Report 2006-002

Penalized empirical risk minimalization

Leila Mohammadi

ISSN 1389-2355

January 4, 2006

Penalized empirical risk minimization

Leila Mohammadi

EURANDOM, P.O.Box 513, 5600 MB Eindhoven, The Netherlands

mohammadi@eurandom.tue.nl

Abstract

In a two-category classification problem labeled by $Y = 1$ and $Y = -1$, we observe a covariate, or feature, $X \in \mathcal{X} \subset \mathbb{R}^d$. We first consider a general loss function and a general penalty and obtain an upper bound for the penalized-risk of the penalized empirical risk minimizer. As an example, we consider the one dimensional case $d = 1$. Let V_0 denote the set on which the label $Y = 1$ is more likely. In the case that V_0 is the union of disjoint intervals and the number of intervals is unknown, we penalize the empirical risk by the number of thresholds, i.e., endpoints of the intervals, and obtain the rate of convergence of the risk minimizer. As another example, we consider the L_1 loss and the total variation penalty. Later we are concerned with obtaining lower bounds in a general setup. As an example, we show that the rate of convergence of the penalized empirical risk minimizer in the threshold estimation problem is optimal within a log factor.

2000 Mathematics Subject Classification: Primary 62G05, secondary 62G20.
Key words: Classification theory, nonparametric models, estimation error, penalization, optimal lower bounds, asymptotic distribution, threshold-based classifiers.

1 Introduction

Adaptivity and penalized risk minimization have recently been developed in a large variety of classification and regression problems, see for example [vdG01], [SN04], [BSR05], [TvdG05] and [SN03]. In this regard, a complex model class of functions or parameters is considered and a penalty on this class is added to the loss function to avoid overfitting.

In [vdG01], a penalized least squares estimator is studied where the penalty is on the complexity of the class of regression functions. It is shown that a rate of convergence for the penalized least squares estimator is determined by the entropy of the sets of regression functions with bounded penalized risks.

Our first result in this paper is related to [vdG01] as we use a similar procedure in classification problems. Instead of the squared error, we consider a general loss function with the penalty on the complexity of the class of classifiers. Similar to [vdG01], we find an upper bound on the probability of large errors.

Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be independent random variables, each with a distribution P_i, where $X_i \in \mathcal{X} \subset \mathbb{R}^d$ is called a feature and $Y_i \in \{-1, 1\}$ is the label of X_i. A classifier h is a function $h : \mathcal{X} \to [-1, 1]$, attaching the label $\text{sign}(h(X))$ to the feature X.

Mohammadi and van de Geer [MvdG05] consider the i.i.d case and give an application of the cube root asymptotics derived by Kim and Pollard [KP90]. In the case where $\mathcal{X}$ is one-dimensional, the set on which the label 1 is more likely is a union of disjoint intervals, and the boundaries of the intervals are estimated. These boundaries are called thresholds (a simple case, with just one threshold, has been presented in Mohammadi and van de Geer [MvdG03]). They establish the asymptotic distributions of estimators of the thresholds, using the set of classifiers with K thresholds as model class with fixed K smaller than or equal to the number of thresholds of the Bayes classifier. Their result is under the assumption that $F_0(x) = P(Y = 1|X = x)$ is differentiable.

Consider the setup of [MvdG05] with the difference that a bound on the number of thresholds is not known. The class of base classifiers will therefore be very large and it is reasonable to add a penalty on the number of thresholds to the risk function. It is shown in this paper that a penalty of this kind yields an estimator that converges to the Bayes rule with rate $(\log n)/n$, provided $2F_0 - 1$ has a jump at its sign change points. We also show that this rate is optimal within a log factor.

In Section 2, we consider a general setup in penalized empirical risk minimization, that covers the case of threshold estimation. Then, we apply the result to two examples. As an example in Section 3, we consider again the setup of [MvdG05] and penalize the empirical risk by the number of thresholds and minimize the penalized risk. Our main result is that the empirical risk minimizer converges to the minimizer of the prediction error with rate $(\log n)/n$. An example with total variation penalty is presented in Section 4.

In Section 5, we deal with lower bounds. We consider again, in Subsection 5.1, a general estimation problem and find two kinds of lower bounds on the error of estimators. As an example in Subsection 5.2, we consider the threshold estimation problem and show that the rate we obtained is optimal up to log factors.

Asymptotics are considered as $n \to \infty$, viewing the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ as the first n of an infinite sequence of independent but not necessarily identically distributed random variables. 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.

2 A general penalized risk minimization

Let $\mathcal{H}$ be a class of classifiers $h : \mathcal{X} \to [-1, 1]$. Consider a loss function $\ell : \mathbb{R} \to \mathbb{R}^+$. For $n \in \mathbb{N}$ and $i = 1, \ldots, n$, consider independent random variables (X_i, Y_i) from a distribution P_i. Set $\mathbf{X} = (X_1, \ldots, X_n)$ and define L_n and L as

$$L_n(h) := \frac{1}{n}\sum_{i=1}^{n} \ell(-Y_i h(X_i)), \quad L_{\mathbf{X}}(h) := \frac{1}{n}\sum_{i=1}^{n} E(\ell(-Y_i h(X_i)) \,|\, X_i),$$

$$L(h) := E(\ell(-Yh(X))).$$

Choose a penalty p, which is a non-negative function of $h \in \mathcal{H}$. Let

$$\hat{h}_n := \arg\min_{h \in \mathcal{H}} (L_n(h) + p^2(h))$$

and suppose

$$h_0 := \arg\min_{h \in \mathcal{H}} L(h)$$

exists. Let

$$\tau^2(h|\tilde{h}) := L_{\mathbf{X}}(h) - L_{\mathbf{X}}(\tilde{h}) + p^2(h)$$

and

$$\mathcal{H}_0(\delta) := \{h \in \mathcal{H} : \tau^2(h|h_0) \leq \delta^2\}.$$

Note that τ^2 and $\mathcal{H}_0$ depend on $\mathbf{X}$. Set

$$U_i(h) := \ell(-Y_i h(X_i)) - E(\ell(-Y_i h(X_i)) \mid X_i).$$

Assume that $d_1, ..., d_n$ are some metrics on $\mathcal{H}$, $d^2 := \frac{1}{n}\sum_{i=1}^n d_i^2$, and

$$|U_i(h) - U_i(\tilde{h})| \leq |W_i| d_i(h, \tilde{h}), \; i = 1, \ldots, n, \; h, \tilde{h} \in \mathcal{H},$$

where $W_1, \ldots, W_n$ are uniformly sub-Gaussian, so that for an M and σ_0^2, the following is satisfied

$$\max_{i=1,\ldots,n} M^2(E(\exp[|W_i|^2/M^2]) - 1) \leq \sigma_0^2. \tag{2.1}$$

Definition Let T be a (subset of a) metric space endowed with a metric m. The u-covering number $N(u, T; m)$ is defined as the number of balls with radius u necessary to cover T with respect to the metric m. The u-entropy is defined as $H(u, T; m) := \log N(u, T; m)$.

For $h \in \mathcal{H}$, let $v_n(h) := L_n(h) - L_{\mathbf{X}}(h)$. A maximal inequality for the empirical process $v_n(h_0) - v_n(h)$ is obtained in the following lemma.

