applied. In this section, we impose the *margin condition* of Tsybakov (2004) (see also Mammen and Tsybakov (1999)). It can also be found on papers concerned with estimation of density level sets, see Polonik (1995) and Tsybakov (1997). In our context, this margin assumption is Assumption B below. Throughout, a neighborhood of a^* is some set of the form $(a^* - \delta, a^* + \delta)$, $\delta > 0$, intersected with $[0,1]$.

Assumption B: Let there exist $c > 0$ and $\varepsilon \geq 0$ such that

$$|1 - 2F_0(x)|g(x) \geq c|x - a^*|^\varepsilon, \tag{19}$$

for all x in a neighborhood of a^*.

In Section 2, we assumed differentiability of F_0 in a neighborhood of $a^* \in (0,1)$, with positive derivative f_0. This corresponds to the case $\varepsilon = 1$. We have $\varepsilon = 0$ if F_0 has a jump at a^*, and also if $a^* \in \{0,1\}$. In general, Assumption B describes how well a^* is identified: large values of ε correspond to less identifiability.

Recall now equality (6):

$$L(\hat{a}_n) - L(a^*) \leq -[\nu_n(\hat{a}_n) - \nu_n(a^*)]/\sqrt{n}. \tag{20}$$

Let $\sigma(a)$ be the standard deviation of $[\nu_n(a) - \nu_n(a^*)]$. Let

$$\psi(r) = \mathbb{E}\left(\sup_{a:\ \sigma(a)\leq r} |\nu_n(a) - \nu(a)|\right), \quad r > 0. \tag{21}$$

It will follow from the proof of Theorem 5 below, that $\psi(r) \sim r$. Moreover, the standard deviation $\sigma(a)$ behaves like $\|a-a^*\|^{1/2}$. Therefore. as we already stated in Section 2, the right hand side of (20) behaves in probability like $\|\hat{a}_n - a^*\|^{1/2}/\sqrt{n}$. From Assumption B, we see that the left hand side behaves like $\|\hat{a}_n - a^*\|^{1+\varepsilon}$. This leads to the rate $n^{-\frac{1+\varepsilon}{1+2\varepsilon}}$.

Theorem 5 *Consider the class $\mathcal{H}$ defined in (8), with $K = 1$ and $b_1 = -1$. Under Assumption B,*

$$\|\hat{a}_n - a^*\| = O_{\mathbf{P}}(n^{-\frac{1}{1+2\varepsilon}}), \quad L(\hat{a}_n) - L(a^*) = O_{\mathbf{P}}(n^{-\frac{1+\varepsilon}{1+2\varepsilon}}).$$

Proof We use the inequality (20):

$$L(\hat{a}_n) - L(a^*) \leq -[\nu_n(\hat{a}_n) - \nu_n(a^*)]/\sqrt{n}, \tag{22}$$

with $\nu_n(a) := \sqrt{n}[L_n(a) - L(a)]$. By Assumption B, we have the lower bound

$$L(\hat{a}_n) - L(a^*) \geq c\|\hat{a}_n - a^*\|^{1+\varepsilon}$$

for the left hand side of of (22).

To find an upper bound for the right hand side of (20), we apply Theorem 5.12 of van de Geer (2000). Define

$$\mathcal{G} := \{\phi : \phi(x,y) := \mathbb{1}(h_a(x) \neq y),\ a \in [0,1]\}$$

and for $\phi^*(x,y) = \mathbb{1}(h_{a^*}(x) \neq y)$ and $\delta > 0$,

$$\mathcal{G}(\delta) := \{\phi - \phi^* :\ \phi \in \mathcal{G}, \|a - a^*\| \leq \delta^2\}.$$

Let $\{H_B(u, \mathcal{G}_1(\delta), P), u > 0\}$ be the entropy with bracketing, for the metric induced by the $L_2(P)$ norm, of the class $\mathcal{G}(\delta)$. It is easy to see that for some constant c_1, and for all $\delta > 0$,

$$H_B(u, \mathcal{G}_1(\delta), P) \leq 2\log\frac{c_1\delta}{u}, \ \ \forall u \in (0, \delta).$$

Set $\delta_n = n^{-1/2}$. We may select T, C, C_0 and C_1 such that for $a := C_1T^2\delta_n$ and $R := T\delta_n$, the conditions of Theorem 5.11 of van de Geer (2000) hold. This theorem then gives that for large T and large n,

$$\mathbf{P}\left(\sup_{\|a-a^*\| \leq \delta_n^2} |\nu_n(a) - \nu_n(a^*)| \geq C_1T^2\delta_n\right) \leq C\exp(-T).$$

