Optimal third root asymptotic bounds in the statistical estimation of thresholds

Franz Merkl and Leila Mohammadi

Abstract

This paper is concerned with estimating the intersection point of two densities, given a sample of both of the densities. This problem arises in classification theory. The main results provide lower bounds for the probability of the estimation errors to be large on a scale determined by the inverse cube root of the sample size. The key to the proof is an entropy estimate. The lower bounds are based on bounds for general estimators, which are applicable in other contexts, as well. Furthermore, we introduce a class of optimal estimators, whose errors asymptotically meet the border permitted by the lower bounds.[1]

1 Introduction

1.1 Motivation and origin of the problem

In this article, we derive lower bounds for the probability of large errors of some estimators to occur. Let $\mathcal{P}$ be a class of probability measures on a measurable space $(\Omega, \mathcal{A})$, and $a : \mathcal{P} \to \mathbb{R}$ be a parameter. Consider an i.i.d. random sample $Z_1, ..., Z_n$ from $P \in \mathcal{P}$ and an estimator $\hat{a}_n(Z_1, ..., Z_n)$, $\hat{a}_n : \Omega^n \to \mathbb{R}$, for a. We are interested in the asymptotic behaviour of $\hat{a}_n$ as $n \to \infty$.

In the theory of empirical processes, one usually considers a deterministic loss function whose minimizer over a particular class is equal to or close to the parameter. In order to get an estimator, one then minimizes the empirical counterpart of the loss function, i.e., one replaces P by the empirical probability measure $\hat{P}_n$ in the minimization procedure.

It is natural to assume that the loss function is convex. For example, one may take the squared error loss function. Under some technical assumptions, if the loss is differentiable with respect to the parameter, the estimators converge to the parameter with the rate $|\hat{a}_n - a(P)| = O_P(n^{-1/2})$ as the sample size n grows to ∞; see van de Geer [vdG00] and van der Vaart and Wellner [vdVW96]. If the loss is not convex, but a parametric form for $\mathcal{P}$ is assumed, one may obtain a similar rate $O_P(n^{-1/2})$; see Mohammadi and van de Geer [MvdG03], sections 2.1 and 2.2. In these cases, one usually gets normal asymptotic distributions of the estimators, using the central limit theorem. When the dimension of the parameter space is bigger than one, higher moments of the loss function may be needed to get the rate $O_P(n^{-1/2})$.

Kim and Pollard [KP90] establish a new functional central limit theorem for empirical processes. They describe an interesting class of asymptotic problems where the estimators

[1] *2000 Mathematics Subject Classification:* Primary 62G05, secondary 62G20.

Key words: Classification theory, nonparametric models, estimation error, entropy bounds, threshold-based classifiers.

converge at a rate different from $O_P(n^{-1/2})$ to limit distributions. The main point in these problems is that the loss function need not be convex; it may not even be continuous.

An important non-continuous loss functions, frequently used in the theory of classification, is the indicator loss function. Let us first describe a general view of classification (or learning theory). We formulate the simplest case, which is a two-class problem. Assume that we have two distributions (on a result space $\mathcal{X}$), labeled by $Y = 1$ and $Y = -1$. The values of Y are called "labels" or "natures". Take an observation X from a mixture of the two distributions. It is sometimes called a "feature". The problem is to predict the unknown nature Y of a feature X. Suppose we have n i.i.d. copies (X_i, Y_i), $i = 1, \dots, n$, of a realisation (X, Y). A classifier h is a measurable function $h : \mathcal{X} \to \{\pm 1\}$. (Here, we do not consider more general $[-1, 1]$-valued classifiers.) The realisation (X, Y) is called misclassified by the classifier h if $h(X) \neq Y$. We take the deterministic loss function $(x, y) \mapsto 1_{\{h(x)\neq y\}}$. For $\mathcal{X} \subseteq \mathbb{R}$ and features X with a continuous distribution (at least close to a point), Mohammadi and van de Geer [MvdG03] apply this setup to the case where the classifier h is varied over the class $\mathcal{H} = (h_a)_{a\in\mathbb{R}}$, where

$$h_a(x) = \begin{cases} 1 & \text{for } x \geq a, \\ -1 & \text{for } x < a. \end{cases} \tag{1.1}$$

Let

$$f_P(x, y) = f_P^+(x)1_{\{y=1\}} + f_P^-(x)1_{\{y=-1\}}, \quad (x, y) \in \Omega = \mathcal{X} \times \{\pm 1\}, \tag{1.2}$$

denote the joint density of (X, Y) (with respect to some reference measure $\lambda\otimes$(counting measure).

If there is a unique point $a(P)$ at which $f_P^+ - f_P^-$ changes its sign, and if this sign change is from "$-$" to "$+$", it suffices to restrict the classifier h to the class $\mathcal{H}$ to minimize the risk $P[h(X) \neq Y]$, since one has

$$\inf_{\text{classifiers}\, h} P[h(X) \neq Y] = \inf_{a\in\mathbb{R}} P[h_a(X) \neq Y] = P[h_{a(P)}(X) \neq Y] \tag{1.3}$$

in this case. Taking a classifier h which minimizes $P[h(X) \neq Y]$ is called "Bayes rule":

$$a(P) = \arg\min_{a\in\mathbb{R}} P[h_a(X) \neq Y]. \tag{1.4}$$

A natural choice for an estimator of a is the threshold $\hat{a}_n$ that minimizes the classification error in the sample, i.e.,

$$\hat{a}_n = \arg\min_{a\in\mathbb{R}} \hat{P}_n[h_a(X) \neq Y] = \arg\min_{a\in\mathbb{R}} \sum_{i=1}^{n} 1_{\{h_a(X_i)\neq Y_i\}}, \tag{1.5}$$

where

$$\hat{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_i, Y_i)} \tag{1.6}$$

denotes the empirical distribution of the sample. (Strictly speaking, the "arg min" in (1.5) is not unique, but one may take any (measurable) choice.) In the theory of classification, this is called empirical risk minimization. Mohammadi and van de Geer [MvdG03] invoke the theory of Kim and Pollard [KP90] to get the rate $O_P(n^{-1/3})$ of $\hat{a}_n$ under some conditions. Under monotonicity assumptions, it is shown that $\hat{a}_n$ is a nonparametric maximum likelihood estimator, and that $n^{1/3}(\hat{a}_n - a(P))$ converges in distribution to a continuous random variable. For more background information about empirical risk minimization in classification, see for instance [FS98], [SS98], [SFBL98], [BB02], and [MBG02].

1.2 Statement of the problem and results

In this article, we address the following question: Is there any sequence of estimators $(\hat{a}_n : \Omega^n \to \mathbb{R})_{n\in\mathbb{N}}$ which converges to $a(P)$ with a rate faster than $O_P(n^{-1/3})$? In a technical sense to be specified below, the answer is "no".

Let us introduce the precise class $\mathcal{P}$ of probability measures that we consider. We assume that the features X take values in the unit interval $\mathcal{X} = [0,1]$.

Let $\tilde{\mathcal{P}}$ denote the set of all probability distributions

$$P = f_P[\lambda_{[0,1]} \otimes (\text{counting measure})] \tag{1.7}$$

on $\Omega = [0,1] \times \{\pm 1\}$ (with the Borel σ algebra) with $f_P \in C^1([0,1] \times \{\pm 1\})$. Here, $\lambda_{[0,1]}$ denotes the Lebesgue measure on $[0,1]$. (This particular choice of the reference measure – at least locally – plays a role in some technical estimates; e.g. in the basic entropy bound (4.14), below.)

We endow $\tilde{\mathcal{P}}$ with the metric d, given by

$$d(P,Q) = \|f_P - f_Q\|_\infty + \|\partial_1 f_P - \partial_1 f_Q\|_\infty, \tag{1.8}$$

where ∂_1 denotes the derivative with respect to the first argument.

Let $\mathcal{P} \subseteq \tilde{\mathcal{P}}$ denote the set of all $P \in \tilde{\mathcal{P}}$, such that $f_P^+ := f_P(\cdot, 1)$ and $f_P^- := f_P(\cdot, -1)$ have a unique intersection point $a(P)$, and this intersection point is contained in the open interval $(0,1)$, and the intersection is transversal with a specified orientation:

$$f_P^+(a(P)) = f_P^-(a(P)), \tag{1.9}$$

$$(f_P^+)'(a(P)) > (f_P^-)'(a(P)). \tag{1.10}$$

We endow $\mathcal{P}$ with the topology induced by the metric d.

Remark. For our results, it is essential to have at least some control on the derivative of f_P, which is reflected by the choice (1.8) of the metric d. In fact, the results heavily depend on the regularity assumptions that one supposes for the densities f_P, at least locally, i.e. close to the parameter $a(P)$.