Lemma 2.1 *Suppose all the above assumptions hold and let* $\sup_{h \in \mathcal{H}} d(h, h_0) \leq R$. *Then, for some* c_1 *depending on* M *and* σ_0, *and for all* $\delta > 0$ *satisfying*

$$\sqrt{n}\delta \geq c_1 \left(\int_0^R H^{1/2}(u, \mathcal{H}; d) du \vee R \right),$$

we have

$$\mathbf{P}\left(\sup_{h \in \mathcal{H}} |v_n(h_0) - v_n(h)| \geq \delta \mid \mathbf{X} \right) \leq c_1 \exp\left[-\frac{n\delta^2}{c_1^2 R^2} \right].$$

Proof Fix $\mathbf{X} = (X_1, \ldots, X_n)$. We have

$$|v_n(h_0) - v_n(h)|$$

$$= \left|\frac{1}{n}\sum_{i=1}^n [\ell(-Y_i h_0(X_i)) - E(\ell(-Y_i h_0(X_i)) \mid X_i) - \ell(-Y_i h(X_i)) + E(\ell(-Y_i h(X_i)) \mid X_i)]\right|$$

$$= \left|\frac{1}{n}\sum_{i=1}^n (U_i(h_0) - U_i(h))\right|.$$

The result follows from Lemma 8.5 of [vdG00], defining $\sigma := \infty$. □

Denote the observed value of $\mathbf{X}$ by $\mathbf{x}$. Assume that the following assumption holds:

Assumption (A) There are $\eta > 0$ and $k > 1$, such that

$$L_{\mathbf{x}}(h) - L_{\mathbf{x}}(h_0) \geq \eta d^k(h, h_0), \quad \forall h \in \mathcal{H}, \quad \forall \mathbf{x} = (x_1, ..., x_n) \in \mathcal{X}^n. \tag{2.2}$$

Assumption (A) is called the identifiability condition. This assumption can be seen in recent developments related to classification and statistics. Massart [Mas00] has used similar conditions in some applications of concentration inequalities to statistics. A discussion of this condition can be found in Tsybakov [Tsy04]. In Tsybakov and van de Geer [TvdG05] it is called the margin condition. See also Bartlett et al. [BJM03] and Blanchard et al. [BLV03].

The following theorem relates the speed of estimation of $\hat{h}_n$ to the entropy of $\mathcal{H}_0$. We use the method of van de Geer [vdG01].

Theorem 2.2 *Suppose Assumption (A). Let*

$$\Psi(\delta) \geq \int_0^\delta H^{1/2}(u, \mathcal{H}_0(\delta); d)du \vee \delta,$$

where Ψ does not depend on $\mathbf{X}$ and assume that $\Psi(\delta^{2/k})/\delta^2$ is a non-increasing function of δ, $\delta > 0$. Take $\epsilon \in (0,1)$. Then for the constant c_1 from Lemma 2.1, for some $c_2 \geq 4c_1$, for

$$\sqrt{n}\delta_n^2\epsilon \geq c_2\Psi((\frac{\delta_n^2}{\eta})^{1/k}), \tag{2.3}$$

and for all $\delta \geq \delta_n$, one has

$$\mathbf{P}(\tau^2(\hat{h}_n|h_0) \geq \frac{p^2(h_0)}{1-\epsilon} + \delta^2) \leq c_2 \exp\left[-\frac{n\epsilon^2\delta^{4(1-1/k)}\eta^{2/k}}{c_2^2}\right].$$

Proof We use

$$L_n(\hat{h}_n) + p^2(\hat{h}_n) \leq L_n(h_0) + p^2(h_0)$$

or

$$L_{\mathbf{X}}(\hat{h}_n) - L_{\mathbf{X}}(h_0) + p^2(\hat{h}_n) \leq [L_n(h_0) - L_{\mathbf{X}}(h_0) - (L_n(\hat{h}_n) - L_{\mathbf{X}}(\hat{h}_n))] + p^2(h_0)$$

or

$$\tau^2(\hat{h}_n|h_0) - p^2(h_0) \leq v_n(h_0) - v_n(\hat{h}_n).$$

We obtain

$$\begin{aligned} &\mathbf{P}(\tau^2(\hat{h}_n|h_0) \geq \frac{p^2(h_0)}{1-\epsilon} + \delta^2 \,|\, \mathbf{X}) \\ &\leq \mathbf{P}(|v_n(h_0) - v_n(\hat{h}_n)| \geq \epsilon\tau^2(\hat{h}_n|h_0) \text{ \& } \tau^2(\hat{h}_n|h_0) \geq \delta^2 \,|\, \mathbf{X}) \\ &\leq \mathbf{P}(\exists s \geq 1 : |v_n(h_0) - v_n(\hat{h}_n)| \geq \epsilon\tau^2(\hat{h}_n|h_0) \text{ \& } 2^{2s-2}\delta^2 \leq \tau^2(\hat{h}_n|h_0) < 2^{2s}\delta^2 \,|\, \mathbf{X}) \\ &\leq \sum_{s=1}^\infty \mathbf{P}\Bigg(\sup_{h\in\mathcal{H}_0(2^s\delta)} |v_n(h_0) - v_n(h)| \geq \epsilon 2^{2s-2}\delta^2 \,|\, \mathbf{X}\Bigg). \end{aligned}$$

If $h \in \mathcal{H}_0(2^s\delta)$, then, $d(h, h_0) \leq (\frac{2^{2s}\delta^2}{\eta})^{1/k}$, by Assumption (A). By Lemma 2.1, we know that for $\sqrt{n}r \geq c_1\Psi((\frac{2^{2s}\delta^2}{\eta})^{1/k})$, where c_1 depends on M, one has

$$\mathbf{P}\Bigg(\sup_{h\in\mathcal{H}_0(2^s\delta)} |v_n(h_0) - v_n(h)| \geq r \,|\, \mathbf{X}\Bigg) \leq c_1 \exp\left[-\frac{nr^2}{c_1^2}(\frac{\eta}{2^{2s}\delta^2})^{2/k}\right]. \tag{2.4}$$

Now, take δ_n as in (2.3) and $\delta \geq \delta_n$. Note that $\delta 2^s/\sqrt{\eta} \geq \delta_n/\sqrt{\eta}$ and hence

$$\frac{\Psi((2^s\delta/\sqrt{\eta})^{2/k})}{(2^s\delta/\sqrt{\eta})^2} \leq \frac{\Psi((\delta_n/\sqrt{\eta})^{2/k})}{(\delta_n/\sqrt{\eta})^2} \leq \epsilon\eta\sqrt{n}/c_2.$$

Thus,

$$c_1\Psi((2^s\delta/\sqrt{\eta})^{2/k}) \leq c_1\sqrt{n}2^{2s}\delta^2\epsilon/c_2 \leq \sqrt{n}2^{2s-2}\delta^2\epsilon,$$

for all $s \in \{1, 2, \ldots\}$. Now, using (2.4) one has

$$\begin{aligned}
\mathbf{P}(\tau^2(\hat{h}_n|h_0) &\geq \frac{p^2(h_0)}{1-\epsilon} + \delta^2 \,|\, \mathbf{X}) \\
&\leq \sum_{s=1}^{\infty} c_1 \exp\Big[-\frac{n\epsilon^2 2^{4s-4}\delta^4}{c_1^2}(\frac{\eta}{2^{2s}\delta^2})^{2/k}\Big] \\
&\leq \sum_{s=1}^{\infty} c_1 \exp\Big[-\frac{n\epsilon^2 2^{4(s-1-s/k)}\delta^{4(1-1/k)}\eta^{2/k}}{c_1^2}\Big] \\
&\leq c_2 \exp\Big[-\frac{n\epsilon^2\delta^{4(1-1/k)}\eta^{2/k}}{c_2^2}\Big],
\end{aligned}$$

for a choice of c_2. The result follows by integrating out. □

Here is a simple consequence of Theorem 2.2.

