Universität Stuttgart

Fachbereich Mathematik

Improved Classification Rates under Refined Margin Conditions

Ingrid Blaschzyk, Ingo Steinwart

Preprint 2016/004

Fachbereich Mathematik
Fakultät Mathematik und Physik
Universität Stuttgart
Pfaffenwaldring 57
D-70 569 Stuttgart

E-Mail: preprints@mathematik.uni-stuttgart.de
WWW: http://www.mathematik.uni-stuttgart.de/preprints

ISSN **1613-8309**

Improved Classification Rates under Refined Margin Conditions

Ingrid Blaschzyk and Ingo Steinwart

Institute for Stochastics and Applications
University of Stuttgart
Pfaffenwaldring 57
D-70569 Stuttgart
e-mail: ingrid.blaschzyk@mathematik.uni-stuttgart.de
e-mail: ingo.steinwart@mathematik.uni-stuttgart.de

Abstract: In this paper we present a simple partitioning based technique to refine the statistical analysis of classification algorithms. The core of our idea is to divide the input space into two parts such that the first part contains a suitable vicinity around the decision boundary, while the second part is sufficiently far away from the decision boundary. Using a set of margin conditions we are then able to control the classification error on both parts separately. By balancing out these two error terms we obtain a refined error analysis in a final step. We apply this general idea to the histogram rule and show that even for this simple method we obtain, under certain assumptions, better rates than the ones known for support vector machines, for certain plug-in classifiers, and for a recently analysed tree based adaptive-partitioning ansatz.

Contents

1. Introduction

Given a dataset $D := ((x_i, y_i), \dots, (x_n, y_n))$ of observations drawn in an i.i.d. fashion from a probability measure P on $X \times Y$, where $X \subset \mathbb{R}^d$ and $Y := \{-1, 1\}$, the learning goal of binary classification is to find a decision function $f_D \colon X \to \{-1, 1\}$ such that for new data (x, y) we have $f_D(x) = y$ with high probability.

The problem of classification is, apart from regression, one of the most considered problems in learning theory and many classical learning methods have been presented in the literature such as histogram rules, nearest neighbor methods or moving window rules. A general reference for these methods is [4]. Several more recent methods use trees to build a classifier, for example the random forest algorithm, introduced in [3], makes a prediction by a majority vote over a collection of random forest trees. Another example is the tree based adaptive-partitioning algorithm, presented in [2]. Here, a classifier is picked by empirical risk minimization over a nested sequence $(S_m)_{m\geq 1}$ of families of sets which consists of dyadic or decorated trees. An example of a non-tree based algorithm is described in [1]. Here, the final classifier is found by empirical risk minimization over a suitable grid of plug-in rules. Another non-tree based algorithm is, for example, the support vector machine (SVM), which solves a regularized empirical risk minimization problem over a reproducing kernel Hilbert space H. For more details on statistical properties of SVM for classification we refer the reader to [7, Chapter 8].

In this paper we discuss a partitioning based technique to analyse the statistical properties of classification algorithms. In particular we show for the histogram rule that under certain assumptions this technique leads to rates, which are faster than the rates obtained in [2],[3], and [7]. To be more precise, we divide the input space X into two overlapping regions that are adjustable by a parameter r in such a way that one set, which we will denote by A_r, contains points near the decision boundary, whereas the other set B_r contains those that are sufficiently bounded far away from the decision boundary. We examine the excess risks over these two sets separately by using an oracle inequality for empirical risk minimizers on both parts. It turns out that under a suitable assumption, which describes the location of critical noise, we have no approximation error as well as an optimal variance bound on B_r, which in turn leads to an $\mathcal{O}(n^{-1})$ behaviour of the excess risk on B_r. However, this bound still depends on the parameter r, namely it increases for $r \to 0$. In contrast our bound on the risk on A_r decreases for $r \to 0$. By balancing out these two risks with respect to r we obtain a refined bound on X under additional assumptions describing the concentration of mass around the decision boundary.

A more detailed discussion on this technique and the statistical result are presented in Section 3. Moreover a comparison of the resulting learning rates to the known ones for the SVM and the tree based adaptive-partitioning algorithm described in [2] can be find at the end of Section 3. We note that all proofs are deferred to Section 4.

2. General assumptions

To describe our learning goal we consider in the following the classification loss $L := L_{\mathrm{class}}(y,t) : Y \times \mathbb{R} \to [0,\infty)$, defined by $L(y,t) := \mathbf{1}_{(-\infty,0]}(y \cdot \mathrm{sign} t)$ for $y \in Y, t \in \mathbb{R}$, where $\mathbf{1}_{(-\infty,0]}$ denotes the indicator function on $(-\infty, 0]$. We define the risk of a measurable estimator $f : X \to \mathbb{R}$ by

$$\mathcal{R}_{L,P}(f) := \int_{X\times Y} L(y, f(x))\, dP(x,y)$$

and the empirical risk by

$$\mathcal{R}_{L,D}(f) := \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)),$$

where $D := \frac{1}{n}\sum_{i=1}^{n} \delta_{(x_i,y_i)}$ denotes the Dirac measure in (x_i, y_i). The smallest possible risk

$$\mathcal{R}^*_{L,P} := \inf_{f\colon X\to\mathbb{R}} \mathcal{R}_{L,P}(f)$$

is called the Bayes risk, and a measurable function $f^*_{L,P}\colon X \to \mathbb{R}$ so that $\mathcal{R}_{L,P}(f^*_{L,P}) = \mathcal{R}^*_{L,P}$ holds is called Bayes decision function. Recall that the Bayes decision function $f^*_{L,P}$ for the classification loss is given by $\operatorname{sign}(2P(y = 1|x) - 1)$ for $x \in X$, where $P(\,\cdot\,|x)$ is the conditional probability on Y given x. Let us now briefly describe a particular histogram rule. To this end, let $\mathcal{A} = (A_j)_{j\geq 1}$ be a partition of $\mathbb{R}^d$ into cubes of side length $s \in (0, 1]$ and $X := [-1, 1]^d$. For $x \in X$ we denote by $A(x)$ the unique cell of $\mathcal{A}$ with $x \in A(x)$ and call the map $h_{P,s}\colon X \to Y$ defined by

$$h_{P,s}(x) := \begin{cases} -1 & \text{if } f_{P,s}(x) < 0, \\ 1 & \text{if } f_{P,s}(x) \geq 0, \end{cases} \tag{1}$$

where $f_{P,s}(x) := P(A(x) \times \{1\}) - P(A(x) \times \{-1\})$, infinite sample histogram rule. For a dataset D we further write

$$f_{D,s} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{y_i=+1\}}\mathbf{1}_{A(x)}(x_i) - \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{y_i=-1\}}\mathbf{1}_{A(x)}(x_i).$$