We prove two versions for lower bounds of estimation errors. The first version considers the estimation error on the "critical scale" const $\cdot n^{-1/3}$, uniformly over (small) open subsets of $\mathcal{P}$:

Theorem 1.1 *Let $\mathcal{U} \subseteq \mathcal{P}$ be a nonempty open set. Then there is $c_1 = c_1(\mathcal{U}) > 0$, such that for every $\delta \in (0, 1/4]$ and for every sequence of estimators $\hat{a}_n : \Omega^n \to \mathbb{R}$, $n \in \mathbb{N}$, one has*

$$\liminf_{n\to\infty} \sup_{P\in\mathcal{U}} P^n \left[n^{1/3} |\hat{a}_n - a(P)| > T \right] \geq \delta, \tag{1.11}$$

where

$$T = T(\mathcal{U}, \delta) = c_1 |\log(11\delta)|^{1/3}. \tag{1.12}$$

Unlike Theorem 1.1, the following theorem considers the asymptotics of the estimation error "point-wise", i.e. it takes the limit as $n \to \infty$ before taking a supremum over open sets $\mathcal{U} \subseteq \mathcal{P}$.

Theorem 1.2 *Let $(\beta_n)_{n\in\mathbb{N}}$ be a sequence of positive numbers with*

$$n^{-1/3}\beta_n \overset{n\to\infty}{\longrightarrow} \infty. \tag{1.13}$$

Then for all non-empty open sets $\mathcal{U} \subseteq \mathcal{P}$ and for all sequences of estimators $\hat{a}_n : \Omega^n \to \mathbb{R}$, $n \in \mathbb{N}$, one has

$$\sup_{P\in\mathcal{U}} \limsup_{n\to\infty} P^n[\beta_n|\hat{a}_n - a(P)| > 1] \geq 1/4. \tag{1.14}$$

The proof of this theorem uses Baire's theorem. A related argument, concerning the equicontinuity and the consistency of substitution estimators with values in a metric space, is presented in [PvZ96].

The "$\limsup_{n\to\infty}$" in (1.14) cannot be replaced by "$\liminf_{n\to\infty}$" for a trivial reason; a simple counter-example is shown at the end of section 5, below.

In contrast to Theorem 1.1, Theorem 1.2 does not consider the critical scale const $\cdot n^{-1/3}$. Indeed, its claim (1.14) breaks down on this critical scale. This is the content of the following theorem:

Theorem 1.3 *There is a family of estimators $(\hat{a}_{n,L} : \Omega^n \to \mathbb{R})_{n\in\mathbb{N}, L>0}$ with the following property: For all $P \in \mathcal{P}$, there is a neighbourhood $\mathcal{N} \subseteq \mathcal{P}$ of P, such that for all $T > 0$ one has*

$$\lim_{L\to\infty} \sup_{Q\in\mathcal{N}} \limsup_{n\to\infty} Q^n[n^{1/3}|\hat{a}_{n,L} - a(Q)| > T] = 0. \tag{1.15}$$

Such estimators $\hat{a}_{n,L}$ are explicitly described in Section 5, below. Speaking very roughly, one estimates the density f_P in a certain neighbourhood of size $Ln^{-1/3}$ of a first approximation of $a(Q)$, using regression lines.

Organisation of this paper. Let us explain how the rest of this paper is organized. In Section 2, we collect some fundamental entropy estimates. Section 3 shows "universal", general counterparts of Theorems 1.1 and 1.2, without assuming the specific form of our model $(\Omega, \mathcal{A}, \mathcal{P})$. We expect these lemmas to be useful for other examples, too. One key idea is the use of Baire's theorem to show that the set of P's with estimation errors being asymptotically large on a given scale is of second Baire category. In Section 4, Theorems 1.1 and 1.2 are proven. The proofs are based on a bound for relative entropies for slightly perturbed densities, described in Lemma 4.1, below. Finally, in Section 5, Theorem 1.3 is proven, by constructing the estimators $\hat{a}_{n,L}$ in a two-step procedure.

2 Preliminaries

In this section, we review some standard estimates to compare probabilities with respect to different measures, based on bounds of the relative entropy. Alternatively (and more or less equivalently), one could use bounds for the Hellinger distance instead of the relative entropy, but we do not follow this alternative approach here. Here, we need not assume any specific form of the model $(\Omega, \mathcal{A}, \mathcal{P})$; we take an arbitrary parameter $a : \mathcal{P} \to \mathbb{R}$, and $(\hat{a}_n : \Omega^n \to \mathbb{R})_{n\in\mathbb{N}}$ denotes any sequence of estimators.

Let $H(P,Q) = E_P[\log \frac{dP}{dQ}]$ denote the relative entropy for $P, Q \in \mathcal{P}$, whenever it is well defined.

Lemma 2.1 *Let P, Q be probability measures with $H(P,Q) < \infty$. For every random variable X with $0 \le X \le 1$, one has*

$$E_Q[X] \ge e^{-2H(P,Q)-1}\left(E_P[X] - \frac{1}{2}\right). \tag{2.1}$$

Proof. For $x \ge 0$, set $\psi(x) = x\log x - x + 1$. Note that $\psi \ge 0$. We set $N = e^{2H(P,Q)+1} \ge e$ and $A = \{dP/dQ > N\}$. Using $\psi \ge 0$, $\psi(1) = 0$, and the convexity of ψ, one sees

$$\psi(x) \ge \frac{\psi(N)}{N} 1_{\{x>N\}} x \tag{2.2}$$

for all $x \ge 0$, and thus

$$H(P,Q) = E_Q\left[\psi\left(\frac{dP}{dQ}\right)\right] \ge \frac{\psi(N)}{N} E_Q\left[1_A \frac{dP}{dQ}\right] = \frac{\psi(N)}{N} P[A]. \tag{2.3}$$

We conclude, using $\psi(N)/N = \log(N/e) + 1/N \ge 2H(P,Q)$:

$$\begin{aligned} E_Q[X] &\ge E_Q[X 1_{A^c}] \ge \frac{1}{N} E_P[X 1_{A^c}] \\ &\ge \frac{1}{N}(E_P[X] - P[A]) \\ &\ge \frac{1}{N}\left(E_P[X] - \frac{N}{\psi(N)} H(P,Q)\right) \\ &\ge \frac{1}{N}\left(E_P[X] - \frac{1}{2}\right), \end{aligned} \tag{2.4}$$

which is the claim (2.1).

□

Remark. The “2” in the exponent of (2.1) could be replaced by any fixed number > 1, if one replaced the $\frac{1}{2}$ in (2.1) by a different constant. This would only change the constants in our main theorems.

Lemma 2.2 *Let $\chi : \mathbb{R} \to [0,1]$ be a measurable function with $\chi(x) = 1$ for $|x| \le 1$ and $\chi(x) = 0$ for $|x| \ge 2$. Take $n \in \mathbb{N}$, $\beta_n > 0$, $P_n, Q_n \in \mathcal{P}$, and $\delta \in (0, 1/4]$. If*

$$nH(P_n, Q_n) \le \frac{1}{2} \log \frac{1}{11\delta} \tag{2.5}$$

and

$$\beta_n |a(P_n) - a(Q_n)| > 4 \tag{2.6}$$

hold, then at least one of the following two bounds,

$$E_{P_n^n}[\chi(\beta_n(\hat{a}_n - a(P_n)))] < 1 - \delta \tag{2.7}$$

or

$$E_{Q_n^n}[\chi(\beta_n(\hat{a}_n - a(Q_n)))] < 1 - \delta, \tag{2.8}$$

is valid.

Proof. (Indirectly.) Assume that both, (2.7) and (2.8), fail to hold. Using Lemma 2.1, we get

$$\begin{aligned} & Q_n^n[|\beta_n(\hat{a}_n - a(P_n))| \le 2] \\ & \ge E_{Q_n^n}[\chi(\beta_n(\hat{a}_n - a(P_n)))] \\ & \ge e^{-2nH(P_n,Q_n)-1} \left(E_{P_n^n}[\chi(\beta_n(\hat{a}_n - a(P_n)))] - \frac{1}{2} \right) \\ & \ge e^{-2nH(P_n,Q_n)-1} \left(1 - \delta - \frac{1}{2} \right) \ge \frac{1}{4e} e^{-2nH(P_n,Q_n)} \ge \frac{11}{4e}\delta > \delta \end{aligned} \tag{2.9}$$

by (2.5); recall that $\delta \le 1/4$. Furthermore, we have

$$Q_n^n[\beta_n|\hat{a}_n - a(Q_n)| \le 2] \ge E_{Q_n^n}[\chi(\beta_n(\hat{a}_n - a(Q_n)))] \ge 1 - \delta \tag{2.10}$$

by the opposite of (2.8) and the choice of χ. The bounds (2.10) and (2.9) imply that the events $\{\beta_n|\hat{a}_n - a(Q_n)| \le 2\}$ and $\{\beta_n|\hat{a}_n - a(P_n)| \le 2\}$ have a non-empty intersection. This implies

$$\beta_n|a(P_n) - a(Q_n)| \le 4, \tag{2.11}$$

which contradicts the hypothesis (2.6).