Corollary 2.3 *Under the conditions of Theorem 2.2, for $k = 2$, we arrive at the inequality*

$$E(\tau^2(\hat{h}_n|h_0)) \leq \frac{p^2(h_0)}{1-\epsilon} + \delta_n^2 + \frac{c_3}{\epsilon^2\eta n}, \quad \forall\epsilon \in (0, 1),$$

where c_3 is a constant depending on M.

For a proof, see Lemma 2.2 of van de Geer [vdG01].

3 An application: threshold estimation

In this section, we consider the one dimensional case $d = 1$ and the indicator loss function $\ell(t) := \mathbb{1}(t > 0)$. We take $\mathcal{X} = [0, 1]$ and assume that the classifier h just takes two values -1 and 1. Note that in this case our choice ℓ is sufficient to cover any loss function. We consider the i.i.d case, so $P_i = P$ for all $i \in \mathbb{N}$. The theoretical error of a classifier h is defined as

$$E(\mathbb{1}(Yh(X) < 0)) = P(h(X) \neq Y).$$

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

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

Let

$$F_0(x) := P(Y = 1|X = x) \tag{3.1}$$

be the conditional probability of the label $Y = 1$ if the feature X has value x.

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

$$U_K := \{a = (a_1, \ldots, a_K) \in [0, 1]^K : a_1 < \ldots < a_K\}. \tag{3.2}$$

Set 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 $b_{k+1} = -b_k$, $k = 2, \ldots, K$. Let $\mathcal{H}_1$ be the collection of classifiers

$$\mathcal{H}_1 = \mathcal{H}_1(K) := \{h_a : a \in U_K\} \tag{3.3}$$

and

$$\mathcal{H} := \cup_{K=1}^{\infty} \mathcal{H}_1(K). \tag{3.4}$$

Define

$$\hat{a}_n := \arg\min_{h_a \in \mathcal{H}} P_n(Y_i \ne h_a(X_i)), \quad a_0 := \arg\min_{h_a \in \mathcal{H}} P(Y \ne h_a(X)).$$

Because $\mathcal{H}$ is too rich, for any sample $(X_i, Y_i), i = 1, \ldots, n$, we can find $\hat{a}_n$ such that $P_n(Y_i \ne h_{\hat{a}_n}(X_i)) = 0$, we say overfitting occurs. By penalizing the empirical risk (on K), we in fact overrule the variance term and prevent overfitting.

One has,

$$\ell(-Yh(X)) = \mathbb{1}(h(X) < 0)\mathbb{1}(Y = 1) + \mathbb{1}(h(X) > 0)\mathbb{1}(Y = -1)$$

$$= Y\mathbb{1}(h(X) < 0) + \mathbb{1}(Y = -1).$$

Fix $\mathbf{X} = (X_1, \ldots, X_n)$ and set $q_i := \mathbb{1}(Y_i = -1) - P(Y_i = -1)$. Then,

$$U_i(h) = (Y_i - E(Y_i \,|\, X_i))\mathbb{1}(h(X_i) < 0) + q_i = W_i \phi(h(X_i)) + q_i,$$

where

$$W_i := Y_i - E(Y_i \,|\, X_i), \quad \phi(t) := \mathbb{1}(t < 0).$$

So,

$$|U_i(h) - U_i(\tilde{h})| = |W_i||\phi(h(X_i)) - \phi(\tilde{h}(X_i))|$$

and we take

$$d_i(h, \tilde{h}) = |\phi(h(X_i)) - \phi(\tilde{h}(X_i))| = \mathbb{1}(h(X_i) \ne \tilde{h}(X_i)).$$

On the other hand, condition (2.1) is satisfied with $M := 1$ and $\sigma_0^2 := e^4 - 1$. Note that

$$h_0(x) = \mathrm{sign}(E(Y \,|\, X)) = \mathrm{sign}(2F_0(x) - 1) \tag{3.5}$$

is the Bayes rule.

Suppose G, the distribution of X, has a bounded density g, write

$$a = (a_1, \ldots, a_K), \quad a_0 = (a_{0,1}, \ldots, a_{0,K})$$

and assume that the following holds:

Assumption (B) There is an $\eta > 0$, such that

$$|2F_0(x) - 1| \geq \eta, \quad \forall x \in (0,1), \quad x \neq a_{0,i}, \quad i = 1, \ldots, K.$$

Then for $V := \{h_a \neq h_{a_0}\}$ one has

$$\begin{aligned} L_{\mathbf{X}}(h_a) - L_{\mathbf{X}}(h_{a_0}) &= \frac{1}{n}\sum_{i=1}^{n} |E(Y_i \mid X_i)| \mathbb{1}(X_i \in V) \\ &= \frac{1}{n}\sum_{i=1}^{n} |2F_0(X_i) - 1| \mathbb{1}(X_i \in V) \\ &\geq \eta \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \in V) \\ &= \eta d^2(h_a, h_{a_0}). \end{aligned} \tag{3.6}$$

Now, condition (2.2) is satisfied with $k = 2$ and if $h \in \mathcal{H}_0(\delta)$, then $d(h_a, h_{a_0}) \leq \eta^{-1/2}\delta$.

Since we do not want the dimension of the parameter to be large, we consider a penalty on the dimension. Let $h = h_a \in \mathcal{H}$, $K = K_a$, and

$$p^2(h_a) := \lambda^2 \frac{K_a}{n}. \tag{3.7}$$

We need the entropy of a class of functions with respect to the L_1 metric.

Lemma 3.1 *Consider the L_1 metric on $\mathbb{R}^m$,*

$$\tilde{d}(a,b) := \sum_{i=1}^{m} |a_i - b_i|, \quad a = (a_1, \ldots, a_m), \quad b = (b_1, \ldots, b_m).$$

Then a ball $\tilde{B}_m(R)$ (with respect to the above metric) in $\mathbb{R}^m$ can be covered by $N \leq (\lfloor \frac{4R}{u} \rfloor + 1)^m$ balls with radius u.

We now show that the penalty yields an estimator with an error bounded by a factor of $(\log n)/n$.