Now, by the peeling device, see for example van de Geer (2000), we can show that

$$\lim_{T\to\infty} \limsup_{n\to\infty} \mathbf{P}\left(\sup_{\sqrt{\|a-a^*\|} > \delta_n} \frac{|\nu_n(a) - \nu_n(a^*)|}{\sqrt{\|a-a^*\|}} \geq T\right) = 0.$$

So,

$$\frac{|\nu_n(\hat{a}_n) - \nu_n(a^*)|}{\sqrt{\|\hat{a}_n - a^*\|} \vee \delta_n} = O_{\mathbf{P}}(1). \tag{23}$$

Combining this with (22) and Assumption B yields

$$c\|\hat{a}_n - a^*\|^{1+\varepsilon} \leq (\sqrt{\|\hat{a}_n - a^*\|} + \delta_n)O_{\mathbf{P}}(1)/\sqrt{n}$$

or $\|\hat{a}_n - a^*\| = O_{\mathbf{P}}(n^{-1/(1+2\varepsilon)})$. Using (23) and (22), we can calculate $L(\hat{a}_n) - L(a^*) = O_{\mathbf{P}}(n^{-\frac{1+\varepsilon}{1+2\varepsilon}})$. ■

Theorem 5 can be refined to a non-asymptotic bound, for example in the following way. Let $\bar{\psi}$ be the smallest concave majorant of ψ defined in (21), and let $w(\cdot)$ be the smallest concave upper-bound of

$$r \mapsto \sup_{L(a)-L(a^*)\leq r^2} \sigma(a).$$

(In our situation, $w(r) \sim r^{\frac{1}{1+\varepsilon}}$.) Let r_* be the positive solution of

$$r^2 = \bar{\psi}(w(r))/\sqrt{n}.$$

Then, from Massart (2003), Koltchinskii (2003b), or Bartlett et al. (2004), we obtain that

$$\mathbf{P}\left(L(\hat{a}_n) - L(a^*) > r_*^2 + \frac{w(r_*)}{r_*^2}\frac{2x}{n} \right) \leq \mathrm{e}^{-x},\ x > 0.$$

When F_0 has a jump at a^*, we have the case $\varepsilon = 0$. Under the conditions of Theorem 5 with $\varepsilon = 0$, we derive the asymptotic distribution of the renormalized empirical risk, locally in a neighborhood of order $1/n$ of a^*, the local empirical risk. The rescaled estimator $n(\hat{a}_n - a^*)$ remains bounded in probability. However, since the local empirical risk has a limit law which has no unique minimum, $n(\hat{a}_n - a^*)$ generally does not converge in distribution. Similar results can be derived when a^* is one of the boundary points 0 or 1. For simplicity we only consider the right hand side limit. We assume that F_0 and g are right continuous.

In Theorem 6 below, convergence in distribution is to be understood in the sense given e.g. in Barbour et al. (1992).

Theorem 6 *Consider the class $\mathcal{H}$ defined in (8), with $K = 1$ and $b_1 = -1$. Assume that $a^* \in (0,1)$, $1/2 < F_0(a^*) < 1$, g and F_0 are right continuous at a^* and $g(a^*) > 0$. Let*

$$\lambda_1 := F_0(a^*)g(a^*),\ \ \lambda_2 := (1 - F_0(a^*))g(a^*).$$

Let $\mathcal{Z}_n(t) = n[L_n(a^* + t/n) - L_n(a^*)]$, $t > 0$. The process $\mathcal{Z}_n$ converges in distribution to $Z_1 - Z_2$, where Z_i is a Poisson process with intensity λ_i, $i = 1,2$, and $Z_1(t)$ and $Z_2(s)$ are independent for all $s,t > 0$.

Proof We have for $t > 0$

$$\mathcal{Z}_n(t) = \sum_{Y_i=1} \mathbb{1}(a^* \leq X_i < a^* + t/n) - \sum_{Y_i=-1} \mathbb{1}(a^* \leq X_i < a^* + t/n).$$

Define

$$I_n(t) := \sum_{Y_i=1} \mathbb{1}(a^* \leq X_i < a^* + t/n),\ \ J_n(t) := \sum_{Y_i=-1} \mathbb{1}(a^* \leq X_i < a^* + t/n). \tag{24}$$

The random variable $I_n(t)$ has a binomial distribution with parameters n and p_1, where

$$p_1 := P(Y = 1, a^* \leq X < a^* + t/n) = \int_{a^*}^{a^*+t/n} F_0 dG. \tag{25}$$

For large n, p_1 is close to $\lambda_1 t/n$. Similarly, for large n, $J_n(t)$ has binomial distribution with parameters n and $p_2 := \lambda_2 t/n$. We know that $B(n, \lambda t/n)$, for large n and small t, is approximately Poisson(λt), i.e. the total variation distance between the two distributions goes to zero as $n \to \infty$.