□

3 General lower bounds

In this section, we prepare the proofs of Theorems 1.1 and 1.2 by providing some general, abstract lower bounds for estimators. Since we expect these lemmas to be useful in other contexts than the estimation of thresholds, too, we do not assume Ω, $\mathcal{P}$, and $a : P \mapsto a(P)$

to have the specific form described in section 1.2. Rather, in this section, $(\Omega, \mathcal{A})$ may be any measurable space, $\mathcal{P}$ may be any set of probability measures on $(\Omega, \mathcal{A})$, endowed with a topology, and $a : \mathcal{P} \to \mathbb{R}$ may be any parameter; only some very general assumptions for the topology on $\mathcal{P}$ and for the parameter a are required in Lemma 3.2, below. The first lemma in this section is a general statement which plays an essential role in the proof of Theorem 1.1.

Lemma 3.1 *Take a sequence $(\beta_n)_{n\in\mathbb{N}}$ of positive numbers, $\delta \in (0, 1/4]$, and a nonempty open set $\mathcal{U} \subseteq \mathcal{P}$. For all large $n \in \mathbb{N}$, assume that there are $P_n, Q_n \in \mathcal{U}$, such that*

$$nH(P_n, Q_n) \le \frac{1}{2} \log \frac{1}{11\delta} \quad \text{and} \quad \beta_n |a(P_n) - a(Q_n)| > 4 \tag{3.1}$$

hold. Then for all sequences $(\hat{a}_n : \Omega^n \to \mathbb{R})_{n\in\mathbb{N}}$ of estimators, one has

$$\liminf_{n\to\infty} \sup_{P\in\mathcal{U}} P^n[\beta_n|\hat{a}_n - a(P)| > 1] \ge \delta. \tag{3.2}$$

Proof. Take $\chi = 1_{[-1,1]}$. By (3.1) and Lemma 2.2, (2.7) or (2.8) holds for all large n. Thus

$$\inf_{P\in\mathcal{U}} P^n[\beta_n|\hat{a}_n - a(P)| \le 1] < 1 - \delta, \tag{3.3}$$

holds for all large n; hence (3.2) is true.

□

The constant "4" on the right hand side in (3.1) is not optimal. However, improving it does not improve our main theorems.

The next lemma provides an abstract key ingredient for the proof of Theorem 1.2.

Lemma 3.2 *Let $a : \mathcal{P} \to \mathbb{R}$ be a parameter, and $(\hat{a}_n : \Omega^n \to \mathbb{R})_{n\in\mathbb{N}}$ be a sequence of estimators. Let $\mathcal{P}$ be endowed with a Baire space metrisizable topology. Assume that $a : \mathcal{P} \to \mathbb{R}$ is continuous. Furthermore, assume that for all $P \in \mathcal{P}$, the total variation distance from P, i.e. $Q \mapsto \|P - Q\|_V = \sup_{A\in\mathcal{A}} |P[A] - Q[A]|$, $Q \in \mathcal{P}$, is continuous, too. Let $(\beta_n)_{n\in\mathbb{N}}$ be a sequence of positive numbers, and take $\delta \in (0, 1/4]$.*

Suppose that for all $P \in \mathcal{P}$, for all neighbourhoods $\mathcal{N}$ of P, and for all $m \in \mathbb{N}$, there are $n \ge m$ and $Q_n \in \mathcal{N}$, such that

$$\beta_n|a(P) - a(Q_n)| > 4 \quad \textit{and} \quad nH(P, Q_n) \le \frac{1}{2}|\log(11\delta)| \tag{3.4}$$

are valid. Then for all non-empty open sets $\mathcal{U} \subseteq \mathcal{P}$, one has

$$\sup_{P\in\mathcal{U}} \limsup_{n\to\infty} P^n[\beta_n|\hat{a}_n - a(P)| > 1] \ge \delta. \tag{3.5}$$

Proof. Let $\chi : \mathbb{R} \to [0, 1]$ be a continuous function with $\chi(x) = 1$ for $|x| \le 1$, and $\chi(x) = 0$ for $|x| \ge 2$. For $m, n \in \mathbb{N}$, we set

$$\mathcal{P}_n = \{P \in \mathcal{P} : E_{P^n}[\chi(\beta_n(\hat{a}_n - a(P)))] \ge 1 - \delta\} \tag{3.6}$$

and

$$\mathcal{F}_m = \bigcap_{n \geq m} \mathcal{P}_n. \tag{3.7}$$

We claim that the map

$$\mathcal{P} \to [0,1], \quad P \mapsto E_{P^n}[\chi(\beta_n(\hat{a}_n - a(P)))] \tag{3.8}$$

is continuous. To prove this claim, let $P \in \mathcal{P}$, and consider a sequence $(Q_k)_k$ in $\mathcal{P}$ converging to P. Then for all $\omega \in \Omega^n$, we have

$$\chi(\beta_n(\hat{a}_n(\omega) - a(Q_k))) \stackrel{k\to\infty}{\longrightarrow} \chi(\beta_n(\hat{a}_n(\omega) - a(P))) \tag{3.9}$$

by the continuity of a and of χ. Using Lebesgue's dominated convergence theorem, this implies

$$E_{P^n}[\chi(\beta_n(\hat{a}_n - a(Q_k)))] \stackrel{k\to\infty}{\longrightarrow} E_{P^n}[\chi(\beta_n(\hat{a}_n - a(P)))]; \tag{3.10}$$

recall that χ takes values in the unit interval. Furthermore,

$$|E_{Q_k^n}[\chi(\beta_n(\hat{a}_n - a(Q_k)))] - E_{P^n}[\chi(\beta_n(\hat{a}_n - a(Q_k)))]| \leq n\|Q_k - P\|_V \stackrel{k\to\infty}{\longrightarrow} 0, \tag{3.11}$$

since by our hypothesis, the total variation distance from P is continuous. Combining (3.10) and (3.11), we get

$$E_{Q_k^n}[\chi(\beta_n(\hat{a}_n - a(Q_k)))] \stackrel{k\to\infty}{\longrightarrow} E_{P^n}[\chi(\beta_n(\hat{a}_n - a(P)))], \tag{3.12}$$

which shows that $E_{P^n}[\chi(\beta_n(\hat{a}_n - a(P)))]$ depends continuously on P. Note that we used in the last step that the chosen topology on $\mathcal{P}$ is metrisizable (or, at least, that sequential continuity on $\mathcal{P}$ implies continuity).

The continuity of the map described in (3.8) implies that the sets $\mathcal{P}_n \subseteq \mathcal{P}$ are closed; thus their intersections $\mathcal{F}_m$ are closed, too.

Next, we show that the sets $\mathcal{F}_m \subseteq \mathcal{P}$, $m \in \mathbb{N}$, are nowhere dense. To check this, take $P \in \mathcal{F}_m$ and a neighbourhood $\mathcal{N}$ of P in $\mathcal{P}$. By the hypothesis of the lemma, there exist $n \geq m$ and $Q_n \in \mathcal{N}$ such that (3.4) holds. Then Lemma 2.2 implies

$$E_{P^n}[\chi(\beta_n(\hat{a}_n - a(P)))] < 1 - \delta \quad \text{or} \quad E_{Q_n^n}[\chi(\beta_n(\hat{a}_n - a(Q_n)))] < 1 - \delta, \tag{3.13}$$

i.e. $P \notin \mathcal{P}_n \supseteq \mathcal{F}_m$ or $Q_n \notin \mathcal{P}_n \supseteq \mathcal{F}_m$, and thus $\mathcal{N} \not\subseteq \mathcal{F}_m$. This shows that indeed $\mathcal{F}_m$ is nowhere dense.

Let $\mathcal{U} \subseteq \mathcal{P}$ be a non-empty open set. Since $\mathcal{P}$ is endowed with a Baire space topology, we conclude that $\mathcal{U}$ is not contained in $\bigcup_{m\in\mathbb{N}} \mathcal{F}_m$; so we can take $P \in \mathcal{U} \setminus \bigcup_{m\in\mathbb{N}} \mathcal{F}_m$. For this P, we know that $P \notin \mathcal{P}_n$ for infinitely many $n \in \mathbb{N}$. Thus we get

$$\liminf_{n\to\infty} E_{P^n}[\chi(\beta_n(\hat{a}_n - a(P)))] \leq 1 - \delta. \tag{3.14}$$

Using

$$P^n[\beta_n|\hat{a}_n - a(P)| > 1] \geq 1 - E_{P^n}[\chi(\beta_n(\hat{a}_n - a(P)))], \tag{3.15}$$

this implies

$$\limsup_{n\to\infty} P^n[\beta_n|\hat{a}_n - a(P)| > 1] \geq \delta, \tag{3.16}$$

and thus the claim (3.5) follows.