Theorem 3.2 *Suppose Assumption (B). Take $\lambda \geq c_4\sqrt{\log n}$, where c_4 is a large constant depending on η. Then for every $\epsilon \in (0,1)$*

$$E(\tau^2(\hat{h}_n | h_0)) \leq \frac{p^2(h_0)}{1-\epsilon} + \frac{c_3}{n\epsilon^2\eta}.$$

Proof It is enough to check the conditions of Theorem 2.2 for $k = 2$. First fix $\mathbf{X}$. For $K_a = K_b = K$, one has

$$d^2(h_a, h_b) = c_5 \sum_{i=1}^{n} |h_a(X_i) - h_b(X_i)|/n,$$

for some constant c_5. If $h \in \mathcal{H}_0(\delta)$, then $\lambda^2 \frac{K_a}{n} \leq \delta^2$ or $K_a \leq \lfloor \frac{n\delta^2}{\lambda^2} \rfloor =: K(\delta)$. There are at most n^K linear subspaces of dimension K in $\mathbb{R}^n$. By Lemma 3.1,

$$H(u, \mathcal{H}_0(\delta); d) \leq K(\delta) \log \frac{5\delta}{u} + K(\delta) \log n$$

$$\leq \frac{n\delta^2}{\lambda^2} \log \frac{5n\delta}{u}.$$

So,

$$\int_0^\delta H^{1/2}(u, \mathcal{H}_0(\delta); d) du \leq A_0 \frac{\sqrt{n}\delta^2}{\lambda} \sqrt{\log n} \vee \delta =: \Psi(\delta),$$

where

$$A_0 := 1 + \int_0^1 \sqrt{\log \frac{5}{u}} du.$$

If

$$\sqrt{n}\delta_n = c_2/(\epsilon\sqrt{\eta}), \quad c_4 \geq A_0 c_2/(\epsilon\sqrt{\eta}),$$

then,

$$\Psi(\delta_n/\sqrt{\eta}) = \delta_n/\sqrt{\eta}$$

and

$$\sqrt{n}\delta_n^2 \epsilon \geq c_2 \Psi(\delta_n/\sqrt{\eta}).$$

□

Corollary 3.3 *By Theorem 3.2, for each fixed $\epsilon \in (0,1)$, we obtain the rate $(\log n)/n$ for the convergence of the error. The best choice for ϵ in Theorem 3.2 will improve the constants but not the rate.*

4 Total variation penalty

Let us first refer to some results on total variation penalties. Least squares penalized regression estimates with total variation penalties are considered in [MvdG97]. These estimators are least squares splines with locally data adaptive placed knot points. The rates of convergence and pointwise limiting distributions are obtained in [MvdG97]. In [Por97] the same problem is discussed with L_1 loss instead of the least squares. See also [DK01].

In this section we apply Theorem 2.2 with total variation penalties to obtain a bound on the error of the penalized risk minimizer.

Let $\mathcal{X} = [0,1]$ and consider the class $\mathcal{H}$ of functions $h : \mathcal{X} \to [-1,1]$ with the derivatives of all orders. Again we consider the i.i.d. case. Set $\ell(t) := |1+t|$ and the penalty

$$p_m^2(h) := \lambda_m^2 TV(h^{(m-1)}), \ m \geq 1, \ 0 < \lambda_m \leq 1,$$

where TV is the total variation function. Note that

$$\begin{aligned} U_i(h) &= |1 - Y_i h(X_i)| - E(|1 - Y_i h(X_i)| \mid X_i) \\ &= (-Y_i + E(Y_i \mid X_i))h(X_i) = W_i \phi(h(X_i)), \end{aligned}$$

where

$$W_i := -Y_i + E(Y_i \mid X_i), \ \phi(t) := t.$$

So,

$$|U_i(h) - U_i(\tilde{h})| = |W_i||h(X_i) - \tilde{h}(X_i)|$$

and we take

$$d_i(h, \tilde{h}) = |\phi(h(X_i)) - \phi(\tilde{h}(X_i))| = |h(X_i) - \tilde{h}(X_i)|.$$

Condition (2.1) is now satisfied with $M := 1$ and $\sigma_0^2 := e^4 - 1$.

Theorem 4.1 *Suppose Assumption (B). One has for all $\epsilon \in (0,1)$*

$$E(\tau^2(\hat{h}_n|h_0)) \leq \lambda_m^2 TV(h_0^{(m-1)})/(1-\epsilon) + \frac{C_m}{n\epsilon^2\eta\lambda_m^{1/m}} + \frac{c_3}{n\epsilon^2\eta}, \tag{4.1}$$

where C_m is a constant.

Proof Set

$$\mathcal{H}_{m,C} := \{h \in \mathcal{H} : p_m^2(h) \leq C\}.$$

It is proved in [BS67] that

$$H_\infty(u, \mathcal{H}_{m,C}) \leq A_m(C/u)^{1/m}, \ \forall u > 0,$$

where H_∞ is the entropy corresponding to the supremum norm. We can show that

$$H(u, \mathcal{H}_{m,\delta/\lambda_m,\delta}) \leq B_m^2(\delta/(\lambda_m u))^{1/m}, \ 0 < u \leq \delta,$$

where

$$\mathcal{H}_{m,C,\delta} := \mathcal{H}_{m,C} \cap \{h : \|h - h_0\|_{\mathbf{x}} \leq \delta\},$$

and where $\|\cdot\|_{\mathbf{x}}$ is the L_2 norm, corresponding with the vector $\mathbf{x}$ (see Lemma 3.4 in [vdG01]). Note that h_0 is the same as in (3.5) and

$$\begin{aligned} L_{\mathbf{X}}(h) - L_{\mathbf{X}}(h_0) &= \frac{1}{n}\sum_{i=1}^{n}(E(1 - Y_i h(X_i) \mid X_i) - E(1 - Y_i h_0(X_i) \mid X_i)) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(Y_i \mid X_i)(h(X_i) - h_0(X_i)) \\ &= \frac{1}{n}\sum_{i=1}^{n} |E(Y_i \mid X_i)||h(X_i) - h_0(X_i)| \\ &\geq \eta\frac{1}{n}\sum_{i=1}^{n} |h(X_i) - h_0(X_i)| \end{aligned}$$

$$\geq \eta \frac{1}{n} \sum_{i=1}^{n} |h(X_i) - h_0(X_i)|^2$$
$$= \eta d^2(h, h_0).$$

Therefore, we apply Corollary 2.3 with $\Phi(\delta) := B_m \delta / \lambda_m^{1/(2m)}$ and take

$$\sqrt{n}\delta_n = \frac{c_2 B_m}{\sqrt{\eta}\epsilon \lambda_m^{1/(2m)}}.$$

We obtain (4.1) with $C_m := (c_2 B_m)^2$. □

Corollary 4.2 *If* $TV(h_0^{(m-1)})$ *remains bounded in* n*, then we may choose*

$$\lambda_m := n^{-m/(2m+1)}$$

to get the rate $n^{-2m/(2m+1)}$ *for the convergence of the error. Again the best choice for* ϵ *will improve the constants but not the rate.*

5 Optimal lower bounds

So far, we have studied upper bounds on the errors of estimators. In this section, we obtain some lower bounds on the errors in estimation problems. Lower bounds are in general not as interesting as upper bounds. However, if we have both upper and lower bounds and if they are of the same order, then we get optimality (for the exact definitions see below).

Lower bounds can be defined in different ways. Our special definitions are referred to [MM03], see Corollary 5.11 below (see also [Moh04], Chapter 6). In fact we first generalize the results in [MM03] and then consider the special case of threshold estimation.

5.1 General theory

In this subsection, we consider a general setup and obtain some lower bound type results. In the next subsection, we apply our theory to classification problems.