Note that for every $0 < t_1 < t_2 < 1$,

$$nP(Y = 1, a^* + t_1/n \le X \le a^* + t_2/n) = n \int_{a^*+t_1/n}^{a^*+t_2/n} F_0 dG \to \lambda_1(t_2 - t_1)$$

and

$$nP(Y = -1, a^* + t_1/n \le X \le a^* + t_2/n) = n \int_{a^*+t_1/n}^{a^*+t_2/n} (1 - F_0) dG \to \lambda_2(t_2 - t_1)$$

as $n \to \infty$. Now by Theorem 5.2.4, Remark 4 and Proposition A2.12 of Embrechts et al. (1997), we conclude that the whole process I_n (J_n) converges weakly to a Poisson process with intensity λ_1 (λ_2). (See also Barbour et al. (1992).) With the method of moment generating functions we can prove that the processes I_n and J_n are asymptotically independent, i.e., for any $t_1, ..., t_m, s_1, ..., s_k$,

$$E(\exp(r_1 I_n(t_1) + ... + r_m I_n(t_m) + l_1 J_n(s_1) + ... + l_k J_n(s_k)))$$

converges to

$$E(\exp(r_1 Z_1(t_1) + ... + r_m Z_1(t_m)))E(\exp(l_1 Z_2(s_1) + ... + l_k Z_2(s_k))).$$

Thus, $I_n - J_n$ converges weakly to the difference of two independent Poisson processes with intensities λ_1 and λ_2. ■

3.2 Extension to Several Thresholds and Sign Changes

Recall that K_0 is the number of sign changes of $2F_0 - 1$, and that K is the number of thresholds in the model class $\mathcal{H}$ defined in (8). Below, whenever we mention the rate $n^{-1/3}$ or n^{-1}, we mean the rate can be obtained under some conditions on F_0 and g (see Theorem 1 (where $\varepsilon = 1$), and Theorem 5 with $\varepsilon = 0$). Recall that a^0 denotes the K_0-vector of the locations of the sign changes of $2F_0 - 1$.

1. Let $K \le K_0$ and a^* is an interior point of U_K. In this case, $\hat{a}_n$ converges to a^*. The rate is $n^{-1/3}$.

2. Let $K = K_0 + 1$. Then, K_0 of the elements of $\hat{a}_n$ converge to a^0, and either $\hat{a}_{1,n}$ converges to 0 or $\hat{a}_{K,n}$ converges to 1. The rate of convergence to the interior points is $n^{-1/3}$ and the rate of convergence to the boundary point is n^{-1}.

3. Let $K > K_0 + 1$. In this case, K_0 of the elements of $\hat{a}_n$ converge to a^0 with rate $n^{-1/3}$. If $K - K_0$ is odd, one element of $\hat{a}_n$ converges to one of the boundary points 0 or 1.

4. Proof of Theorem 1 and Theorem 4

We start out with presenting the Kim Pollard Theorem (Kim and Pollard (1990)) in a general context. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent copies of a random variable ξ, with values in some space

$\mathcal{S}$. Let $\phi(\cdot,a):\mathcal{S}\to\mathbb{R}$ be a collection of functions indexed by a parameter $a\in\mathcal{A}\subset\mathbb{R}^r$. Define $L_n(a)=\sum_{i=1}^n\phi(\xi_i,a)/n$ and $L(a)=E\phi(\xi,a)$. Moreover, let

$$\nu_n(a)=\sqrt{n}[L_n(a)-L(a)],\ a\in\mathcal{A}.$$

Define

$$\mathcal{G}_R:=\{\phi(\cdot,a):\ |a_k-a_k^*|\le R,\ k=1,...,r\},\ R>0. \tag{26}$$

The envelope G_R of this class is defined as

$$G_R(\cdot)=\sup_{\phi\in\mathcal{G}_R}|\phi(\cdot)|.$$

Theorem 1.1 in Kim and Pollard (1990) requires uniform manageability of a class of functions. The definition of uniform manageability can be found in Pollard (1989) and Pollard (1990). If $\mathcal{G}$ is VC-subgraph, then a sufficient condition for the class $\mathcal{G}_R$ to be uniformly manageable is that its envelope function G_R is uniformly square integrable for R near zero.