□

4 Lower bounds for errors in threshold estimation

In this section, we prove Theorems 1.1 and 1.2. Thus, let $\mathcal{P}$ denote again the concrete space of probability measures defined in subsection 1.2, endowed with the metric d, defined in (1.8). Note that $(\tilde{\mathcal{P}}, d)$ is a *complete* metric space. Perturbing $P \in \mathcal{P}$ slightly (with respect to the metric d) will move the intersection point $a(P)$ only slightly, and small enough perturbations can destroy neither the uniqueness of the intersection point $a(P)$ nor the transversality condition; thus $a : \mathcal{P} \to [0, 1]$ is a *continuous* parameter, and $\mathcal{P}$ is an *open* subset of the space $\tilde{\mathcal{P}}$; recall (1.7). In particular, by Baire's category theorem, $\mathcal{P}$ is a Baire space.

The following lemma contains the basic entropy estimate for perturbed densities. Here is the idea: A given probability density f_P is slightly modified by a perturbation of order $O(\epsilon)$ in a neighbourhood of size $O(\epsilon)$ of the transversal intersection point $a(P)$. Let Q denote the probability measure corresponding to the modified density f_Q; then we show that the entropy $H(P, Q)$ has roughly the order $O(\epsilon^3)$, but the parameter $a(Q)$ deviates from $a(P)$ on a scale of order ϵ. The cube of ϵ arising in the entropy bound is the key to derive the cube root asymptotic lower bounds in this article.

Lemma 4.1 *Let $P \in \mathcal{P}$, and let $\mathcal{U} \subseteq \mathcal{P}$ be an open neighbourhood of P. Then there is $c_1 = c_1(P,\mathcal{U}) > 0$, such that for every $\delta \in (0, 1/4]$ and for all large n (say for $n \geq n_0(P,\mathcal{U}, \delta)$), there is $Q_n \in \mathcal{U}$, such that the following holds:*

$$nH(P, Q_n) \leq \frac{1}{2}|\log(11\delta)| \quad \text{and} \quad \frac{n^{1/3}}{c_1|\log(11\delta)|^{1/3}}|a(P) - a(Q_n)| > 4. \tag{4.1}$$

Proof. Choose a ball $\mathcal{N} \subseteq \mathcal{U}$ (with respect to the metric d) centered at P. Let r denote the radius of $\mathcal{N}$. Let $c_1 > 0$ be small enough (to be specified below). Take $\delta \in (0, 1/4]$. We abbreviate

$$\beta_n = \frac{n^{1/3}}{c_1|\log(11\delta)|^{1/3}} > 0. \tag{4.2}$$

Take a fixed, compactly supported $\phi \in C^1(\mathbb{R})$ with $\phi(0) = 1$, $0 \leq \phi \leq 1$, and

$$\|\phi'\|_\infty \|f_P\|_\infty < r. \tag{4.3}$$

For $\epsilon > 0$, we set

$$\Xi_\epsilon(x) = \epsilon\phi\left(\frac{x - a(P)}{\epsilon}\right). \tag{4.4}$$

We set

$$\rho_P^+ = \frac{f_P^+}{f_P^+ + f_P^-}, \qquad \rho_P^- = \frac{f_P^-}{f_P^+ + f_P^-}; \tag{4.5}$$

at least in some compact neighbourhood V_P of $a(P)$, these functions are well defined with values in $[1/3, 2/3]$. For all small $\epsilon > 0$, Ξ_ϵ is supported in such a neighbourhood V_P. We set

$$\epsilon_n = c_2|\log(11\delta)|^{1/3} n^{-1/3}, \tag{4.6}$$

where $c_2 = c_2(P, \phi) = (\|f_P\|_\infty \|\phi\|_2^2)^{-1/3}$. Let Q_n be defined by its density f_{Q_n}, where

$$f_{Q_n}^+ = (1 + \Xi_{\epsilon_n} \rho_P^-) f_P^+, \tag{4.7}$$

$$f_{Q_n}^- = (1 - \Xi_{\epsilon_n} \rho_P^+) f_P^-, \tag{4.8}$$

$$f_{Q_n}(x, y) = f_{Q_n}^+(x) 1_{\{y=1\}} + f_{Q_n}^-(x) 1_{\{y=-1\}}. \tag{4.9}$$

Here $\Xi_{\epsilon_n} \rho_P^\pm$ is to be interpreted as 0 outside the support of Ξ_{ϵ_n}. Note that for all large n, $f_{Q_n}^\pm$ is well defined. As a consequence of the assumptions (1.9) and (1.10), one sees that $f_P^+(a(P)) = f_P^-(a(P)) > 0$. For large n, ϵ_n is small, thus $f_{Q_n}^\pm$ is non-negative, and f_{Q_n} is a probability density. Furthermore, using (4.3) and $|\rho_P^\mp f_P^\pm| \leq \|f_P\|_\infty$, one sees that $d(Q_n, P) < r$ and thus $Q_n \in \mathcal{N} \subseteq \mathcal{U}$ hold for all large n. We calculate for $(x, y) \in \Omega$:

$$\frac{dQ_n}{dP}(x, y) = 1 + (1_{\{y=1\}} \rho_P^-(x) - 1_{\{y=-1\}} \rho_P^+(x)) \Xi_{\epsilon_n}(x). \tag{4.10}$$

For $|t - 1| \leq 1/2$, one has $-\log t + t - 1 \leq |t - 1|^2$. So

$$H(P, Q_n) = E_P \left[-\log \frac{dQ_n}{dP} + \frac{dQ_n}{dP} - 1 \right] \leq E_P \left[\left(\frac{dQ_n}{dP} - 1 \right)^2 \right] \tag{4.11}$$

for $|\frac{dQ_n}{dP} - 1| \leq 1/2$, which holds for all large n. Note that $|\Xi_{\epsilon_n}| \leq \epsilon_n$. For all large n, $\rho_P^\pm \in [1/3, 2/3]$ holds on the support of Ξ_{ϵ_n}. Then one has

$$\frac{1}{3} \leq |1_{\{y=1\}} \rho_P^-(x) - 1_{\{y=-1\}} \rho_P^+(x)| \leq \frac{2}{3} \tag{4.12}$$

on the support of Ξ_{ϵ_n}, and therefore

$$\left(\frac{dQ_n}{dP} - 1 \right)^2 \leq \left(\frac{2}{3} \right)^2 \Xi_{\epsilon_n}^2(X). \tag{4.13}$$

We get the following "$O(\epsilon_n^3)$" estimate for the relative entropy:

$$nH(P, Q_n) \leq \frac{4}{9} n E_P[\Xi_{\epsilon_n}^2(X)] \leq \frac{4}{9} n \|f_P\|_\infty \|\Xi_{\epsilon_n}\|_2^2 \leq \frac{1}{2} n \|f_P\|_\infty \|\phi\|_2^2 \epsilon_n^3 \leq \frac{1}{2} |\log(11\delta)| \tag{4.14}$$

by the choice (4.6) of ϵ_n.

(As a side remark, note that the estimate (4.14) relies on our choice to take the Lebesgue measure $\lambda_{[0,1]}$ in the reference measure. In some cases with arbitrary reference measures, it would break down.)