Consider a statistical model $\mathcal{P}$, i.e. a class of probability measures over a measurable space $(\Omega, \mathcal{A})$. Consider another measurable space $(\mathcal{H}, \mathcal{B}(\mathcal{H}))$ and a loss function

$$L : \mathcal{P} \times \mathcal{H} \to \mathbb{R}, \quad (P, h) \mapsto L_P(h). \tag{5.1}$$

Assume that for each $P \in \mathcal{P}$, there exists a minimizer $h_P \in \mathcal{H}$, i.e.

$$L_P(h_P) \leq L_P(h), \quad \text{for all } h \in \mathcal{H}. \tag{5.2}$$

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. The following lemma is Lemma 2.1 in [MM03].

Lemma 5.1 *Let* P *and* Q *be probability measures with* $H(P, Q) < \infty$*. For every random variable* X *with* $0 \leq X \leq 1$*, one has*

$$E_Q[X] \geq e^{-2H(P,Q)-1} \left(E_P[X] - \frac{1}{2} \right). \tag{5.3}$$

Set $\Delta_P(h) := L_P(h) - L_P(h_P)$. This function is called the excess risk. For $\gamma > 0$, set $\Delta_P^\gamma(h) := \min\{\gamma, \Delta_P(h)\}$. Our aim is to find lower bounds on the probability of $\Delta_P^\gamma(\hat{h}_n)$ being large for every estimator $\hat{h}_n : \Omega^n \to \mathcal{H}$. Fix $n \in \mathbb{N}$.

Lemma 5.2 *Let $P, Q \in \mathcal{P}$ and $\gamma > 0$ such that*

$$\Delta_P(h) + \Delta_Q(h) \geq \gamma, \quad \forall h \in \mathcal{H}. \tag{5.4}$$

Then for any $\delta \in (0, 1/2)$ and for any estimator $\hat{h}_n : \Omega^n \to \mathcal{H}$, at least one of the following two statements holds:

$$E_{P^n}(\Delta_P^\gamma(\hat{h}_n)) \geq \delta\gamma \tag{5.5}$$

or

$$E_{Q^n}(\Delta_Q^\gamma(\hat{h}_n)) \geq (\frac{1}{2} - \delta)\gamma \exp(-2nH(P,Q) - 1). \tag{5.6}$$

Proof Note that Δ_P^γ and Δ_Q^γ take values in $[0, \gamma]$. It is easily seen that $\Delta_P^\gamma(h) + \Delta_Q^\gamma(h) \geq \gamma$, for all $h \in \mathcal{H}$. Hence,

$$E_{P^n}(\Delta_P^\gamma(\hat{h}_n) + \Delta_Q^\gamma(\hat{h}_n)) \geq \gamma, \tag{5.7}$$

thus for any $\delta \in (0, 1/2)$, we have $E_{P^n}(\Delta_P^\gamma(\hat{h}_n)) \geq \delta\gamma$ or $E_{P^n}(\Delta_Q^\gamma(\hat{h}_n)) \geq (1 - \delta)\gamma$. By Lemma 5.1, we know that

$$\frac{1}{\gamma} E_{Q^n}(\Delta_Q^\gamma(\hat{h}_n)) \geq \frac{1}{\gamma}\left(E_{P^n}(\Delta_Q^\gamma(\hat{h}_n)) - \frac{\gamma}{2}\right) \exp(-2nH(P,Q) - 1). \tag{5.8}$$

So, if $E_{P^n}(\Delta_Q^\gamma(\hat{h}_n)) \geq (1 - \delta)\gamma$, then

$$E_{Q^n}(\Delta_Q^\gamma(\hat{h}_n)) \geq ((1 - \delta)\gamma - \frac{\gamma}{2}) \exp(-2nH(P,Q) - 1) \tag{5.9}$$

$$= (1/2 - \delta)\gamma \exp(-2nH(P,Q) - 1). \tag{5.10}$$

□

Lemma 5.2 is the first step to prove lower rate of convergence. We use the following immediate consequence of this lemma.

Lemma 5.3 *Let $P, Q \in \mathcal{P}$ and $\gamma > 0$ with $\inf_{h \in \mathcal{H}}(\Delta_P(h) + \Delta_{Q_n}(h)) \geq \gamma$ and $H(P,Q) < \infty$. Let $n \in \mathbb{N}$, then for any estimator $\hat{h} : \Omega^n \to \mathcal{H}$, we have*

$$\max_{\mu \in \{P,Q\}} E_{\mu^n}(\Delta_\mu^\gamma(\hat{h})) \geq \frac{1}{4} \exp(-2nH(P,Q) - 1)\gamma. \tag{5.11}$$

Proof By Lemma 5.2,

$$E_{P^n}(\Delta_P^\gamma(\hat{h})) \geq \frac{\gamma}{4}, \text{ or } E_{Q^n}(\Delta_Q^\gamma(\hat{h})) \geq \frac{\gamma}{4} \exp(-2nH(P,Q) - 1). \tag{5.12}$$

Thus, using $\exp(2nH(P,Q)+1) \geq 1$,

$$\max_{\mu\in\{P,Q\}} \frac{\exp(2nH(P,Q)+1)}{\gamma} E_{\mu^n}(\Delta_\mu^\gamma(\hat{h})) \geq 1/4. \tag{5.13}$$

Hence, (5.11) holds. □

A lower bound can be derived from (5.11) by considering γ dependent on n. The following theorem is the first lower bound result. It is proved by (5.11).

Theorem 5.4 *Let $(\hat{h}_n : \Omega^n \to \mathbb{R})_{n\in\mathbb{N}}$ be a sequence of estimators. Take sequences $(\delta_n)_{n\in\mathbb{N}}$ and $(\gamma_n)_{n\in\mathbb{N}}$ of positive numbers and a non-empty 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*

$$H^2(P_n, Q_n) \leq \delta_n \quad \textit{and} \quad \inf_h(\Delta_{P_n}^{\gamma_n}(h) + \Delta_{Q_n}^{\gamma_n}(h)) \geq \gamma_n. \tag{5.14}$$

One has for all non-empty open sets $\mathcal{U} \subseteq \mathcal{P}$

$$\liminf_n \sup_{P\in\mathcal{U}} \frac{E_{P^n}(\Delta_P^{\gamma_n}(\hat{h}_n))}{\gamma_n \exp(-2n\delta_n)} \geq \frac{1}{4\mathrm{e}}. \tag{5.15}$$

The second lower bound result in this paper is presented in the theorem below. Consider positive sequences $(\gamma_n)_{n\in\mathbb{N}}$ and $(\delta_n)_{n\in\mathbb{N}}$. Define the total variation distance

$$(P,Q) \mapsto \|P-Q\|_{\mathcal{A}} := \sup_{A\in\mathcal{A}}(P(A) - Q(A)).$$

Theorem 5.5 *Let $(\hat{h}_n : \Omega^n \to \mathbb{R})_{n\in\mathbb{N}}$ be a sequence of estimators. Let $\mathcal{P}$ be endowed with a Baire space metrizable topology. Assume that the map $\mathcal{P} \ni P \mapsto \Delta_P(h)$ is continuous for all $h \in \mathcal{H}$ with respect to the total variation distance.*

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

$$H(P, Q_n) \leq \delta_n \quad \textit{and} \quad \inf_h(\Delta_P^{\gamma_n}(h) + \Delta_{Q_n}^{\gamma_n}(h)) \geq \gamma_n. \tag{5.16}$$