Theorem 7 *(Kim and Pollard (1990)) Let $\{\hat{a}_n\}$ be a sequence of estimators for which*
(i) $L_n(\hat{a}_n)\le\inf_{a\in\mathcal{A}}L_n(a)+o_P(n^{-2/3})$,
(ii) $\hat{a}_n$ converges in probability to the unique a^* that minimizes $L(a)$,
(iii) a^* is an interior point of $\mathcal{A}$.
Let $\phi(\cdot,a^*)=0$ and suppose
(iv) $L(a)$ is twice differentiable with positive definite second derivative matrix $\mathcal{V}$ at a^*,
(v) $H(s,t)=\lim_{\tau\to\infty}\tau E\phi(\xi,a+s/\tau)\phi(\xi,a+t/\tau))$ exists for each s,t in R^d and

$$\lim_{\tau\to\infty}\tau E\phi(\xi,a^*+s/\tau)^2\mathbb{1}\{|\phi(\xi,a^*+s/\tau)|>\eta\tau\})=0$$

for each $\eta>0$ and s in $\mathbb{R}^r$,
(vi) $E|\phi(\xi,a)-\phi(\xi,b)|=O(\|a-b\|)$ near a^*,
(vii) the classes $\mathcal{G}_R$ in (26), for R near zero, are uniformly manageable for the envelopes G_R and satisfy $E(G_R^2)=O(R)$ as $R\to0$, and for each $\eta>0$ there is a constant C such that $E(G_R^2\mathbb{1}\{G_R>C\})<\eta R$ for R near zero.
Then the process $\{n^{2/3}[L_n(a^*+tn^{-1/3})-L_n(a^*)]:t\in\mathbb{R}^r\}$, (where we take $L_n(a)=0$ if $a\notin\mathcal{A}$}), converges in distribution to a Gaussian process $\{Z(t):\ t\in\mathbb{R}^r\}$ with continuous sample paths, expected value $\mathbb{E}Z(t)=t^T\mathcal{V}t/2$ and covariance kernel H. If Z has non-degenerate increments, then $n^{1/3}(\hat{a}_n-a^*)$ converges in distribution to the (almost surely unique) random vector that minimizes $\{Z(t):\ t\in\mathbb{R}^r\}$.

Proof of Theorem 1 We apply the Kim Pollard Theorem to the function

$$\phi(x,y,a):=\mathbb{1}(h_a(x)\neq y)-\mathbb{1}(h_{a^*}(x)\neq y),$$

Condition (i) is met by the definition of $\hat{a}_n$. To check Condition (ii), we note that, because $\{\phi(\cdot,a):\ a\in U_K\}$ is a uniformly bounded VC-subgraph class, we have the uniform law of large numbers

$$\sup_{a\in U_K}|L_n(a)-L(a)|\to0,\ \text{a.s.}.$$

Since we assume that $a^* \in U_K$ is unique and L is continuous., this implies

$$\hat{a}_n \to a^*, \text{ a.s..}$$

Condition (iii) is satisfied by assumption.

To check Condition (iv), for odd i, we have

$$\frac{\partial}{\partial a_i} P(h_a(X) \neq Y) = (2F_0(a_i) - 1)g(a_i)$$

so

$$\begin{aligned}\frac{\partial^2}{\partial a_i^2} P(h_a(X) \neq Y)\Big|_{a_i = a_i^*} &= \Big(\big[2 f_0(a_i) g(a_i) + (2F_0(a_i) - 1) g'(a_i) \big] \Big)\Big|_{a_i = a_i^*} \\ &= -2 f_0(a_i^*) g(a_i^*).\end{aligned}$$

For even i, these terms are symmetric. Thus (iv) is satisfied with

$$\mathcal{V} := \begin{bmatrix} 2 f_0(a_1^*) g(a_1^*) & 0 & \dots & 0 \\ 0 & -2 f_0(a_2^*) g(a_2^*) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & (-1)^{K-1} 2 f_0(a_K^*) g(a_K^*) \end{bmatrix}.$$

Now, a^* minimizes $L(a)$ for a in the interior of U_K, so $2F_0 - 1$ changes sign from negative to positive at a_k^* for odd k, and it changes sign from positive to negative at a_k^* for even k. Hence $f_0(a_k^*) > 0$ for odd k and $f_0(a_k^*) < 0$ for even k and therefore $\mathcal{V}$ is positive definite.

Next, we study the existence of the covariance kernel H, required in Condition (v). Consider $t, s \in \mathbb{R}$ and large $\tau > 0$ so that $a^* + t/\tau, a^* + s/\tau \in U_K$. First we note that the product of the brackets is the same for $Y = 1$ and for $Y = -1$. For $a_1 < a_2, b_1 < b_2, a_1^* < a_2^*$, we have

$$\begin{aligned}&\Big[\mathbb{1}(a_1^* \le x < a_2^*) - \mathbb{1}(a_1 \le x < a_2) \Big] \Big[\mathbb{1}(a_1^* \le x < a_2^*) - \mathbb{1}(b_1 \le x < b_2) \Big] \\ &= \Big[\mathbb{1}(x \ge a_1) - \mathbb{1}(x \ge a_2) - \mathbb{1}(x \ge a_1^*) + \mathbb{1}(x \ge a_2^*) \Big] \\ &\quad \times \Big[\mathbb{1}(x \ge b_1) - \mathbb{1}(x \ge b_2) - \mathbb{1}(x \ge a_1^*) + \mathbb{1}(x \ge a_2^*) \Big] \\ &= A(x) - B(x) - C(x) + D(x),\end{aligned}$$