On the other hand, defining $\rho_{Q_n}^\pm$ in analogy to (4.5),

$$\rho_{Q_n}^+ = (1 + \Xi_{\epsilon_n} \rho_P^-) \rho_P^+, \tag{4.15}$$

$$\rho_{Q_n}^- = (1 - \Xi_{\epsilon_n} \rho_P^+) \rho_P^- \tag{4.16}$$

follows from (4.7) and (4.8), using $f^+_{Q_n} + f^-_{Q_n} = f^+_P + f^-_P$. Using $\rho^-_P(a(P)) = \rho^+_P(a(P)) = \rho^+_{Q_n}(a(Q_n)) = 1/2$ and (4.15), we get

$$\begin{aligned}\frac{\epsilon_n}{4} &= \frac{1}{4}|\Xi_{\epsilon_n}(a(P))| \\ &= |\Xi_{\epsilon_n}(a(P))\, \rho^-_P(a(P))\, \rho^+_P(a(P))| \\ &= |\rho^+_P(a(P)) - \rho^+_{Q_n}(a(P))| \\ &= |\rho^+_{Q_n}(a(Q_n)) - \rho^+_{Q_n}(a(P))| \\ &\leq \|(\rho^+_{Q_n})'\|_{\infty,V_P}|a(Q_n) - a(P)|. \end{aligned} \tag{4.17}$$

Taking the derivative of (4.15) and taking a supremum over V_P, we see that $\|(\rho^+_{Q_n})'\|_{\infty,V_P}$ is bounded by a constant $c_3(P,\phi) > 0$ for all large n; note that $\|\Xi'_{\epsilon_n}\|_\infty = \|\phi'\|_\infty$ does not depend on n. We obtain

$$\beta_n|a(Q_n) - a(P)| \geq \beta_n \frac{\epsilon_n}{4c_3(P)} = \frac{n^{1/3}}{c_1|\log(11\delta)|^{1/3}} \frac{c_2|\log(11\delta)|^{1/3} n^{-1/3}}{4c_3(P,\phi)} > 4 \tag{4.18}$$

when we choose

$$0 < c_1(P,\mathcal{U}) < \frac{c_2(P,\phi)}{16c_3(P,\phi)}; \tag{4.19}$$

recall the choices (4.2) and (4.6) of β_n and ϵ_n. The statements (4.18) and (4.14) together are just the claim (4.1).

□

Proof of Theorem 1.1. Take a fixed $P \in \mathcal{U}$, and take $c_1(\mathcal{U}) = c_1(P,\mathcal{U}) > 0$ from Lemma 4.1. Then Lemma 4.1 guarantees that the hypothesis (3.1) of Lemma 3.1 holds with $P_n = P$, where β_n is again given by (4.2). Thus Lemma 3.1 yields the claim (1.11).

□

Proof of Theorem 1.2. $\mathcal{P}$ with the metric d is indeed a Baire space, and $a : \mathcal{P} \to (0,1)$ is continuous. Note that the total variation distance is continuous with respect to d.

We check that Lemma 3.2 is applicable with $\delta = 1/4$. Let $P \in \mathcal{P}$, and let $\mathcal{N}$ be a neighbourhood of P in $\mathcal{P}$. We apply Lemma 4.1 to obtain a sequence $(Q_n)_n$ in $\mathcal{P}$ such that (4.1) holds. Hence, we get for all large n (say for $n \geq c_4(P,\mathcal{U})$):

$$\beta_n|a(P) - a(Q_n)| \geq \frac{n^{1/3}}{c_1(P,\mathcal{U})|\log(11\delta)|^{1/3}}|a(P) - a(Q_n)| > 4 \tag{4.20}$$

by (1.13). Together with the entropy bound in (4.1), this shows that Lemma 3.2 is indeed applicable, and it yields the claim (1.14).

□

Remark. Note that this proof would break down if we had taken β_n on the "critical" scale: "$\beta_n = \text{const} \cdot n^{1/3}$": Indeed, the constant $c_1(P,\mathcal{U})$ depends on $\mathcal{U}$ (and it really diverges as $\mathcal{U}$ gets smaller), but β_n must not depend on the choice of $\mathcal{U}$. This breakdown has to occur, since Theorem 1.3 shows that the claim (1.14) of Theorem 1.2 cannot hold anymore on the critical scale $\beta_n = \text{const} \cdot n^{1/3}$.

5 Optimal estimators for thresholds

In this section, we prove Theorem 1.3. The optimal estimators, whose errors asymptotically meet the border permitted by the lower bounds, are constructed by a two-step procedure: In the first step (Lemma 5.1 below), we use the empirical risk minimizer to obtain a "starting estimator", which yields error terms roughly on the scale $n^{-1/3}$. The second step (Lemma 5.2 below) constructs a refined family of estimators, based on a starting approximation a_0.

Let $(X_i, Y_i) : \Omega^n \to \Omega$ again denote the canonical projections.

Lemma 5.1 *There is a sequence of estimators $\hat{a}_n : \Omega^n \to [0,1]$, $n \in \mathbb{N}$, such that for all $P \in \mathcal{P}$, there is a neighbourhood $\mathcal{N}(P)$ of P in $\mathcal{P}$, such that the following holds:*

$$\lim_{L\to\infty} \sup_{Q\in\mathcal{N}(P)} \lim_{n\to\infty} Q^n[|\hat{a}_n - a(Q)| \geq Ln^{-1/3}] = 0. \tag{5.1}$$

Proof. Consider the empirical risk minimizers $\hat{a}_n$, $n \in \mathbb{N}$, given by (1.5). We abbreviate $f_Q^\Sigma = f_Q^+ + f_Q^-$. Take $P \in \mathcal{P}$ and

$$\mathcal{N}(P) = \{Q \in \mathcal{P} : 2|(\rho_Q^+)'(a(Q))| > |(\rho_P^+)'(a(P))|,\ 2f_Q^\Sigma(a(Q)) > f_P^\Sigma(a(P))\}, \tag{5.2}$$

where ρ_Q^+ is again defined as in (4.5). By the transversality of the intersection point $a(Q)$ of f_Q^+ and f_Q^-, the maps $Q \mapsto (\rho_Q^+)'(a(Q))$ and $Q \mapsto f_Q^\Sigma(a(Q))$, $Q \in \mathcal{N}(P)$, are continuous. Furthermore, $f_P^\Sigma(a(P)) > 0$ and $|(\rho_P^+)'(a(P))| > 0$ hold. Using these facts, one sees that $\mathcal{N}(P)$ is a neighbourhood of P.

Take $Q \in \mathcal{N}(P)$. By Theorem 2.2 in Mohammadi and van de Geer's article [MvdG03], we know that

$$n^{1/3}(\hat{a}_n - a(Q)) \xrightarrow{\mathcal{L}} \left[(\rho_Q^+)'(a(Q))\sqrt{f_Q^\Sigma(a(Q))}\right]^{-2/3} Z \tag{5.3}$$

as $n \to \infty$ for some continuous random variable Z, not depending on Q (with respect to some probability measure $\mathbb{P}$). We set

$$\alpha = \inf_{Q\in\mathcal{N}(P)} \left[(\rho_Q^+)'(a(Q))\sqrt{f_Q^\Sigma(a(Q))}\right]^{2/3}. \tag{5.4}$$

Note that $\alpha > 0$ by the choice (5.2) of $\mathcal{N}(P)$. We obtain the claim (5.1):

$$\begin{aligned}\sup_{Q\in\mathcal{N}(P)} \lim_{n\to\infty} Q[|\hat{a}_n - a(Q)| \geq Ln^{-1/3}] &= \sup_{Q\in\mathcal{N}(P)} \mathbb{P}\left[Z \geq \left[(\rho_Q^+)'(a(Q))\sqrt{f_Q^\Sigma(a(Q))}\right]^{2/3} L\right] \\ &\leq \mathbb{P}[Z \geq \alpha L] \overset{L\to\infty}{\longrightarrow} 0. \end{aligned} \tag{5.5}$$

□

The next lemma is used to construct the refined estimators in Theorem 1.3. Here is the idea: Given a starting approximation a_0 for $a(Q)$ with an error on the scale $n^{-1/3}$, consider all data points in a neighbourhood of size $Ln^{-1/3}$, where L is large, but fixed. Then construct a regression line through the data points in this neighbourhood, and take the intersection of this regression line with the x-axis as the refined estimator.

Lemma 5.2 *There is a family of estimators* $(\hat{a}_{n,L,a_0} : \Omega^n \to \mathbb{R})_{n\in\mathbb{N},\, L>0,\, a_0\in(0,1)}$ *with the following property: For all* $P \in \mathcal{P}$*, there is a neighbourhood* $\mathcal{N} \subseteq \mathcal{P}$ *of* P*, such that for all* $T > 0$ *one has*

$$\lim_{L\to\infty} \sup_{Q\in\mathcal{N}} \limsup_{n\to\infty} \sup_{\substack{a_0\in(0,1)\\ |a_0-a(Q)|\le Ln^{-1/3}}} Q^n[n^{1/3}|\hat{a}_{n,L,a_0} - a(Q)| > T] = 0. \tag{5.6}$$

Proof. Given $n \in \mathbb{N}$, $L > 0$, and $a_0 \in (0,1)$, we introduce the abbreviations $\tilde{X}_i = X_i - a_0$, $M = Ln^{-1/3}$, $I_M = [-M, M]$, and we define the random set

$$J = \{j :\ 1 \le j \le n,\ \tilde{X}_j \in I_M\}. \tag{5.7}$$

We define the estimator $\hat{a}_{n,L,a_0}$ as follows: Consider the regression line ℓ (equation $y = \hat{b}_1 x + \hat{b}_2$) through the points $(\tilde{X}_j, Y_j)$, $j \in J$, provided it is well defined, i.e. provided there are at least two different $\tilde{X}_{j_1} \ne \tilde{X}_{j_2}$, $j_1, j_2 \in J$. If $\hat{b}_1 \ne 0$, we set

$$\hat{a}_{n,L,a_0} = a_0 - \frac{\hat{b}_2}{\hat{b}_1}. \tag{5.8}$$

Geometrically, this means that $\hat{a}_{n,L,a_0}$ is the intersection of the real axis with the regression line through the points (X_j, Y_j), where only the points with $|X_j - a_0| \le Ln^{-1/3}$ are taken, whenever this intersection is well defined. If the regression line ℓ is not well-defined, or if $\hat{b}_1 = 0$, we set $\hat{a}_{n,L,a_0} = a_0$, just to have a definite value in this case, too.