Thus, the empirical histogram is defined by $h_{D,s} := \operatorname{sign} f_{D,s}$. We define the set $\mathcal{F}$ by

$$\mathcal{F} := \left\{ \sum_{A_j\cap[-1,1]^d\neq\emptyset} c_j \mathbf{1}_{A_j} : c_j \in \{-1, 1\} \right\}.$$

Then, it is easy to show that the empirical histogram rule $h_{D,s}$ is an empirical risk minimizer over $\mathcal{F}$ for the classification loss, that means

$$\mathcal{R}_{L,D}(h_{D,s}) = \inf_{f\in\mathcal{F}} \mathcal{R}_{L,D}(f).$$

Since we aim in a further step to examine the risk on subsets of X consisting of cells, we have to specify the loss on those subsets. Therefore, we define for an arbitrary index set $J \subset \{1, \ldots, m\}$ the set

$$T_J := \bigcup_{j\in J} A_j \tag{2}$$

and the related loss $L_{T_J} : X \times Y \times \mathbb{R} \to [0, \infty)$ by

$$L_{T_J}(x, y, t) := \mathbf{1}_{\bigcup_{j \in J} A_j}(x) L_{\mathrm{class}}(y, t). \tag{3}$$

Furthermore, we define the risk over T by

$$\mathcal{R}_{L_{T_J}, P}(f) := \int_{X \times Y} L_{T_J}(x, y, f(x))\, dP(x, y).$$

As mentioned in the introduction, we have to make assumptions on P to obtain rates. Therefore we recall some notions from [7, Chapter 8], which describe the behaviour of P in the vicinity of the decision boundary. To this end, let $\eta\colon X \to [0, 1]$, defined by $\eta(x) := P(y = 1|x), x \in X$ be a version of the posterior probability of P, that means that the probability measures $P(\,\cdot\,|x)$ form a regular conditional probability of P. We write

$$\begin{aligned} X_1 &:= \{\, x \in X \colon \eta(x) > 1/2 \,\}, \\ X_{-1} &:= \{\, x \in X \colon \eta(x) < 1/2 \,\}. \end{aligned}$$

Then, the function $\Delta_\eta\colon X \to [0, \infty]$ defined by

$$\Delta_\eta(x) := \begin{cases} d(x, X_1) & \text{if } x \in X_{-1}, \\ d(x, X_{-1}) & \text{if } x \in X_1, \\ 0 & \text{otherwise}, \end{cases} \tag{4}$$

where $d(x, A) := \inf_{x' \in A} d(x, x')$, is called distance to the decision boundary. This helps us to describe the mass of the marginal distribution P_X of P around the decision boundary by the following exponents. We say that P has strong margin exponent (SME) $\alpha \in (0, \infty]$ if there exists a constant $c_{\mathrm{SME}} > 0$ such that

$$P_X(\{\Delta_\eta(x) < t\}) \leq (c_{\mathrm{SME}} t)^\alpha$$

for all $t > 0$. Descriptively, the strong margin exponent α measures the amount of mass close to the decision boundary. Therefore, large values of α are better since they reflect a low concentration of mass in this region, which makes the classification easier. Furthermore, we say that P has margin-noise exponent (MNE) $\beta \in (0, \infty]$ if there exists a constant $c_{\mathrm{MNE}} > 0$ such that

$$\int_{\{\Delta_\eta(x) < t\}} |2\eta(x) - 1|\, dP_X(x) \leq (c_{\mathrm{MNE}} t)^\beta$$

for all $t > 0$. The margin-noise exponent β measures the mass and the noise, that means the amount of points $x \in X$ with $\eta(x) \approx 1/2$, around the decision boundary. That is, we have high margin-noise exponent if we have low mass and/or high noise around the decision boundary. Next, we say that the distance to the decision boundary Δ_η controls the noise from below by the exponent γ if there exist a $\gamma \in [0, \infty)$ and a constant $c_{\mathrm{LC}} > 0$ with

$$\Delta_\eta^\gamma(x) \leq c_{\mathrm{LC}} |2\eta(x) - 1| \tag{5}$$

for P_X-almost all $x \in X$. That means, if $\eta(x)$ is close to $1/2$ for some $x \in X$, this x is close to the decision boundary. For examples of typical values of these exponents and relations between them we refer the reader to [7, Chapter 8].

Finally, in order to describe the region of the decision boundary in a more geometrical way, we say according to [5, 3.2.14(1)] that a general set $T \subset X$ is m-rectifiable for an integer $m > 0$ if there exists a Lipschitzian function mapping some bounded subset of $\mathbb{R}^m$ onto T. Moreover, we denote by $\mathcal{H}^{d-1}$ the $(d-1)$-dimensional Hausdorff measure on $\mathbb{R}^d$.

The following lemma, which is based on [6, Lemma A.10.4], describes the Lebesgue measure of the decision boundary in terms of the Hausdorff measure. Its result will be necessary for the analysis of the main theorem in Section 3.