where

$$\begin{aligned}A(x) &:= (\mathbb{1}(x \ge a_1) - \mathbb{1}(x \ge a_1^*))(\mathbb{1}(x \ge b_1) - \mathbb{1}(x \ge a_1^*)) \\ &= \mathbb{1}[\min(a_1, a_1^*), \max(a_1, a_1^*))\mathbb{1}[\min(b_1, a_1^*), \max(b_1, a_1^*)) \\ &= \mathbb{1}[a_1^*, \min(a_1, b_1))\mathbb{1}(a_1^* < \min(a_1, b_1)) + \mathbb{1}[\max(a_1, b_1), a_1^*)\mathbb{1}(a_1^* > \max(a_1, b_1)), \\ D(x) &:= (\mathbb{1}(x \ge a_2) - \mathbb{1}(x \ge a_2^*))(\mathbb{1}(x \ge b_2) - \mathbb{1}(x \ge a_2^*)) \\ &= \mathbb{1}[\min(a_2, a_2^*), \max(a_2, a_2^*))\mathbb{1}[\min(b_2, a_2^*), \max(b_2, a_2^*)) \\ &= \mathbb{1}[a_2^*, \min(a_2, b_2))\mathbb{1}(a_2^* < \min(a_2, b_2)) + \mathbb{1}[\max(a_2, b_2), a_2^*)\mathbb{1}(a_2^* > \max(a_2, b_2)),\end{aligned}$$

$$B(x) := (\mathbb{1}(x \geq a_1) - \mathbb{1}(x \geq a_1^*))(\mathbb{1}(x \geq b_2) - \mathbb{1}(x \geq a_2^*)),$$

and

$$C(x) := (\mathbb{1}(x \geq a_2) - \mathbb{1}(x \geq a_2^*))(\mathbb{1}(x \geq b_1) - \mathbb{1}(x \geq a_1^*)).$$

Assume that $a_1 = a_1^* + s_1/\tau, a_2 = a_2^* + s_2/\tau, b_1 = a_1^* + t_1/\tau, b_2 = a_2^* + t_2/\tau$. When τ tends to infinity, we have $\int BdG = \int CdG = 0$. Moreover,

$$\begin{aligned}
&\int (A+D)dG \\
&= \left[\mathbb{1}(0 < s_1, t_1) \int_{a_1^*}^{a_1^* + \min(s_1,t_1)/\tau} dG + \mathbb{1}(0 > s_1, t_1) \int_{a_1^* + \max(s_1,t_1)/\tau}^{a_1^*} dG \right. \\
&\left. + \mathbb{1}(0 < s_2, t_2) \int_{a_2^*}^{a_2^* + \min(s_2,t_2)/\tau} dG + \mathbb{1}(0 > s_2, t_2) \int_{a_2^* + \max(s_2,t_2)/\tau}^{a_2^*} dG \right] \\
&= \min(s_1,t_1) g(a_1^*) \mathbb{1}(0 < s_1, t_1) - \max(s_1,t_1) g(a_1^*) \mathbb{1}(0 > s_1, t_1) \\
&+ \min(s_2,t_2) g(a_2^*) \mathbb{1}(0 < s_2, t_2) - \max(s_2,t_2) g(a_2^*) \mathbb{1}(0 > s_2, t_2). \qquad (27)
\end{aligned}$$

Let m be the integer part of $(K+1)/2$. Now, we obtain

$$\begin{aligned}
&E\phi(X,Y,a^* + s/\tau)\phi(X,Y,a^* + t/\tau) \\
&= E\left[\mathbb{1}\left(X \in \cup_{i=1}^m [a_{2i-1}^* + s_{2i-1}/\tau, a_{2i}^* + s_{2i}/\tau)\right) - \mathbb{1}\left(X \in \cup_{i=1}^m [a_{2i-1}^*, a_{2i}^*)\right)\right] \\
&\times \left[\mathbb{1}\left(X \in \cup_{i=1}^m [a_{2i-1}^* + t_{2i-1}/\tau, a_{2i}^* + t_{2i}/\tau)\right) - \mathbb{1}\left(X \in \cup_{i=1}^m [a_{2i-1}^*, a_{2i}^*)\right)\right] \\
&= \sum_{k=1}^K E\left[\mathbb{1}\left(X \in [a_k^*, a_k^* + \min(s_k,t_k))\right)\mathbb{1}(0 < s_k, t_k)\right. \\
&\left. - \mathbb{1}\left(X \in [a_k^* + \max(s_k,t_k), a_k^*)\right)\mathbb{1}(0 > s_k, t_k)\right] \qquad (28)
\end{aligned}$$