We abbreviate for $Q \in \mathcal{P}$:

$$s_Q = f_Q^{\Sigma}(a(Q)), \tag{5.9}$$
$$t_Q = (f_Q^+)'(a(Q)) - (f_Q^-)'(a(Q)), \tag{5.10}$$

where again $f_Q^{\Sigma} = f_Q^+ + f_Q^-$. Let $P \in \mathcal{P}$, and take the following neighbourhood of P:

$$\mathcal{N} = \left\{ Q \in \mathcal{P} :\ \begin{array}{l} |t_Q| > |t_P|/2,\ s_Q < 2s_P,\ d(P,Q) < 1, \\ a(Q) > a(P)/2,\ 1 - a(Q) > (1 - a(P))/2 \end{array} \right\}. \tag{5.11}$$

Take $T > 0$ and $\delta > 0$. Let L be large enough (more specifically, so large that

$$2L \ge T, \tag{5.12}$$
$$L^3 \ge S^2 := \frac{5c_5}{\delta}, \tag{5.13}$$
$$c_6 \frac{S}{\sqrt{L}} < \frac{T}{5} \tag{5.14}$$

hold with some positive constants $c_5 = c_5(\mathcal{N})$ and $c_6 = c_6(\mathcal{N})$, to be specified below). We claim that for all $Q \in \mathcal{N}$, one has

$$\limsup_{n\to\infty} \sup_{\substack{a_0\in(0,1)\\ |a_0-a(Q)|\le Ln^{-1/3}}} Q^n[n^{1/3}|\hat{a}_{n,L,a_0} - a(Q)| > T] \le \delta. \tag{5.15}$$

The claim (5.6) of the lemma then follows immediately from the statement (5.15).

To prove (5.15), let $Q \in \mathcal{N}$. For every $n \in \mathbb{N}$ and for every $a_0 \in (0,1)$ with

$$|a_0 - a(Q)| \leq L n^{-1/3}, \tag{5.16}$$

we are going to define an event $B = B(Q, n, a_0, S, L) \subseteq \Omega^n$ with the property

$$Q^n[B(Q, n, a_0, S, L)] \geq 1 - \delta. \tag{5.17}$$

(This is done in (5.29), below.) We then show that for all large n (say for $n \geq n_0(Q, \mathcal{N}, S, L)$, but uniformly in the choice of a_0), one has

$$B(Q, n, a_0, S, L) \subseteq \{n^{1/3} |\hat{a}_{n,L,a_0} - a(Q)| \leq T\}. \tag{5.18}$$

Once we have proven (5.17) and (5.18), the claim (5.15) is an immediate consequence.

Take $Q \in \mathcal{N}$ and $a_0 \in (0,1)$ with the constraint (5.16). It is convenient to shift quantities by a_0: We set

$$\tilde{a} = \tilde{a}(Q) = a(Q) - a_0, \tag{5.19}$$
$$\tilde{f}_Q^{\pm}(x) = f_Q^{\pm}(x + a_0), \tag{5.20}$$
$$\tilde{f}_Q^{\Sigma}(x) = f_Q^{\Sigma}(x + a_0). \tag{5.21}$$

In particular, note that

$$|\tilde{a}| \leq M = L n^{-1/3}, \tag{5.22}$$

and that for all large n (uniformly in a_0 and $Q \in \mathcal{N}$ with the constraint (5.16)), $\tilde{f}_Q^{\pm}$ is defined at least on I_M.

The coefficients $\hat{b}_1$ and $\hat{b}_2$ of the regression line ℓ are determined by the linear system

$$A\hat{b} = c, \tag{5.23}$$

where

$$A = \begin{pmatrix} \sum_{j \in J} \tilde{X}_j^{\,2} & \sum_{j \in J} \tilde{X}_j \\ \sum_{j \in J} \tilde{X}_j & |J| \end{pmatrix}, \quad \hat{b} = \begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \end{pmatrix}, \quad c = \begin{pmatrix} \sum_{j \in J} \tilde{X}_j Y_j \\ \sum_{j \in J} Y_j \end{pmatrix}. \tag{5.24}$$

We introduce the (normalized) difference of the coefficient matrix from its expected value:

$$\begin{pmatrix} M^{5/2}\Delta_{11} & M^{3/2}\Delta_{12} \\ M^{3/2}\Delta_{21} & M^{1/2}\Delta_{22} \end{pmatrix} = \frac{1}{\sqrt{n}}(A - E_{Q^n}[A]), \tag{5.25}$$
$$\begin{pmatrix} M^{3/2}\Gamma_1 \\ M^{1/2}\Gamma_2 \end{pmatrix} = \frac{1}{\sqrt{n}}(c - E_{Q^n}[c]). \tag{5.26}$$

Our reason to normalize the Δ_{ij} and Γ_i in this way is the following bound on the variances (to be read element-wise):

$$\mathrm{Var}_{Q^n}\left[\begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{pmatrix}\right] \leq E_{Q^n}\left[1_{\{\tilde{X}_1 \in I_M\}} \begin{pmatrix} \tilde{X}_1^4/M^5 & \tilde{X}_1^2/M^3 \\ \tilde{X}_1^2/M^3 & 1/M \end{pmatrix}\right] \leq \begin{pmatrix} c_5 & c_5 \\ c_5 & c_5 \end{pmatrix}, \tag{5.27}$$

$$\mathrm{Var}_{Q^n}\left[\begin{pmatrix}\Gamma_1\\ \Gamma_2\end{pmatrix}\right] \le E_{Q^n}\left[1_{\{\tilde X_1\in I_M\}}\begin{pmatrix}\tilde X_1^2 Y_1^2/M^3\\ Y_1^2/M\end{pmatrix}\right] \le \begin{pmatrix}c_5\\ c_5\end{pmatrix}, \tag{5.28}$$

with some constant $c_5 = c_5(\mathcal{N}) > 0$. We used that the density f_Q of (X_i, Y_i) is uniformly bounded for $Q \in \mathcal{N}$.

Here is the definition of the event B:

$$B(Q, n, a_0, S, L) = \{|\Delta_{i,j}| \le S,\ |\Gamma_i| \le S \quad (i, j \in \{1, 2\})\}. \tag{5.29}$$

By Chebyshev's inequality, (5.27), (5.28), and the choice (5.13) of S imply the claim (5.17):

$$Q^n[B(Q, n, a_0, S, L)^c] \le \frac{5c_5}{S^2} = \delta. \tag{5.30}$$

The factor "5" arises, since there are 5 random variables involved; (recall $\Delta_{12} = \Delta_{21}$).

In the rest of this proof, we verify the claim (5.18) for all large n (uniformly in a_0). So assume that the event $B(Q, n, a_0, S, L)$ holds.

The system (5.23) is equivalent to

$$\begin{aligned}&\begin{pmatrix}\int_{I_M} x^2\tilde f_Q^\Sigma(x)\,dx + \frac{M^{5/2}\Delta_{11}}{\sqrt n} & \int_{I_M} x\tilde f_Q^\Sigma(x)\,dx + \frac{M^{3/2}\Delta_{12}}{\sqrt n}\\ \int_{I_M} x\tilde f_Q^\Sigma(x)\,dx + \frac{M^{3/2}\Delta_{21}}{\sqrt n} & \int_{I_M} \tilde f_Q^\Sigma(x)\,dx + \frac{M^{1/2}\Delta_{22}}{\sqrt n}\end{pmatrix}\hat b\\ &= \begin{pmatrix}\int_{I_M} x(\tilde f_Q^+(x) - \tilde f_Q^-(x))\,dx + \frac{M^{3/2}\Gamma_1}{\sqrt n}\\ \int_{I_M} (\tilde f_Q^+(x) - \tilde f_Q^-(x))\,dx + \frac{M^{1/2}\Gamma_2}{\sqrt n}\end{pmatrix}.\end{aligned} \tag{5.31}$$