Lemma 2.1. *Let $X := [-1,1]^d$ and P be a probability measure on $X \times \{-1,1\}$ with fixed version $\eta\colon X \to [0,1]$ of its posterior probability. Moreover let λ^d be the d-dimensional Lebesgue measure and $\mathcal{H}^{d-1}$ be the $(d-1)$-dimensional Hausdorff measure on $\mathbb{R}^d$. Furthermore, let X_0 equal the relative boundary of X_1 in X, that means $X_0 = \delta_X X_1$, with $\mathcal{H}^{d-1}(X_0) > 0$ and let X_0 be $(d-1)$-rectifiable. Then, there exists a $\delta^* > 0$ such that for all $\delta \in (0, \delta^*]$ we have*

$$\lambda^d(\{\, x \in X \mid \Delta(x) \le \delta \,\}) \le 4\mathcal{H}^{d-1}(\{\, x \in X \mid \eta(x) = 1/2 \,\}) \cdot \delta.$$

3. Oracle inequality and learning rates

Our goal is to find an upper bound for the excess risk $\mathcal{R}_{L,P}(h_{D,s}) - \mathcal{R}^*_{L,P}$. The idea is to split X into two overlapping sets and to find a bound on the risks over these sets by using information on P. To this end, we denote the set of indices of cubes that intersect X by

$$J := \{\, j \ge 1 \mid A_j \cap [-1,1]^d \neq \emptyset \,\}.$$

Next, we split this set into cubes that lie near the decision boundary and into cubes that are bounded away from the decision boundary. To be more precisely, we define, for $r > 0$ and a version η for which the assumptions at the end of Section 2 hold, the set of indices of cubes near the decision boundary by

$$J_A^r := \{\, j \in J \mid \forall\, x \in A_j : \Delta_\eta(x) \le 3r \,\}$$

and the set of indices of cubes that are sufficiently bounded away by

$$J_B^r := \{\, j \in J \mid \forall\, x \in A_j : \Delta_\eta(x) \ge r \,\}.$$

Moreover, we write

$$A_r := \bigcup_{j \in J_A^r} A_j, \tag{6}$$

$$B_r := \bigcup_{j \in J_B^r} A_j. \tag{7}$$

As the following lemma shows, we need to define requirements on the side length of the cells to ensure that $X \subset A_r \cup B_r$. Besides that, it shows that we are able to assign all $x \in A_j$, where $j \in J_B^r$, either to the class X_{-1} or to X_1.

Lemma 3.1. *Let $\mathcal{A} = (A_j)_{j\geq 1}$ be a partition of $\mathbb{R}^d$ into cubes of side length $s \in (0,1]$ and let $X := [-1,1]^d$. For $r \geq s/2$ define the sets A_r and B_r by (6) and (7). Then,*

i) we have $X \subset A_r \cup B_r$,
ii) we have either $A_j \cap X_1 = \emptyset$ or $A_j \cap X_{-1} = \emptyset$ for $j \in J_B^r$.

Since the excess risk is non-negative, we obtain under the assumption of Lemma 3.1(i) that

$$\begin{aligned} &\mathcal{R}_{L,P}(h_{D,s}) - \mathcal{R}^*_{L,P} \\ &\leq \left(\mathcal{R}_{L_{A_r},P}(h_{D,s}) - \mathcal{R}^*_{L_{A_r},P}\right) + \left(\mathcal{R}_{L_{B_r},P}(h_{D,s}) - \mathcal{R}^*_{L_{B_r},P}\right). \end{aligned} \tag{8}$$

That means, we can bound the excess risk $\mathcal{R}_{L,P}(h_{D,s}) - \mathcal{R}^*_{L,P}$ if we find bounds on the excess risks over the sets A_r and B_r. For that purpose, we use an oracle inequality for empirical risk minimizer separately on both error terms, see [7, Theorem 7.2]. This is possible, since the following lemma shows that, considering the loss L_{T_J} for any set T_J constructed as in (2), the empirical histogram rule $h_{D,s}$ is still an empirical risk minimizer over $\mathcal{F}$.

Lemma 3.2. *Consider for an arbitrary index set $J \subset \{1,\dots,m\}$ the set $T_J := \bigcup_{j\in J} A_j$ and the related loss $L_{T_J} : X \times Y \times \mathbb{R} \to [0,\infty)$ defined in (3). Then, the empirical histogram rule $h_{D,s}$ is an empirical risk minimizer over $\mathcal{F}$ for the loss L_{T_J}, that means*

$$\mathcal{R}_{L_{T_J},D}(h_{D,s}) = \inf_{f\in\mathcal{F}} \mathcal{R}_{L_{T_J},D}(f).$$

Before we state our oracle inequality, we discuss in a more detailed way the improvement that we gained by our separation technique described above. First, we make no approximation error on the set B_r, which consists of cells that are sufficiently bounded away from the decision boundary. This follows from the circumstance that $h_{D,s}$ learns correctly on those cells and follows even intuitively inasmuch as the noise concentration is rather low in this region. We refer the reader to Part 1 of the proof of Lemma 3.4 for details. Second, the main refinement arises from the fact that we achieve, under the condition that the decision boundary controls the noise from above, in the variance bound on B_r, a bound of the form

$$\mathbb{E}_P(L \circ f - L \circ f^*_{L,P})^2 \leq V \cdot \mathbb{E}_P(L \circ f - L \circ f^*_{L,P})^\theta$$

where $V > 0$, the best possible exponent, $\theta = 1$. This bound plays an important part in the analysis of the risk terms, since we have small variance if the right-hand side of the latter bound is small, as the next lemma shows.