(for large τ). Finally, by (27) and (28), the limit of $\tau E\phi(X,Y,a^* + s/\tau)\phi(X,Y,a^* + t/\tau)$ as $\tau \to \infty$ becomes

$$\begin{aligned}
H(s,t) = \sum_{k=1}^K \Big[&\min(s_k,t_k) g(a_k^*) \mathbb{1}(0 < s_k, t_k) \\
&- \max(s_k,t_k) g(a_k^*) \mathbb{1}(0 > s_k, t_k)\Big].
\end{aligned}$$

So, the first part of condition (v) is satisfied. As for the second part of condition (v), for any ε and $\tau > 1/\varepsilon$, and $t \in \mathbb{R}$, we have

$$E\left[\mathbb{1}^2(h_{a^* + \frac{t}{\tau}}(X) \neq Y)\mathbb{1}(\mathbb{1}(h_{a^* + \frac{t}{r}}(X) \neq Y) > \tau\varepsilon)\right] = 0.$$

To show that Condition (vi) is satisfied, we note that for any $a, b \in U_K$,

$$E\left[\left|\mathbb{1}(h_a(X) \neq Y) - \mathbb{1}(h_b(X) \neq Y)\right|\right] \leq \sum_{k=1}^K E\left[\mathbb{1}\left(X \in [\min(a_k,b_k), \max(a_k,b_k))\right)\right]$$

$$\leq \sum_{k=1}^{K} |a_k - b_k| g(\xi_k)$$

for some $\xi_k \in [\min(a_k, b_k), \max(a_k, b_k)]$. Hence

$$E\left(\left| \mathbb{1}(h_a(X) \neq Y) - \mathbb{1}(h_b(X) \neq Y) \right| \right) = O(\|a - b\|),$$

for a and b near a^*.

Now we calculate an upper bound for the envelope function. Fix $(x,y) \in \mathcal{X} \times \{-1,1\}$. To maximize the function $\phi(x,y,a) = \mathbb{1}(h_a(x) \neq y) - \mathbb{1}(h_{a^*}(x) \neq y)|$, note that for $y = 1$, this function is increasing in a_k's for even k and decreasing in a_k's for odd k. To simplify, assume K is odd. Over $\mathcal{G}_R$, $\phi(x,y,a)$ is maximized when

$$a_1 = a_1^* - R, \;\; a_2 = a_2^* + R, \;\; a_3 = a_3^* - R, \;\; \ldots, \;\; a_K = a_K^* - R. \tag{29}$$

For $y = -1$, it is maximized when

$$a_1 = a_1^* + R, \;\; a_2 = a_2^* - R, \;\; a_3 = a_3^* + R, \;\; \ldots, \;\; a_K = a_K^* + R. \tag{30}$$

Similarly, $\mathbb{1}(h_{a^*}(x) \neq y) - \mathbb{1}(h_a(x) \neq y)$ is maximized for $y = 1$, in case (30) and for $y = -1$, it is maximized in case (29). So, the maximum of $|\phi(x,y,a))|$ is the maximum of

$$\mathbb{1}\left(x \in [a_1^* - R, a_1^*] \cup [a_2^*, a_2^* + R] \cup \ldots \cup [a_K^* - R, a_K^*] \right)$$

and

$$\mathbb{1}\left(x \in [a_1^*, a_1^* + R] \cup [a_2^* - R, a_2^*] \cup \ldots \cup [a_K^*, a_K^* + R] \right).$$

So the envelope G_R of $\mathcal{G}_R$ satisfies

$$G_R \leq G'_R$$

where

$$G'_R = \mathbb{1}\left(x \in \cup_{k=1}^{K} [a_k^* - R, a_k^* + R] \right).$$

Now, note that

$$E(G_R'^2) \leq \sum_{k=1}^{K} P(a_k^* - R \leq X \leq a_k^* + R)$$

and

$$\frac{P(a_k^* - R \leq X \leq a_k^* + R)}{R} = \frac{2Rg(a'_k)}{R} < R^*, \;\; \exists R^* < \infty,$$

for some $a'_k \in (a_k^* - R, a_k^* + R)$, when R is close to zero. We thus have $E(G_R'^2) = O(R)$. Since G'_R is bounded by one, it is also easy to see that G'_R is uniformly square integrable for R close to zero. Finally, since $\mathcal{G}$ is VC-subgraph, we conclude that $\mathcal{G}_R$ is uniformly manageable for the envelope G_R. ■