Let us introduce some notation, used in the Taylor approximations below. Variables $\xi_j = \xi_j(x, \tilde a, Q, a_0)$ denote some values between x and $\tilde a$. Variables ϵ_j denote error terms which are bounded by $\mathcal{N}$-dependent constants:

$$|\epsilon_j| \le \mathrm{const}(\mathcal{N}), \tag{5.32}$$

and δ_j denote error terms which are bounded by

$$|\delta_j| \le \mathrm{const}\cdot\sigma(Q, 2M), \tag{5.33}$$

where

$$\sigma(Q, r) = \max_{y=\pm1}\sup_{|x_1-x_2|\le r} |\partial_1 f_Q(x_1, y) - \partial_1 f_Q(x_2, y)| \xrightarrow{r\to 0} 0 \tag{5.34}$$

denotes the modulus of continuity of $\partial_1 f_Q$; recall that $f_Q \in C^1(\Omega)$. We emphasize that the constants "const($\mathcal{N}$)", "const" in (5.32) and (5.33) must not depend on a_0. Recall that $|\tilde a| \le M$. Let us approximate the integrals in (5.31) by Taylor's formula.

$$\int_{I_M} x^2\tilde f_Q^\Sigma(x)\,dx = \int_{-M}^{M} x^2(s_Q + (x-\tilde a)(\tilde f_Q^\Sigma)'(\xi_1))\,dx = \frac{2}{3}M^3 s_Q + \epsilon_1 M^4, \tag{5.35}$$

$$\int_{I_M} x\tilde f_Q^\Sigma(x)\,dx = \int_{-M}^{M} x(s_Q + (x-\tilde a)(\tilde f_Q^\Sigma)'(\xi_2))\,dx = \epsilon_2 M^3, \tag{5.36}$$

$$\int_{I_M} \tilde f_Q^\Sigma(x)\,dx = \int_{-M}^{M} (s_Q + (x-\tilde a)(\tilde f_Q^\Sigma)'(\xi_3))\,dx = 2M s_Q + \epsilon_3 M^2, \tag{5.37}$$

$$\int_{I_M} x(\tilde f_Q^+(x)-\tilde f_Q^-(x))\,dx = \int_{-M}^{M} x(x-\tilde a)((\tilde f_Q^+)'(\xi_4)-(\tilde f_Q^-)'(\xi_4))\,dx$$
$$= \frac{2}{3}M^3 t_Q + \delta_2 M^3 = \epsilon_4 M^3, \tag{5.38}$$
$$\int_{I_M} (\tilde f_Q^+(x)-\tilde f_Q^-(x))\,dx = \int_{-M}^{M} (x-\tilde a)((\tilde f_Q^+)'(\xi_5)-(\tilde f_Q^-)'(\xi_5))\,dx$$
$$= -2\tilde a M t_Q + \delta_1 M^2 = \epsilon_5 M^2. \tag{5.39}$$

We rewrite the system (5.31) as

$$\begin{pmatrix} \frac{2}{3}M^3 s_Q + \epsilon_1 M^4 + \frac{M^{5/2}\Delta_{11}}{\sqrt{n}} & \epsilon_2 M^3 + \frac{M^{3/2}\Delta_{12}}{\sqrt{n}} \\ \epsilon_2 M^3 + \frac{M^{3/2}\Delta_{21}}{\sqrt{n}} & 2M s_Q + \epsilon_3 M^2 + \frac{M^{1/2}\Delta_{22}}{\sqrt{n}} \end{pmatrix} \hat b$$
$$= \begin{pmatrix} \frac{2}{3}M^3 t_Q + \delta_2 M^3 + \frac{M^{3/2}\Gamma_1}{\sqrt{n}} \\ -2\tilde a M t_Q + \delta_1 M^2 + \frac{M^{1/2}\Gamma_2}{\sqrt{n}} \end{pmatrix} = \begin{pmatrix} \epsilon_4 M^3 + \frac{M^{3/2}\Gamma_1}{\sqrt{n}} \\ \epsilon_5 M^2 + \frac{M^{1/2}\Gamma_2}{\sqrt{n}} \end{pmatrix}; \tag{5.40}$$

both forms of the right hand side are useful below. Dividing the first row in (5.40) by $(2/3)M^3 s_Q$ and the second row by $2Ms_Q$, we get the normalized system

$$\begin{pmatrix} 1 + M\left[\frac{3}{2}\frac{\epsilon_1}{s_Q} + \frac{3}{2s_Q}\frac{\Delta_{11}}{L^{3/2}}\right] & \frac{3}{2}\frac{\epsilon_2}{s_Q} + \frac{3}{2s_Q}\frac{\Delta_{12}}{L^{3/2}} \\ M^2\left[\epsilon_6 + \frac{1}{2s_Q}\frac{\Delta_{21}}{L^{3/2}}\right] & 1 + M\left[\epsilon_7 M + \frac{1}{2s_Q}\frac{\Delta_{22}}{L^{3/2}}\right] \end{pmatrix} \hat b$$
$$= \frac{t_Q}{s_Q}\begin{pmatrix} 1 + \frac{3}{2}\frac{\delta_2}{t_Q} + \frac{3}{2t_Q}\frac{\Gamma_1}{L^{3/2}} \\ -\tilde a + M\left(\frac{\delta_1}{2t_Q} + \frac{1}{2t_Q}\frac{\Gamma_2}{L^{3/2}}\right) \end{pmatrix} = \frac{t_Q}{s_Q}\begin{pmatrix} \epsilon_8 + \epsilon_9 \frac{\Gamma_1}{L^{3/2}} \\ M\left[\epsilon_{10} + \epsilon_{11}\frac{\Gamma_2}{L^{3/2}}\right] \end{pmatrix}. \tag{5.41}$$

Heuristically, one should view (5.41) as a perturbation of the system

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \frac{t_Q}{s_Q}\begin{pmatrix} 1 \\ -\tilde a \end{pmatrix}, \tag{5.42}$$

for which one knows $-b_2/b_1 = \tilde a$.

By (5.13), (5.14), and the definition (5.29) of the event B, we know for $i,j \in \{1,2\}$:

$$\frac{|\Delta_{i,j}|}{L^{3/2}}, \frac{|\Gamma_i|}{L^{3/2}} \le \frac{S}{L^{3/2}} \le 1 \tag{5.43}$$

and

$$\left|\frac{3}{2t_Q}\frac{\Gamma_i}{L^{3/2}}\right| \le c_6 \frac{S}{L^{3/2}} \le \frac{T}{5L}, \tag{5.44}$$

when we choose the constant in (5.14) to be $c_6 = c_6(\mathcal{N}) = \sup_{Q\in\mathcal{N}} |3/(2t_Q)| \le 3|t_P|^{-1} < \infty$; recall the definition (5.11) of $\mathcal{N}$, and recall that we assume the event B to hold.

For $M \le 1$, we rewrite (5.41) in the following form:

$$\begin{pmatrix} 1 + M\epsilon_{12} & \epsilon_{13} \\ M^2\epsilon_{14} & 1 + M\epsilon_{15} \end{pmatrix}\hat b = \frac{t_Q}{s_Q}\begin{pmatrix} 1 + \epsilon_{16}\delta_2 + \epsilon_{17}\frac{T}{L} \\ -\tilde a + M\left[\epsilon_{18}\delta_1 + \epsilon_{19}\frac{T}{L}\right] \end{pmatrix} = \frac{t_Q}{s_Q}\begin{pmatrix} \epsilon_{20} \\ M\epsilon_{21} \end{pmatrix}, \tag{5.45}$$

where $|\epsilon_{19}| < 1/5$ and $|\epsilon_{17}| \leq 1/5$.

Let us consider the asymptotics of $\hat{b}_2/\hat{b}_1$ as $n \to \infty$, i.e. as $M = Ln^{-1/3} \to 0$. For all large n (uniformly in a_0), the system (5.45) is nonsingular; recall that the error terms ϵ_j are bounded uniformly in a_0. We get for all large n:

$$\frac{\hat{b}_2}{\hat{b}_1} = \frac{\begin{vmatrix} 1 + M\epsilon_{12} & \epsilon_{20} \\ M^2\epsilon_{14} & -\tilde{a} + M\left[\epsilon_{18}\delta_1 + \epsilon_{19}\frac{T}{L}\right] \end{vmatrix}}{\begin{vmatrix} 1 + \epsilon_{16}\delta_2 + \epsilon_{17}\frac{T}{L} & \epsilon_{13} \\ M\epsilon_{21} & 1 + M\epsilon_{15} \end{vmatrix}}. \tag{5.46}$$

Let d_2 and d_1 denote the determinant in the numerator and denominator of the right hand side in (5.46), respectively. Recall that the error terms δ_1, δ_2 converge to 0 as $n \to \infty$, uniformly in a_0; see (5.33) and (5.34). Using $|\epsilon_{19}|, |\epsilon_{17}| \leq 1/5 < 1/4$, we get for all large n (uniformly in a_0):

$$|-\tilde{a} - d_2| \leq \frac{MT}{4L}, \tag{5.47}$$

$$|1 - d_1| \leq \frac{T}{4L}, \tag{5.48}$$

$$\left|1 - \frac{1}{d_1}\right| \leq \frac{T}{2L}. \tag{5.49}$$

We used $T \leq 2L$ in the last step.