Lemma 3.3. *Let $X := [-1,1]^d$ and P be a probability measure on $X \times \{-1,1\}$ with fixed version $\eta\colon X \to [0,1]$ of its posterior probability. Assume that the associated distance to the decision boundary Δ_η controls the noise from below by the exponent $\gamma \in [0,\infty)$ and consider for some fixed $r > 0$ the set B_r, defined in (7). Furthermore let $L := L_{class}$ be the classification loss and let $f^*_{L,P}$ be a fixed Bayes decision function. Then, for all measurable $f\colon X \to \{-1,1\}$ we have*

$$\mathbb{E}_P(L_{B_r} \circ f - L_{B_r} \circ f^*_{L,P})^2 \leq \frac{c_{LC}}{r^\gamma} \mathbb{E}_P(L_{B_r} \circ f - L_{B_r} \circ f^*_{L,P}).$$

We remark that the right-hand side of the variance bound on B_r depends on the separation parameter r. This dependence is also reflected in the risk term on B_r. In particular, we show in the proof of our main theorem by applying [7, Theorem 7.2] on the risk term on the set B_r that the improvements mentioned above lead to

$$\mathcal{R}_{L_{B_r},P}(h_{D,s}) - \mathcal{R}^*_{L_{B_r},P} \leq \frac{32 c_1 (8^{d+1} s^{-d} + \tau)}{r^\gamma n}$$

with probability $P^n \geq 1 - e^{-\tau}$, where $\tau \geq 1$ and c_1 is a positive constant. Whereas this error term increases for $r \to 0$, the error term on the set A_r behaves exactly the opposite way, that is, it decreases for $r \to 0$. In fact, bounding the risk on A_r requires additional knowledge of the behaviour of P in the vicinity of the decision boundary. By applying [7, Theorem 7.2] on the risk on the set A_r we show under the assumption that P has strong margin exponent α and margin-noise exponent β that

$$\mathcal{R}_{L_A,P}(h_D) - \mathcal{R}^*_{L_A,P} \leq 6 c_{\mathrm{MNE}} s^\beta + 4 \left(\frac{8V(c_5 r s^{-d} + \tau)}{n} \right)^{\frac{\alpha+\gamma}{\alpha+2\gamma}}$$

holds with probability $P^n \geq 1 - e^{-\tau}$. Here, c_5 is a positive constant, $\tau \geq 1$ and V is the prefactor of the variance bound on A_r, shown in Part 2 of the proof of Lemma 3.4. If we balance the obtained risk terms over A_r and B_r with respect to r, we obtain the oracle inequality presented in our main theorem. For this purpose, we define the positive constant

$$\tilde{c}_{\alpha,\gamma,d} := \left(\frac{16\gamma(\alpha+2\gamma) \cdot 8^{d+1} \max\{c_{\mathrm{LC}}, 2^\gamma\} \cdot (\alpha+\gamma)^{-1}}{\hat{c}^{\frac{\alpha+\gamma}{\alpha+2\gamma}}} \right)^{\frac{\alpha+\gamma}{\alpha+\gamma+\gamma(\alpha+2\gamma)}},$$

which depends on α, γ and d and where $\hat{c} := 24 \max\{12\mathcal{H}^{d-1}(\{\eta = 1/2\}), 1\} \cdot \max\left\{1, \frac{\alpha+\gamma}{\gamma} c_{\mathrm{SME}}^{\frac{\alpha\gamma}{\alpha+\gamma}} \left(\frac{\gamma c_{\mathrm{LC}}}{\alpha}\right)^{\frac{\alpha}{\alpha+\gamma}}\right\}$.

Theorem 3.4. *Let $\mathcal{A} = (A_j)_{j\geq 1}$ be a partition of $\mathbb{R}^d$ into cubes of side length $s \in (0,1]$. Let $X := [-1,1]^d$ and P be a probability measure on $X \times \{-1,1\}$ with fixed version $\eta\colon X \to [0,1]$ of its posterior probability. Assume that the associated distance to the decision boundary Δ_η controls the noise from below by the exponent $\gamma \in [0,\infty)$ and assume as well that P has MNE $\beta \in [0,\infty)$ and*

SME $\alpha \in (0,\infty]$. Furthermore, let X_0 equal the relative boundary of X_1 in X, that means $X_0 = \delta_X X_1$, with $\mathcal{H}^{d-1}(X_0) > 0$ and let X_0 be $(d-1)$-rectifiable. Let L be the classification loss and let for fixed $n \geq 1$ and $\tau \geq 1$ the bounds

$$s \leq \tilde{c}_{\alpha,\gamma,d}^{\frac{(1+\gamma)(\alpha+\gamma)+\gamma^2}{(1+\gamma)(\alpha+\gamma)+\gamma^2+d\gamma}} \left(\frac{\tau}{n}\right)^{\frac{\gamma}{(1+\gamma)(\alpha+\gamma)+\gamma^2+d\gamma}}, \tag{9}$$

and

$$s^d n \geq \tau \left(\frac{\tilde{c}_{\alpha,\gamma,d}}{\min\{\frac{\delta^*}{3},1\}}\right)^{\frac{(1+\gamma)(\alpha+\gamma)+\gamma^2}{\gamma}} \tag{10}$$

be satisfied, where the constant $\tilde{c}_{\alpha,\gamma,d} > 0$ depends on α, γ, d and the constant $\delta^* > 0$ is the one of Lemma 2.1. Then, there exists a constant $c_{\alpha,\gamma,d} > 0$ such that

$$\mathcal{R}_{L,P}(h_{D,s}) - \mathcal{R}^*_{L,P} \leq 6\left(c_{MNE}s\right)^{\beta} + c_{\alpha,\gamma,d}\left(\frac{\tau}{s^d n}\right)^{\frac{(1+\gamma)(\alpha+\gamma)}{(1+\gamma)(\alpha+\gamma)+\gamma^2}} \tag{11}$$

holds with probability $P^n \geq 1-2e^{-\tau}$, where the constant $c_{\alpha,\gamma,d}$ only depends on α, γ and d.