Proof of Theorem 4

Checking the Conditions (i)-(vii) of the Kim Pollard Theorem is very similar to the proof of Theorem 1. We consider again $\phi(x,y,a) = P(h_a(X) \neq Y) - P(h_{a^*}(X) \neq Y)$. Condition (i) is clearly true. Because the class $\{C_a : a \in \mathbb{R}^r\}$ is VC and L is continuous at a^*, we know by the same argument as in the proof of Theorem 1 that $\hat{a}_n \to a^*$ almost surely. So, Condition (ii) is met. Condition (iii) is met because $\mathbb{R}^r$ is open. The function L is twice differentiable with positive definite second derivative matrix $\mathcal{V}$ at a^*. So, (iv) is satisfied. To show that (v) is satisfied, we consider the covariance structure of $\phi(X,Y,a)$. Now,

$$Cov\big(\phi(X,Y,a)],\phi(X,Y,\tilde{a})\big) = I - II,$$

where

$$I := E\left[\phi(X,Y,a)\phi(X,Y,\tilde{a})\right]$$

and

$$II := \left[L(a) - L(a^*)\right]\left[L(\tilde{a}) - L(a^*)\right] = O(\tau^{-4}),$$

for $\|a - a^*\| = O(1/\tau)$ and $\|\tilde{a} - a^*\| = O(1/\tau)$. As for I, write $C = C_a$, $\tilde{C} = C_{\tilde{a}}$, and $C_* = C_{a^*}$, then

$$\begin{aligned} I = {} & P(Y=1, X \in C^c \cap \tilde{C}^c) - P(Y=1, X \in C^c \cap C_0^c) \\ & - P(Y=1, X \in C_0^c \cap \tilde{C}^c) + P(Y=1, X \in C_*^c) \\ & + P(Y=-1, X \in C \cap \tilde{C}) - P(Y=-1, X \in C \cap C_*) \\ & - P(Y=-1, X \in C_0 \cap \tilde{C}) + P(Y=-1, X \in C_*). \end{aligned}$$

It is easy to see that

$$\begin{aligned} I = {} & \int \Bigg[\int_{v \geq k_a(u), v \geq k_{\tilde{a}}(u)} F_0(u,v) - \int_{v \geq k_a(u), v \geq k_{a^*}(u)} F_0(u,v) \\ & - \int_{v \geq k_{a^*}(u), v \geq k_{\tilde{a}}(u)} F_0(u,v) + \int_{v \geq k_{a^*}(u)} F_0(u,v) \\ & + \int_{v < k_a(u), v < k_{\tilde{a}}(u)} (1 - F_0(u,v)) - \int_{v < k_a(u), v < k_{a^*}(u)} (1 - F_0(u,v)) \\ & - \int_{v < k_{a^*}(u), v < k_{\tilde{a}}(u)} (1 - F_0(u,v)) + \int_{v < k_{a^*}(u),} (1 - F_0(u,v)) \Bigg] g(u,v) du dv. \\ = {} & \int_{k_a(u) \leq k_{\tilde{a}}(u) \leq k_{a^*}(u)} \int_{k_{\tilde{a}}(u)}^{k_{a_0}(u)} g(u,v) dv du + \int_{k_{\tilde{a}}(u) \leq k_a(u) \leq k_{a^*}(u)} \int_{k_a(u)}^{k_{a^*}(u)} g(u,v) dv du \\ & + \int_{k_{a^*}(u) \leq k_a(u) \leq k_{\tilde{a}}(u)} \int_{k_{a^*}(u)}^{k_a(u)} g(u,v) dv du + \int_{k_{a^*}(u) \leq k_{\tilde{a}}(u) \leq k_a(u)} \int_{k_{a^*}(u)}^{k_{\tilde{a}}(u)} g(u,v) dv du. \end{aligned}$$

For each $s,t \in \mathbb{R}^r$, and for sequences $\{\overline{a}(\tau)\}$ and $\{\underline{a}(\tau)\}$ with

$$\lim_{\tau \to \infty} \overline{a}(\tau) = \lim_{\tau \to \infty} \underline{a}(\tau) = a^*,$$

we have

$$\lim_{\tau \to \infty} \tau \int_{k_{a^*+s/\tau}(u) \leq k_{a^*+t/\tau}(u) \leq k_{a^*}(u)} \int_{k_{a^*+t/\tau}(u)}^{k_{a^*}(u)} g(u,v) dv du$$

$$= \lim_{\tau\to\infty} \tau \int_{k_{a^*+s/\tau}(u)\le k_{a^*+t/\tau}(u)\le k_{a^*}(u)} \Big(k_{a^*}(u) - k_{a^*+t/\tau}(u) \Big) g(u, k_{\overline{a}(\tau)}(u)) du$$

$$= \lim_{\tau\to\infty} \tau \int_{k_{a^*+s/\tau}(u)\le k_{a^*+t/\tau}(u)\le k_{a^*}(u)} \big(-t^T/\tau\big) k'_{\underline{a}(\tau)}(u) g(u, k_{\overline{a}(\tau)}(u)) du. \tag{31}$$