Using the definition (5.8) of $\hat{a}_{n,L,a_0}$ and (5.46), we conclude, still for large n (uniformly in a_0):

$$\begin{aligned} |\hat{a}_{n,L,a_0} - a(Q)| &= \left|-\tilde{a} - \frac{\hat{b}_2}{\hat{b}_1}\right| = \left|-\tilde{a} - \frac{d_2}{d_1}\right| \\ &\leq |\tilde{a}|\frac{T}{2L} + \frac{MT}{4L} + \frac{MT}{4L}\frac{T}{2L} \leq \frac{MT}{L} = Tn^{-1/3}; \end{aligned} \tag{5.50}$$

recall that $|\tilde{a}| \leq M$ and $T \leq 2L$. Thus the claim (5.18) holds for all large n, uniformly in a_0.

$\square$

We now use Lemma 5.1 to obtain the starting approximation a_0 required by Lemma 5.2:

Proof of Theorem 1.3. Let us abbreviate $Z_j = (X_j, Y_j)$. Let $n \in \mathbb{N}$, $n \geq 2$, and $L > 0$. We construct the estimator $\hat{a}_{n,L}$ by a two-step procedure: We split the sample $Z_1, \ldots, Z_n$ into two halves $Z_1, \ldots, Z_m$ and $Z_{m+1}, \ldots, Z_{2m}$, where we abbreviate $m = m(n) = \lfloor n/2 \rfloor$. (For odd n, we drop one data point at the end.) We then use the first half of the data to get a "rough" estimate $\hat{a}_m = \hat{a}_m(Z_1, \ldots, Z_m)$ by Lemma 5.1. Using this as a "starting estimate", we refine it by Lemma 5.2, applied to the second half of the data:

$$\hat{a}_{n,L} = \hat{a}_{m,L,\hat{a}_m}(Z_{m+1}, \ldots, Z_{2m}). \tag{5.51}$$

It is important to split the data into two disjoint pieces, since then $\hat{a}_m$ is independent from $(Z_{m+1}, \ldots, Z_{2m})$; thus we have a good control on the distribution of $\hat{a}_{m,L,\hat{a}_m}(Z_{m+1}, \ldots, Z_{2m})$ conditioned on $\hat{a}_m(Z_1, \ldots, Z_m)$. By a slight abuse of notation, we abbreviate $\hat{a}_{m,L,a_0} = \hat{a}_{m,L,a_0}(Z_{m+1}, \ldots, Z_{2m})$.

Let $P \in \mathcal{P}$, and let $\mathcal{N}$ be the intersection of the two neighbourhoods of P that were constructed in Lemma 5.1 and Lemma 5.2. Let $T > 0$ and $\delta > 0$. By the same two lemmas, we know for all large L:

$$\sup_{Q \in \mathcal{N}(P)} \lim_{n \to \infty} Q^n[A(Q, n, L)^c] < \frac{\delta}{2} \tag{5.52}$$

with the events

$$A(Q, n, L) = \{|\hat{a}_m - a(Q)| \leq L m^{-1/3}\}, \tag{5.53}$$

and

$$\sup_{Q \in \mathcal{N}} \limsup_{n \to \infty} \sup_{\substack{a_0 \in (0,1) \\ |a_0 - a(Q)| \leq L m^{-1/3}}} Q^n \left[m^{1/3} |\hat{a}_{m,L,a_0} - a(Q)| > \frac{T}{2} \right] < \frac{\delta}{2}. \tag{5.54}$$

Note that for all large n (say for $n \geq n_0(Q, L)$), we have $\hat{a}_m \in (0, 1)$ on the event $A(Q, n, L)$. Let $Q \in \mathcal{N}$, and let $n \in \mathbb{N}$ be large enough. Using $n^{1/3} \leq 2m^{1/3}$ in the first step, we get

$$\begin{aligned}
& Q^n \left[n^{1/3} |\hat{a}_{n,L} - a(Q)| > T \right] \\
& \leq Q^n \left[m^{1/3} |\hat{a}_{n,L} - a(Q)| > \frac{T}{2} \right] \\
& \leq E_{Q^n} \left[Q^n \left[m^{1/3} |\hat{a}_{n,L} - a(Q)| > \frac{T}{2} \,\middle|\, \hat{a}_m \right] 1_{A(Q,n,L)} \right] + Q^n \left[A(Q, n, L)^c \right] \\
& \leq E_{Q^n} \left[Q^n \left[m^{1/3} |\hat{a}_{m,L,\hat{a}_m} - a(Q)| > \frac{T}{2} \,\middle|\, \hat{a}_m \right] 1_{A(Q,n,L)} \right] + \frac{\delta}{2}.
\end{aligned} \tag{5.55}$$

Using the independence structure and (5.54), we know for almost all a_0 with $|a_0 - a(Q)| \leq L m^{-1/3}$ ("almost all" with respect to the law of $\hat{a}_m$):

$$Q^n \left[m^{1/3} |\hat{a}_{m,L,\hat{a}_m} - a(Q)| > \frac{T}{2} \,\middle|\, \hat{a}_m = a_0 \right] = Q^n \left[m^{1/3} |\hat{a}_{m,L,a_0} - a(Q)| > \frac{T}{2} \right] \leq \frac{\delta}{2}. \tag{5.56}$$

Combining (5.55) and (5.56), we conclude

$$Q^n \left[n^{1/3} |\hat{a}_{n,L} - a(Q)| > T \right] \leq \delta. \tag{5.57}$$

This finishes the proof of the claim (1.15).

□

Let us finally describe a simple counter-example, showing that the "$\limsup_{n \to \infty}$" in the claim (1.14) of Theorem 1.2 cannot be replaced by "$\liminf_{n \to \infty}$".

Let us take $\hat{a}_n$ to be *constant* estimators in the following way: For all $k \in \mathbb{N}_0$ and $n \in [2^k, 2^{k+1}[\cap\mathbb{N}$, set $\hat{a}_n = (n-2^k)2^{-k}$. Then, whatever the value $a(P) \in [0,1]$ is, we have the following: For all $k \in \mathbb{N}_0$ there is $n \in [2^k, 2^{k+1}[\cap\mathbb{N}$ with $|\hat{a}_n - a(P)| \le 2^{-k} \le 2/n$. But then

$$\liminf_{n\to\infty} P^n[\beta_n|\hat{a}_n - a(P)| > 1] = 0 \tag{5.58}$$

whenever $\beta_n = o(n)$.

Intuitively speaking, the counter-example uses the following idea: If you have many non-running clocks, one for every minute of a day, all showing different times, then one of them will show the correct time, up to 1 minute.

Acknowledgment: We are grateful to S. A. van de Geer and W. R. van Zwet for their support and for interesting discussions. Furthermore, the authors would like to thank D. Donoho for interesting discussions, too.

References

[BB02] Peter L. Bartlett and Jonathan Baxter. Estimation and approximation bounds for gradient-based reinforcement learning. *J. Comput. System Sci.*, 64(1):133–150, 2002. Special issue on COLT 2000 (Palo Alto, CA).

[FS98] Yoav Freund and Robert E. Schapire. Large margin classification using the perceptron algorithm. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (Madison, WI, 1998)*, pages 209–217 (electronic), New York, 1998. ACM.

[KP90] JeanKyung Kim and David Pollard. Cube root asymptotics. *Ann. Statist.*, 18(1):191–219, 1990.

[MBG02] Llew Mason, Peter L. Bartlett, and Mostefa Golea. Generalization error of combined classifiers. *J. Comput. System Sci.*, 65(2):415–438, 2002.

[MvdG03] Leila Mohammadi and Sara 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.

[PvZ96] Hein Putter and Willem R. van Zwet. Resampling: consistency of substitution estimators. *Ann. Statist.*, 24(6):2297–2318, 1996.

[SFBL98] Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann. Statist.*, 26(5):1651–1686, 1998.

[SS98] Robert E. Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (Madison, WI, 1998)*, pages 80–91 (electronic), New York, 1998. ACM.

[vdG00] Sara A. van de Geer. *Applications of empirical process theory.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2000.

[vdVW96] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes.* Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

University of Leiden,
P.O. Box 9512,
2300 RA Leiden,
The Netherlands.
e-mails: merkl@math.leidenuniv.nl, leila@math.leidenuniv.nl