The proof shows that the constants $c_{\alpha,\gamma,d}$ is given by

$$c_{\alpha,\gamma,d} := 128 \cdot 8^{d+1} \max\{c_{\mathrm{LC}}, 2^{\gamma}\} \cdot \max\left\{\frac{\gamma(\alpha+2\gamma)}{\alpha+\gamma}, 1\right\} \cdot \tilde{c}_{\alpha,\gamma,d}^{-\gamma}. \tag{12}$$

Note that the assumptions (9) and (10) on the side length s of the cubes are natural assumptions, since s has to be small enough given a specific number of observations, but yet should not shrink too fast for grown observations. By choosing an appropriate sequence of s_n in dependence of our data length n and setting a constraint on the margin-noise exponent β we state learning rates in the next theorem. Prior to that, we define the positive constant

$$\tilde{c}_{\alpha,\beta,\gamma,\tau,d} := \left(\frac{d(1+\gamma)(\alpha+\gamma)\cdot c_{\alpha,\gamma,d}\cdot \tau^{\frac{(1+\gamma)(\alpha+\gamma)}{(1+\gamma)(\alpha+\gamma)+\gamma^2}}}{6\beta c_{\mathrm{MNE}}^{\beta}((1+\gamma)(\alpha+\gamma)+\gamma^2)}\right)^{\frac{(1+\gamma)(\alpha+\gamma)+\gamma^2}{\beta((1+\gamma)(\alpha+\gamma)+\gamma^2)+d(1+\gamma)(\alpha+\gamma)}}$$

that depends on $\alpha, \beta, \gamma, \tau$ and d and where $c_{\alpha,\gamma,d}$ is the constant from (12).

Theorem 3.5. *Assume that X and P satisfy the assumptions of Theorem 3.4 for $\beta \leq \gamma^{-1}\kappa$, where $\kappa := (1+\gamma)(\alpha+\gamma)$. In addition assume that the side length s_n in Theorem 3.4 is given by*

$$s_n = \tilde{c}_{\alpha,\beta,\gamma,\tau,d}\, n^{-\frac{\kappa}{\beta(\kappa+\gamma^2)+d\kappa}}.$$

Then, there exists a constant $c_{\alpha,\beta,\gamma,\tau,d} > 0$ such that for all $n \geq n_0$

$$\mathcal{R}_{L,P}(h_{D,s_n}) - \mathcal{R}^*_{L,P} \leq c_{\alpha,\beta,\gamma,\tau,d}\, n^{-\frac{\beta\kappa}{\beta(\kappa+\gamma^2)+d\kappa}}$$

holds with probability $P^n \geq 1 - 2e^{-\tau}$, where n_0 and the constant $c_{\alpha,\beta,\gamma,\tau,d}$ only depend on $\tau, \alpha, \beta, \gamma$ and d.

The proof shows that the constant $c_{\alpha,\beta,\gamma,\tau,d}$ is given by

$$c_{\alpha,\beta,\gamma,\tau,d} := 2\max\left\{\frac{d\cdot\kappa}{\beta(\kappa+\gamma^2)}, 1\right\} c_{\alpha,\gamma,\delta}\cdot\tau^{\frac{\kappa}{\kappa+\gamma^2}}\cdot\tilde{c}_{\alpha,\beta,\gamma,\tau,d}^{-\frac{d\kappa}{\kappa+\gamma^2}}.$$

To obtain the rates we have to know the P describing parameters. However, it is possible to yield the rates in Theorem 3.5 by a training validation ansatz, that is by splitting the dataset into two parts and considering a suitable set of candidates s_n. In order to compare our rate obtained in Theorem 3.5, we now consider, besides our geometric assumption on X, namely

(i) X_0 is $(d-1)$-rectifiable with $\mathcal{H}^{d-1}(X_0) > 0$ and X_0 equals the relative boundary of X_1 in X,

the following two assumptions on P:

(ii) P has SME $\alpha \in (0,\infty]$,
(iii) there exists a $\gamma \in [0,\infty)$ and constants $c_1, c_2, c_{\mathrm{UC}} > 0$ such that
   a) $c_1|2\eta(x)-1| \geq c_{\mathrm{LC}}\Delta_\eta^\gamma(x)$,
   b) $c_2|2\eta(x)-1| \leq c_{\mathrm{UC}}\Delta_\eta^\gamma(x)$.

Here, assumption $(iii)_{a)}$ coincides up to the constant c_1 with the definition in (5). Furthermore, assumption $(iii)_{b)}$ indicates that we have an upper control by Δ_η on the noise, which is a kind of inverse to $(iii)_{a)}$. Then, [7, Lemma 8.17] shows under the assumptions (ii) and (iii) that P has MNE $\beta = \alpha+\gamma$. Hence, we find in Theorem 3.5 with $\kappa := (1+\gamma)(\alpha+\gamma)$ and a suitable cell-width that h_{D,s_n} learns with a rate with exponent

$$\begin{aligned}\frac{\beta(1+\gamma)(\alpha+\gamma)}{\beta[(1+\gamma)(\alpha+\gamma)+\gamma^2]+d(1+\gamma)(\alpha+\gamma)} &= \frac{(\alpha+\gamma)(1+\gamma)(\alpha+\gamma)}{(\alpha+\gamma)[(1+\gamma)(\alpha+\gamma)+\gamma^2]+d(1+\gamma)(\alpha+\gamma)}\\ &= \frac{(1+\gamma)(\alpha+\gamma)}{(1+\gamma)(\alpha+\gamma)+\gamma^2+d(1+\gamma)}.\end{aligned}$$

A simple transformation shows that this exponent equals

$$\frac{(1+\gamma)(\alpha+\gamma)}{(1+\gamma)(\alpha+\gamma)+\gamma^2+d(1+\gamma)} = \frac{\alpha+\gamma}{\alpha+\gamma+\frac{\gamma^2}{1+\gamma}+d} = \frac{\alpha+\gamma}{\alpha+2\gamma+\frac{\gamma^2}{1+\gamma}+d-\gamma} = \frac{\alpha+\gamma}{\alpha+2\gamma+d-\frac{\gamma}{1+\gamma}}. \tag{13}$$

First, we compare the rate with exponent (13) with the rate achieved by support vector machines (SVM) for the hinge loss by considering the assumptions $(i), (ii)$ and (iii). For this purpose, [7, Chapter 8.3(8.18)] shows that the best possible rate for the SVM is obtained by

$$n^{-\frac{\alpha+\gamma}{\alpha+2\gamma+d}+\rho},$$

where $\rho > 0$ is an arbitrary small number. Hence, our rate in (13) is better by $-\frac{\gamma}{1+\gamma}$ in the denominator. For the typical value of $\gamma = 1$, indicating a moderate control of noise by the decision boundary, our rate is better by $-1/2$ in the denominator. Finally, we remark that for both results less assumptions are sufficient.