When $\tau \to \infty$, the conditions $k_{a_0+s/\tau}(u) \le k_{a^*+t/\tau}(u)$ and $k_{a^*+t/\tau}(u) \le k_{a^*}(u)$ becomes $(-s^T + t^T)k'_{a^*}(u) \ge 0$ and $-t^T k'_{a^*}(u) \ge 0$, respectively. So the limit in (31) becomes

$$-\int_{0\ge t^T k'_{a^*}(u)\ge s^T k'_{a^*}(u)} t^T k'_{a^*}(u) g(u, k_{a^*}(u)) du.$$

Hence, have shown that

$$\lim_{\tau\to\infty} \tau Cov\big(\phi(X,Y,a^*+s/\tau), \phi(X,Y,a^*+t/\tau)\big)$$

$$= \int \alpha^T(u,t,s) k'_{a^*}(u) g(u, k_{a^*}(u)) du,$$

where α is defined in (18). The second part of Condition (v) is true because the functions $\phi(\cdot, a)$ are bounded. We conclude that Condition (v) is satisfied.

Conditions (vi) and (vii) are verified in the same way as in the proof of Theorem 1. ■

References

A. Barbour, L. Holst, and S. Janson. *Poisson Approximation*. Oxford Studies in Probability. Clarendon Press, Oxford, 1992.

P. L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

P. L. Bartlett, S. Mendelson, and P. Philipps. Empirical risk minimization. *Proceedings of COLT 2004, Springer Verlag*, 2004.

L. Birgé. Estimating a density under order restrictions. *Ann. Statist.*, 15(3):995–1012, 1987.

S. Boucheron, G. Bousquet, and G. Lugosi. Theory of classification: some recent advances. *To appear in ESAIM Probability and statistics*, 2005.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA, 1984.

P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling extremal events. For insurance and finance*, volume 33 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1997.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning. Data mining, Inference, and Prediction*. Springer Series in Statistics. Springer-Verlag, New York, 2001.

M. Kearns, Y. Mansour, A. Ng, and D. Ron. An experimental and theoretical comparison of model selection methods. *Machine Learning*, 27:7–50, 1997.

J. Kim and D. Pollard. Cube root asymptotics. *Ann. Statist.*, 18(1):191–219, 1990.

V. Koltchinskii. Bounds on margin distributions in learning problems. *Ann. Inst. H. Poincaré Probab. Statist.*, 39(6):943–978, 2003a.

V. Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *Preprint*, 2003b.

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statist.*, 30(1):1–50, 2002.

G. Lugosi and A. B. Nobel. Adaptive model selection using empirical complexities. *Ann. Statist.*, 27(6):1830–1864, 1999.

G. Lugosi and N. Vayatis. On the bayes-risk consistency of regularized boosting methods. *Ann. Statist.*, 32, 2004.

G. Lugosi and M. Wegkamp. Complexity regularization via localized random penalties. *Ann. Statist.*, 32(4):1679–1697, 2004.

E. Mammen and A. B. Tsybakov. Smooth discriminant analysis. *Ann. Statist.*, 27(6):1808–1829, 1999.

P. Massart. *Concentration inequalities and model selection.* Ecole d'Eté Probabilité de Saint Flour XXXIII. Springer Verlag, 2003.

L. Mohammadi. *Estimation of thresholds in classification.* PhD thesis, University of Leiden, 2004.

L. Mohammadi and S. A. van de Geer. On threshold-based classification rules. *Institute of Mathematical Statistics, Lecture Notes Monograph Series, Mathematical Statistics and Applications: Festschrift for Constance van Eeden*, 42:261–280, 2003.

D. Pollard. Asymptotics via empirical processes. *Statist. Sci.*, 4(4):341–366, 1989. With comments and a rejoinder by the author.

D. Pollard. *Empirical Processes: Theory and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, 2. Institute of Mathematical Statistics, Hayward, CA, 1990.

W. Polonik. Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *Ann. Statist.*, 23(3):855–881, 1995.

A. B. Tsybakov. On nonparametric estimation of density level sets. *Ann. Statist.*, 25(3):948–969, 1997.

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1): 135–166, 2004.

A. B. Tsybakov and S. A. van de Geer. Square root penalty: adaptation to the margin in classification and in edge estimation. *Ann. Statist.*, 33, 2005.

S. A. van de Geer. *Empirical Processes in M-Estimation.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2000.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

V. N. Vapnik. *Statistical Learning Theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. Springer-Verlag, New York, second edition, 2000.