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

$$\sup_{\mu\in\mathcal{U}} \limsup_{n\to\infty} \frac{E_{\mu^n}(\Delta_\mu^{\gamma_n}(\hat{h}_n))}{\gamma_n \exp(-2n\delta_n)} \geq \frac{1}{4\mathrm{e}}. \tag{5.17}$$

Proof Set

$$\mathcal{P}_n = \{P \in \mathcal{P} : E_{P^n}(\Delta_P^{\gamma_n}(\hat{h}_n)) \leq \frac{1}{4\mathrm{e}}\gamma_n \exp(-2n\delta_n)\} \tag{5.18}$$

and

$$\mathcal{F}_m := \bigcap_{n\geq m} \mathcal{P}_n. \tag{5.19}$$

We claim that the map

$$\mathcal{P} \to [0,1], \quad P \mapsto E_{P^n}(\Delta_P^{\gamma_n}(\hat{h}_n)) \tag{5.20}$$

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

$$\Delta_{Q_k}^{\gamma_n}(\hat{h}_n(\omega)) \overset{k\to\infty}{\longrightarrow} \Delta_P^{\gamma_n}(\hat{h}_n(\omega)) \tag{5.21}$$

by the continuity of $P \mapsto \Delta_P(\omega)$. Using Lebesgue's dominated convergence theorem and $0 \leq \Delta_P^{\gamma_n}(\hat{h}_n) \leq \gamma_n$, this implies that for each n

$$E_{P^n}[\Delta_{Q_k}^{\gamma_n}(\hat{h}_n)] \overset{k\to\infty}{\longrightarrow} E_{P^n}[\Delta_P^{\gamma_n}(\hat{h}_n)]. \tag{5.22}$$

Furthermore,

$$|E_{Q_k^n}[\Delta_{Q_k}^{\gamma_n}(\hat{h}_n)] - E_{P^n}[\Delta_{Q_k}^{\gamma_n}(\hat{h}_n)]| \leq \gamma_n \|Q_k - P\|_{\mathcal{A}} \overset{k\to\infty}{\longrightarrow} 0. \tag{5.23}$$

Combining (5.22) and (5.23), we get

$$E_{Q_k^n}[\Delta_{Q_k}^{\gamma_n}(\hat{h}_n)] \overset{k\to\infty}{\longrightarrow} E_{P^n}[\Delta_P^{\gamma_n}(\hat{h}_n)], \tag{5.24}$$

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

The continuity of the map described in (5.20) 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 (5.16) holds. Then Lemma 5.3 implies (5.11), 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

$$\limsup_{n\to\infty} \frac{1}{\gamma_n \exp(-2n\delta_n)} E_{P^n}[\Delta_P^{\gamma_n}(\hat{h}_n)] \geq \frac{1}{4\mathrm{e}} \tag{5.25}$$

and therefore,

$$\sup_{\mu\in\mathcal{U}} \limsup_{n\to\infty} \frac{1}{\gamma_n \exp(-2n\delta_n)} E_{\mu^n}(\Delta_\mu^{\gamma_n}(\hat{h}_n)) \geq \frac{1}{4\mathrm{e}}. \tag{5.26}$$

□

We use the following version of Chebyshev's theorem.

Lemma 5.6 *Consider a random variable $X \leq 1$ with probability measure P. Then, for any $c_7 < E_P(X)$,*

$$P(X \leq c_7) \leq \frac{1 - E_P(X)}{1 - c_7}. \tag{5.27}$$

Proof We have

$$E_P(X) = E_P(X\mathbb{1}(X > c_7)) + E_P(X\mathbb{1}(X \le c_7)) \tag{5.28}$$

$$\le P(X > c_7) + c_7 P(X \le c_7) \tag{5.29}$$

$$= 1 - P(X \le c_7)(1 - c_7). \tag{5.30}$$

Hence, (5.27) holds. □

We have the following corollary from Theorems 5.4 and 5.5 and Lemma 5.6.

Corollary 5.7 *Under the conditions of Theorem 5.5, for any $c_7 < \frac{1}{4e}$ one has*

$$\sup_{\mu\in\mathcal{U}} \limsup_{n\to\infty} \mu^n(\Delta_\mu^{\gamma_n}(\hat{h}_n) > c_7\gamma_n \exp(-2n\delta_n))$$

$$\ge \frac{\sup_{\mu\in\mathcal{U}} \limsup_{n\to\infty} \frac{E_{\mu^n}(\Delta_\mu^{\gamma_n}(\hat{h}_n))}{\gamma_n \exp(-2n\delta_n)} - c_7}{1 - c_7} \tag{5.31}$$

$$\ge \frac{\frac{1}{4e} - c_7}{1 - c_7}. \tag{5.32}$$

Similarly, under the conditions of Theorem 5.4, for any $c_7 < \frac{1}{4e}$ one has

$$\liminf_n \sup_{\mu\in\mathcal{U}} \mu^n(\Delta_\mu^{\gamma_n}(\hat{h}_n) > c_7\gamma_n \exp(-2n\delta_n)) \ge \frac{\frac{1}{4e} - c_7}{1 - c_7}. \tag{5.33}$$

5.2 Optimality in threshold estimation

In this subsection, we prove that the rate in the threshold estimation problem obtained in Section 3, is optimal up to logrithmic factors. Our example is even more general so that the threshold estimation is a very specail case.

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

$$P = f_P[\lambda_{[0,1]} \otimes (\text{counting measure})] \tag{5.34}$$

on $\Omega = [0, 1] \times \{\pm 1\}$ (with the Borel σ algebra $\mathcal{A}$). Here, $\lambda_{[0,1]}$ denotes the Lebesgue measure on $[0, 1]$.

We endow $\tilde{\mathcal{P}}$ with the total variation metric d

$$d(P, Q) := \|P - Q\|_{\mathcal{A}}. \tag{5.35}$$

Let $\mathcal{P} \subseteq \tilde{\mathcal{P}}$ denote the set of all $P \in \tilde{\mathcal{P}}$, such that for some $a(P) \in (0, 1)$, $\theta = \theta_P \ge 0, \epsilon = \epsilon_P > 0$ and some constants $c_8, c_9, c_{10} \in (0, \infty)$, $c_8 \le c_9$, one has

$$f_P^{\pm}(x) \ge c_{10} \tag{5.36}$$

and

$$c_8\,|\,X - a(P)|^\theta \le |f_P^+(x) - f_P^-(x)| \le c_9\,|\,X - a(P)|^\theta, \tag{5.37}$$

for all $x \in I_\epsilon := (a(P) - \epsilon, a(P) + \epsilon)$.

We take $\mathcal{H}$ as in (3.4),

$$L_P(h) := P(Y \ne h(X)) \tag{5.38}$$

and

$$V^h := \{x : h(x) = 1\}.$$

The following lemma provides sufficient conditions for the lower bounds.