Second, we compare our rate with the ones for certain plug-in classifiers, see [1], and with the rates of a classification scheme, described in [2]. In both cases, the authors use the so called margin assumption, which is comparable to the Definition of the noise exponent in [7]. Hence, we find under assumptions (ii) and (iii) with [7, Exercise 8.5] that the margin condition is fulfilled for α/γ. In addition to (i) and (iii) we impose the following two conditions:

(iv) η is Hoelder-continuous with exponent γ,
(v) P_X is the Lebesgue measure.

Under condition (i) and (v) we find with Lemma 2.1 that assumption (ii) is fulfilled for $\alpha = 1$. Furthermore, we find under condition (iv) with Lemma A.1 that assumption (iii) is fulfilled with exponent γ. Hence, the conditions (i) and $(iii) - (v)$ yield in (13) a rate with exponent

$$\frac{1+\gamma}{1+2\gamma+d-\frac{\gamma}{1+\gamma}}.$$

Furthermore, [1, Theorem 4.3] shows that certain plug-in classifier yield under the same conditions (i) and $(iii) - (v)$ the rate

$$n^{-\frac{1+\gamma}{1+2\gamma+d}} \tag{14}$$

and we find that our rate is better by $-\frac{\gamma}{1+\gamma}$ in the denominator. Under the same assumptions, [2, Corollary 5.2(ii)] shows that the described classification scheme obtains the rate

$$\left(\frac{(\log n)^{\frac{1}{2+d}}}{n}\right)^{-\frac{1+\gamma}{2\gamma+d}}.$$

Hence, our rate is worse by $\frac{1}{1+\gamma}$. However, the results from [2] are also comparable under another set of assumptions. Indeed, if we assume that the conditions $(i) - (iv)$ hold, then, our rate given in (13) holds and [2, Corollary 5.2(i)] shows that the described classification scheme yields the rate

$$\left(\frac{\log n}{n}\right)^{-\frac{\alpha+\gamma}{\alpha+2\gamma+d}}$$

and our rate is again better by $-\frac{\gamma}{1+\gamma}$ in the denominator. Note that for $\alpha = 1$ this rate equals (14) up to the logarithm. Finally, we remark that for our results as well as for the results from [1] and [2] less assumptions are sufficient and in the comparisons above we tried to formulate reasonable sets of common assumptions.

4. Proofs

Proof of Lemma 2.1: For a set $T \subset X$ and $\delta > 0$ we define as in [6] the sets

$$T^{+\delta} := \{\, x \in X \,|\, d(x,T) \leq \delta \,\},$$
$$T^{-\delta} := X \setminus (X \setminus T)^{+\delta}.$$

Since $X_1 := \{\, x \in X \,|\, \eta(x) \leq 1/2 \,\}$ is bounded and measurable, we find with [6, Lemma A.10.3] and the proof of [6, Lemma+A.10.4(ii)] that there exists a $\delta^* > 0$, such that for all $\delta \in (0, \delta^*]$ we have

$$\lambda^d(X_1^{+\delta} \setminus X_1^{-\delta}) \leq 4\mathcal{H}^{d-1}(\partial X_1) \cdot \delta = 4\mathcal{H}^{d-1}(\{\, x \in X \,|\, \eta(x) = 1/2 \,\}) \cdot \delta. \quad (15)$$

Next, we show that

$$\{\, x \in X \,|\, \Delta(x) \leq \delta \,\} \subset X_1^{+\delta} \setminus X_1^{-\delta} \cup X_0. \quad (16)$$

For this purpose, we remark that according to (4) we have

$$\begin{aligned}&\{\, x \in X \,|\, \Delta(x) \leq \delta \,\}\\ &= \{\, x \in X_1 \,|\, d(x, X_{-1}) \leq \delta \,\} \cup \{\, x \in X_{-1} \,|\, d(x, X_1) \leq \delta \,\} \cup X_0.\end{aligned}$$

Let us first show that $\{\, x \in X_1 \,|\, d(x, X_{-1}) \leq \delta \,\} \subset X_1^{+\delta} \setminus X_1^{-\delta}$. To this end, consider an $x \in X_1$ with $d(x, X_{-1}) \leq \delta$, where we check at once that $x \in X_1^{+\delta}$. Now, assume that $x \in X_1^{-\delta} = X \setminus (X \setminus X_1)^{+\delta}$. Then, we find that $x \notin (X \setminus X_1)^{+\delta}$ such that $d(x, X \setminus X_1) = d(x, X_{-1} \cup X_0) > \delta$. Hence, $x \notin X_1^{-\delta}$. Next, let us show that $\{\, x \in X_{-1} \,|\, d(x, X_1) \leq \delta \,\} \subset X_1^{+\delta} \setminus X_1^{-\delta}$. To this end, consider an $x \in X_{-1}$ with $d(x, X_1) \leq \delta$. Then, it is clear that $x \in X_1^{+\delta}$ by definition of $X_1^{+\delta}$. Furthermore, $x \notin X_1^{-\delta}$ since $X_1^{-\delta} = X \setminus (X_{-1})^{+\delta} \subset X_1$. Having showed (16), we find together with the fact that $\lambda^d(X_0) = 0$ since $X_0 = \partial X_1$ is $(d-1)$-rectifiable that

$$\lambda^d(\{\, x \in X \,|\, \Delta(x) \leq \delta \,\}) \leq \lambda^d(X_1^{+\delta} \setminus X_1^{-\delta}).$$

Finally, with (15) we obtain that

$$\lambda^d(\{\, x \in X \,|\, \Delta(x) \leq \delta \,\}) \leq \lambda^d(X_1^{+\delta} \setminus X_1^{-\delta}) \leq 4\mathcal{H}^{d-1}(\{\, x \in X \,|\, \eta(x) = 1/2 \,\}) \cdot \delta$$

for all $\delta \in (0, \delta^*]$. □

Proof of Lemma 3.1:

i) We define the set of indices

$$J_C^r := \{\, j \in J \,|\, \exists\, x \in A_j : \Delta_\eta(x) < r \,\}$$

and define the set

$$C_r := \bigcup_{j \in J_C^r} A_j.$$

Since $X \subset B_r \cup C_r$, it suffices to show that $C_r \subset A_r$. To show the latter we fix an $x \in C_r$. If $x \in X_0$ we immediately have $\Delta_\eta(x) = 0 \leq 3r$, hence we assume w.l.o.g. that $x \in X_1$. Then, there exists a $j \in J_C^r$ such that $x \in A_j$. Furthermore, there exists an $x^* \in A_j$ with $\Delta_\eta(x^*) < r$ and we find

$$\begin{aligned}\Delta_\eta(x) &= \inf_{x' \in X_{-1}} \|x - x'\|_\infty \\ &\leq \inf_{x' \in X_{-1}} (\|x - x^*\|_\infty + \|x^* - x'\|_\infty) \\ &\leq s + \Delta_\eta(x^*) \\ &< s + r,\end{aligned}$$

where $\|\cdot\|_\infty$ is the supremum norm in $\mathbb{R}^d$. Since $s \leq 2r$, it follows that $\Delta_\eta(x) \leq 3r$ and therefore $x \in A_r$.

ii) We assume for A_j with $j \in J_B^r$ that we have an $x_1 \in A_j \cap X_1 \neq \emptyset$ and an $x_{-1} \in A_j \cap X_{-1} \neq \emptyset$. Then, the connecting line $\overline{x_{-1}x_1}$ from x_{-1} to x_1 is contained in A_j since A_j is convex and we have $\|x_{-1} - x_1\|_\infty \leq s$. Moreover, since $\Delta_\eta(x) \geq r$ for all $x \in B_r$ we have $X_0 \cap B_r = \emptyset$. Next, pick an $m > 1$ such that

$$t_0 = 0, \qquad t_m = 1, \qquad t_i = \frac{i}{m}$$

and

$$x_i := t_i x_{-1} + (1 - t_i)x_1$$

for $i = 0, \ldots, m$. Clearly, $x_i \in \overline{x_{-1}x_1}$ and $x_i \in X_{-1} \cup X_1$. Since $x_0 \in X_1$ and $x_m \in X_{-1}$, there exists an i with $x_i \in X_1$ and $x_{i+1} \in X_{-1}$ and we find that

$$\|x_i - x_{i+1}\|_\infty \geq \Delta_\eta(x_i) \geq r.$$

On the other hand,

$$\|x_i - x_{i+1}\|_\infty = \frac{1}{m}\|x_{-1} - x_1\|_\infty \leq \frac{s}{m} \leq \frac{2r}{m}$$

such that $r \leq \frac{2r}{m}$, which is not true for $m \geq 3$. Hence, we can not have an $x_1 \in A_j \cap X_1 \neq \emptyset$ and an $x_{-1} \in A_j \cap X_{-1} \neq \emptyset$ for $j \in J_B^r$.

$\square$

Proof of Lemma 3.2: For $f \in \mathcal{F}$ we have

$$\begin{aligned}&\mathcal{R}_{L_{T_J},D}(f) \\ &= \int_{X \times Y} L_{T_J}(x, y, f(x))\, dD(x, y) \\ &= \sum_{j \in J} \int_{A_j \times Y} L_{\mathrm{class}}(y, f(x))\, dD(x, y).\end{aligned}$$

Next, we take a closer look at the risk on a single cell A_j for a $j \in J$. That is,

$$\int_{A_j \times Y} L_{\text{class}}(y, f(x)) \, dD(x,y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{A_j}(x_i) \mathbf{1}_{y_i \neq c_i}.$$

The risk on a cell is the smaller the less often we have $y_i \neq c_i$ such that the best classifier on a cell is the one which decides by majority. This is true for the histogram rule by definition. Since the risk is zero on A_j with $j \notin J$, the histogram rule minimizes the risk with respect to L_{T_J}. □

Proof of Lemma 3.3: We define $h_f := L_{B_r} \circ f - L_{B_r} \circ f^*_{L,P}$ for a measurable $f \colon X \to \{-1, 1\}$. Since $(L_{B_r} \circ f - L_{B_r} \circ f^*_{L,P})^2 = \mathbf{1}_{B_r} \frac{|f - f^*_{L,P}|}{2}$ we obtain

$$\begin{aligned}
&\mathbb{E}_P(h_f - \mathbb{E}_P h_f)^2 \\
&\leq \mathbb{E}_P(h_f)^2 \\
&= \mathbb{E}_P(L_{B_r} \circ f - L_{B_r} \circ f^*_{L,P})^2 \\
&= \frac{1}{2} \mathbb{E}_P \mathbf{1}_{B_r} |f - f^*_{L,P}|.
\end{aligned}$$