Lemma 5.8 *Let $P \in \mathcal{P}$ and $\mathcal{U} \subseteq \mathcal{P}$ be an open neighbourhood of P. Then there are $c_{11} := c_{11}(P,\mathcal{U}) > 0$ and $c_{12} := c_{12}(P,\mathcal{U}) > 0$, such that for all large n (say for $n \ge n_0(P,\mathcal{U})$), there is $Q_n \in \mathcal{U}$, such that the following holds*

$$nH(P,Q_n) \le c_{11} \quad \text{and} \quad \inf_h(\Delta_P^{\gamma_n}(h) + \Delta_{Q_n}^{\gamma_n}(h)) \ge \gamma_n, \tag{5.39}$$

where $\gamma_n := c_{12} n^{-\frac{1+\theta}{1+2\theta}}$.

Proof Choose a ball $\mathcal{N} \subseteq \mathcal{U}$, with respect to the metric d, centered at P with the radius r. Consider $\epsilon_n := n^{-\frac{1}{2\theta+1}}$. Define the densities

$$f_{Q_n}^{\pm}(x) := f_P^{\pm}(x)\mathbb{1}(x \notin I_{\epsilon_n}) + f_P^{\mp}(x)\mathbb{1}(x \in I_{\epsilon_n}). \tag{5.40}$$

Note that for $x \in [0,1]$,

$$|f_{Q_n}^+(x) - f_{Q_n}^-(x)| = |f_P^+(x) - f_P^-(x)|.$$

On the other hand

$$\begin{aligned}\int |f_{Q_n}^+ - f_P^+|dx &= \int_{I_{\epsilon_n}} |f_P^- - f_P^+|dx \\ &\le \int_{I_{\epsilon_n}} c_9\,|\,X - a(P)|^\theta dx \\ &= c_9 \int_{-\epsilon_n}^{\epsilon_n} |y|^\theta dx \\ &= 2c_9\epsilon_n^{\theta+1}/(\theta+1) = 2(c_9/(\theta+1))n^{-\frac{\theta+1}{2\theta+1}}.\end{aligned} \tag{5.41}$$

Similarly,

$$\int |f_{Q_n}^- - f_P^-|dx \le 2(c_9/(\theta+1))n^{-\frac{\theta+1}{2\theta+1}}.$$

It shows that $d(P,Q_n) \le r$, if we take n large enough. Also, for large n, $f_{Q_n}^{\pm} \ge c_{10}$ over I_{ϵ_n} and hence, $Q_n \in \mathcal{N}$. From now on, we suppose that n is large enough. One has

$$\int (\frac{f_{Q_n}^+}{f_P^+} - 1)^2 f_P^+ dx \le \frac{1}{c_{10}} \int_{I_{\epsilon_n}} (c_9\,|\,X - a(P)|^\theta)^2 dx$$

$$\leq \frac{c_9^2}{c_{10}} \int_{-\epsilon_n}^{\epsilon_n} |y|^{2\theta} dx$$
$$\leq \frac{2c_9^2}{c_{10}(2\theta+1)} \epsilon_n^{2\theta+1}.$$

Similarly,

$$\int (\frac{f_{Q_n}^-}{f_P^-} - 1)^2 f_P^- dx \leq \frac{2c_9^2}{c_{10}(2\theta+1)} \epsilon_n^{2\theta+1}.$$

Hence,

$$nH(P, Q_n) \leq nE((\frac{dQ_n}{dP} - 1)^2) \leq \frac{4c_9^2}{c_{10}(2\theta+1)} n\epsilon_n^{2\theta+1}$$
$$= \frac{4c_9^2}{c_{10}(2\theta+1)} =: c_{11}.$$

On the other hand

$$\int |f_{Q_n}^+ - f_P^+| f_P^+ dx \geq \int_{I_{\epsilon_n}} c_8 \,|X - a(P)|^\theta f_P^+ dx$$
$$\geq c_8 c_{10} \int_{I_{\epsilon_n}} |X - a(P)|^\theta dx = c_8 c_{10} \int_{-\epsilon_n}^{\epsilon_n} |y|^\theta dx$$
$$\geq 2c_8 c_{10} \epsilon_n^{\theta+1}/(\theta+1) = \frac{2c_8 c_{10}}{\theta+1} n^{-\frac{\theta+1}{2\theta+1}}. \tag{5.42}$$

Take an arbitrary $h \in \mathcal{H}$. One has

$$\Delta_P(h) + \Delta_{Q_n}(h) = \int_{V^h \Delta V^{h_P}} |f_P^+ - f_P^-| dx + \int_{V^h \Delta V^{h_{Q_n}}} |f_{Q_n}^+ - f_{Q_n}^-| dx$$
$$\geq \int_{(V^h \Delta V^{h_P}) \cup (V^h \Delta V^{h_{Q_n}})} \min\{|f_P^+ - f_P^-|, |f_{Q_n}^+ - f_{Q_n}^-|\} dx$$
$$\geq \int_{V^{h_P} \Delta V^{h_{Q_n}}} \min\{|f_P^+ - f_P^-|, |f_{Q_n}^+ - f_{Q_n}^-|\} dx$$
$$\geq \int_{V^{h_P} \Delta V^{h_{Q_n}}} |f_P^+ - f_P^-| dx.$$

Note that over I_{ϵ_n},

$$(f_{Q_n}^+ - f_{Q_n}^-)(f_P^+ - f_P^-) < 0,$$

so that $I_{\epsilon_n} \subset V^{h_P} \Delta V^{h_{Q_n}}$. Hence, using (5.42),

$$\Delta_P(h) + \Delta_{Q_n}(h) \geq c_{12} n^{-\frac{1+\theta}{1+2\theta}},$$

where $c_{12} := \frac{2c_8 c_{10}}{\theta+1}$.

□

To prove the lower bounds, we need to show the continuity of the function Δ_P. It is shown in the next lemma.

Lemma 5.9 *For any $h \in \mathcal{H}$, the function $\mathcal{P} \ni P \mapsto \Delta_P(h) = L_P(h) - L_P(h_P)$ is continuous with respect to the total variation distance $\| \cdot \|_{\mathcal{A}}$.*

Proof We prove that both, $P \mapsto L_P(h)$ and $P \mapsto L_P(h_P)$ are continuous. Using that $L_P(h) = P(A)$ for the event $A = \{h(X) \neq Y\}$, continuity of $P \mapsto L_P(h)$ follows immediately from the definition of $\|\cdot\|_{\mathcal{A}}$.

For any finite measures $\mu_1, \dots, \mu_k$ over $\mathcal{X}$, define $\mu_1 \vee \dots \vee \mu_k$ by its Radon-Nikodym derivative

$$\frac{d(\mu_1 \vee \dots \vee \mu_k)}{d\nu} = \max\left(\frac{d\mu_1}{d\nu}, \dots, \frac{d\mu_k}{d\nu}\right), \tag{5.43}$$

where ν denotes any dominating measure for $\mu_1 \dots, \mu_k$. Note that this definition is independent of the choice of a dominating measure ν and of the choice of representatives of the Radon-Nikodym derivatives. Furthermore, the maximum operation $\vee$ is continuous with respect to the total variation distance.

Denote the probability distributions corresponding to $f_P^\pm$ by $P^\pm$. We get by (5.38)

$$\begin{aligned} 1 - L_P(h_P) &= P(h_P(X) = Y) = P^+(h_P(X) = 1) + P^-(h_P(X) = -1) \\ &= \int\limits_{\{h_P = 1\}} \frac{dP^+}{d\nu}\, d\nu + \int\limits_{\{h_P = -1\}} \frac{dP^-}{d\nu}\, d\nu = (P^+ \vee P^-)(\mathcal{X}). \end{aligned} \tag{5.44}$$

Since the maximum operation $\vee$ and the map $P \mapsto P^\pm$ are continuous with respect to the total variation distance, this implies that $P \mapsto L_P(h_P)$ is continuous. This proves our claim.

□

Corollary 5.10 *By Lemmata 5.5 and 5.8, we have the following. For all non-empty open sets $\mathcal{U} \subseteq \mathcal{P}$ and all estimators $\hat{h}_n$, one has for any $c_7 < \frac{1}{4e}$*

$$\begin{aligned} &\sup_{\mu \in \mathcal{U}} \limsup_{n \to \infty} \mu^n (L_P(\hat{h}_n) - L_P(h_P) \geq c_7 n^{-\frac{1+\theta}{1+2\theta}}) \\ &= \sup_{\mu \in \mathcal{U}} \limsup_{n \to \infty} \mu^n (\min\{L_P(\hat{h}_n) - L_P(h_P), c_7 n^{-\frac{1+\theta}{1+2\theta}}\} \geq c_7 n^{-\frac{1+\theta}{1+2\theta}}) \\ &= \sup_{\mu \in \mathcal{U}} \limsup_{n \to \infty} \mu^n (\Delta_\mu^{c_7 n^{-\frac{1+\theta}{1+2\theta}}}(\hat{h}_n) \geq c_7 n^{-\frac{1+\theta}{1+2\theta}}) \\ &\geq \frac{\frac{1}{4e} - c_7}{1 - c_7}. \end{aligned} \tag{5.45}$$

Similarly, by Lemma 5.9, for any $c_7 < \frac{1}{4e}$ one has

$$\liminf_n \sup_{\mu \in \mathcal{U}} \mu^n (L_P(\hat{h}_n) - L_P(h_P) \geq c_7 n^{-\frac{1+\theta}{1+2\theta}}) \geq \frac{\frac{1}{4e} - c_7}{1 - c_7}. \tag{5.46}$$

Corollary 5.11 *The case $\theta = 0$ corresponds to the threshold problem described in Section 3. Consider the setup introduced in Section 3 with Assumption (B). Let $P(Y = 1) = 1/2$. Consider the case that $2P(Y = 1 \mid X = x) - 1$ has a sign change $a(P) \in (0,1)$ and suppose that for an $\epsilon > 0$, the densities $f^\pm := P(\cdot|Y = \pm 1)$ are bounded from below by some $c_{10} \in (0, \infty)$ over $I_\epsilon = (a(P) - \epsilon, a(P) + \epsilon)$. Let g, the density of G, be bounded from below by c_{13} over I_ϵ. Define $c_8 := c_{13}\eta$ and*

$$c_9 := \sup_{x \in I_\epsilon} |f^+(x) - f^-(x)|.$$

Then (5.36) and (5.37) hold and hence $P \in \mathcal{P}$ and by Corollary 5.10, for all non-empty open sets $\mathcal{U} \subseteq \mathcal{P}$ and all estimators $\hat{h}_n$, one has for any $c_7 < \frac{1}{4e}$

$$\sup_{\mu \in \mathcal{U}} \limsup_{n \to \infty} \mu^n(L_P(\hat{h}_n) - L_P(h_P) \geq c_7(1/n)) \geq \frac{\frac{1}{4e} - c_7}{1 - c_7}.$$

Similarly, for any $c_7 < \frac{1}{4e}$ one has

$$\liminf_n \sup_{\mu \in \mathcal{U}} \mu^n(L_P(\hat{h}_n) - L_P(h_P) \geq c_7(1/n)) \geq \frac{\frac{1}{4e} - c_7}{1 - c_7}. \quad (5.47)$$

Corollary 5.12 *Upper bounds in the threshold estimation problem with known K for the case $\theta \geq 0$ is discussed in [MvdG05] (See Theorem 5 in that article). They consider, without loss of generality, $K = 1$ and the assumption*

$$|2P(Y = 1 \,|\, X = x) - 1| g(x) \geq c\,|\,x - a^*|^\theta,$$

for some $c > 0$, $\theta \geq 0$ and all x in a nighbourhood of a^*, where a^* is the threshold of the Bayes rule. It is shown that

$$P(Y\hat{h}_n(X) < 0) - P(Yh_{a^*}(X) < 0) = O_{\mathbf{P}}(n^{-\frac{1+\theta}{1+2\theta}}),$$

where $\hat{h}_n$ is the empirical risk minimizer and h_{a^*} is Bayes rule. Similar to Corollary 5.11, it can be shown that the rate $n^{-\frac{1+\theta}{1+2\theta}}$ is optimal for $\theta \geq 0$, for the case that the number of thresholds K is known.

Acknowledgment

I am extremely grateful to Prof. F. Merkl for his useful suggestions on Section 5.

References

[BJM03] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification and risk bounds. *Tech. Rep. 638, University of California at Berkeley*, 2003.

[BLV03] G. Blanchard, G. Lugosi, and N. Vayatis. On the rate of convergence of regularized boosting classifiers. *Journal of Machine Learning Research*, (4):861–894, 2003.

[BS67] M. Š. Birman and M. Z. Solomjak. Piecewise polynomial approximations of functions of classes $W_p{}^\alpha$. *Mat. Sb. (N.S.)*, 73 (115):331–355, 1967.

[BSR05] G. Blanchard, C. Schafer, and Y. Rozenhole. Oracle bounds and exact algorithm for dyadic classification trees. 2005.

[DK01] P. L. Davies and A. Kovac. Local extremes, runs, strings and multiresolution. *Ann. Statist.*, 29(1):1–65, 2001. With discussion and rejoinder by the authors.

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

[Mas00] P. Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math. (6)*, 9(2):245–303, 2000. Probability theory.

[MM03] F. Merkl and L. Mohammadi. Optimal third root asymptotic bounds in the statistical estimation of thresholds. *Tech. Rep. MI-11 University of Leiden*, June, 2003.

[Moh04] L. Mohammadi. *Estimation of thresholds in classification.* PhD thesis, University of Leiden, 2004.

[MvdG97] E. Mammen and S. A. van de Geer. Penalized quasi-likelihood estimation in partial linear models. *Ann. Statist.*, 25(3):1014–1035, 1997.

[MvdG03] 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.

[MvdG05] L. Mohammadi and S. A. van de Geer. Asymptotics in empirical risk minimization. *JMLR*, 6:2027–2047, 2005.

[Por97] S. Portnoy. Local asymptotics for quantile smoothing splines. *Ann. Statist.*, 25(1):414–434, 1997.

[SN03] C. Scott and R. Nowak. Near-minimax optimal classification with dyadic classification trees. Preprint, 2003.

[SN04] C. Scott and R. Nowak. Minimax-optimal classification with dyadic decision trees. Preprint, 2004.

[Tsy04] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004.

[TvdG05] 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.

[vdG00] S. A. van de Geer. *Empirical Processes in M-Estimation.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2000.

[vdG01] S. A. van de Geer. Least squares estimation with complexity penalties. *Math. Methods Statist.*, 10(3):355–374, 2001. Meeting on Mathematical Statistics (Marseille, 2000).