For $x \in B_r$ we have $\Delta_\eta(x) \geq r$ and thus we find with our lower-control assumption that

$$r^\gamma \leq \Delta_\eta^\gamma(x) \leq c_{LC} |2\eta(x) - 1|$$

and therefore

$$1 \leq c_{LC} r^{-\gamma} |2\eta(x) - 1|.$$

By using $\mathbf{1}_{B_r} \frac{|f - f^*|}{2} = \mathbf{1}_{(X_{-1} \triangle \{f<0\}) \cap B_r}$, where $\triangle$ denotes the symmetric difference defined by $C \triangle D := (C \setminus D) \cup (D \setminus C)$ for sets $C, D \subset X$ and by using Lemma A.1 we obtain for the variance bound

$$\begin{aligned}
\mathbb{E}_P(h_f - \mathbb{E}_P h_f)^2 &\leq \frac{1}{2} \int \mathbf{1}_{B_r}(x) |f(x) - f^*_{L,P}(x)| \, dP_X(x) \\
&\leq \frac{c_{LC}}{2 r^\gamma} \int \mathbf{1}_{B_r}(x) |2\eta(x) - 1| |f(x) - f^*_{L,P}(x)| \, dP_X(x) \\
&= \frac{c_{LC}}{r^\gamma} \int_{(X_{-1} \triangle \{f<0\}) \cap B_r} |2\eta(x) - 1| \, dP_X(x) \\
&= \frac{c_{LC}}{r^\gamma} (\mathcal{R}_{L_{B_r},P}(f) - \mathcal{R}^*_{L_{B_r},P}) \\
&= \frac{c_{LC}}{r^\gamma} \mathbb{E}_P h_f.
\end{aligned}$$

□

Proof of Theorem 3.4: We define the set of cubes A_r and B_r as in (6), (7) for the choice of

$$r := \tilde{c}_{\alpha,\gamma,d} \left(\frac{\tau}{s^d n} \right)^{\frac{1-\theta}{1+\gamma(2-\theta)}}, \tag{17}$$

where

$$\theta := \frac{\alpha}{\alpha+\gamma}. \tag{18}$$

To estimate the excess risk $\mathcal{R}_{L,P}(h_{D,s}) - \mathcal{R}^*_{L,P}$, we split the risk as in (8) by

$$\begin{aligned}
&\mathcal{R}_{L,P}(h_{D,s}) - \mathcal{R}^*_{L,P} \\
&\leq \left(\mathcal{R}_{L_{A_r},P}(h_{D,s}) - \mathcal{R}^*_{L_{A_r},P}\right) + \left(\mathcal{R}_{L_{B_r},P}(h_{D,s}) - \mathcal{R}^*_{L_{B_r},P}\right).
\end{aligned}$$

This separation is valid by Lemma 3.1(i), since $s \leq r$. To see that, we remark that

$$\begin{aligned}
s \leq \tilde{c}_{\alpha,\gamma,d}\left(\frac{\tau}{s^d n}\right)^{\frac{1-\theta}{1+\gamma(2-\theta)}} &\iff s^{\frac{1+\gamma(2-\theta)+d(1-\theta)}{1+\gamma(2-\theta)}} \leq \tilde{c}_{\alpha,\gamma,d}\left(\frac{\tau}{n}\right)^{\frac{1-\theta}{1+\gamma(2-\theta)}} \\
&\iff s \leq \left(\tilde{c}_{\alpha,\gamma,d}\left(\frac{\tau}{n}\right)^{\frac{1-\theta}{1+\gamma(2-\theta)}}\right)^{\frac{1+\gamma(2-\theta)}{1+\gamma(2-\theta)+d(1-\theta)}}
\end{aligned}$$

and conclude by replacing θ by (18) that $s \leq r$ holds if

$$s \leq \tilde{c}_{\alpha,\gamma,d}^{\frac{(1+\gamma)(\alpha+\gamma)+\gamma^2}{(1+\gamma)(\alpha+\gamma)+\gamma^2+d\gamma}}\left(\frac{\tau}{n}\right)^{\frac{\gamma}{(1+\gamma)(\alpha+\gamma)+\gamma^2+d\gamma}},$$

which equals (9). The rest of the proof is structured in three parts, where we establish error bounds on A_r and B_r in the first two parts and combine the results obtained in the third and last part of the proof. In the following we write $A := A_r$ and $B := B_r$ and keep in mind, that these sets depend on a parameter r. Furthermore we write $h_D := h_{D,s}$.

Part 1: In the first part we establish an oracle inequality for $\mathcal{R}_{L_B,P}(h_{D,s}) - \mathcal{R}^*_{L_B,P}$. Therefore we define $h_f^B := L_B \circ f - L_B \circ f^*_{L_B,P}$ and find that

$$\|h_f^B\|_\infty = \|L_B \circ f - L_B \circ f^*_{L_B,P}\|_\infty \leq 1$$

for all $f \in \mathcal{F}$. Furthermore with Lemma 3.3 we obtain

$$\mathbb{E}_P(h_f^B)^2 \leq \frac{c_{\mathrm{LC}}}{r^\gamma}\mathbb{E}_P h_f^B \leq \frac{c_1}{r^\gamma}\mathbb{E}_P h_f^B, \tag{19}$$

where $c_1 := \max\{c_{\mathrm{LC}}, 2^\gamma\}$. We observe that $r^\gamma \leq c_1$, since with assumption (10), where we rewrite the exponent by $\frac{(1+\gamma)(\alpha+\gamma)+\gamma^2}{\gamma} = \frac{1-\theta}{1+\gamma(2-\theta)}$, we find

$$\begin{aligned}
r &= \tilde{c}_{\alpha,\gamma,d}\left(\frac{\tau}{s^d n}\right)^{\frac{1-\theta}{1+\gamma(2-\theta)}} \\
&\leq \tilde{c}_{\alpha,\gamma,d}\left(\left(\frac{\min\{\frac{\delta^*}{3},1\}}{\tilde{c}_{\alpha,\gamma,d}}\right)^{\frac{1+\gamma(2-\theta)}{1-\theta}}\right)^{\frac{1-\theta}{1+\gamma(2-\theta)}} \\
&= \min\left\{\frac{\delta^*}{3}, 1\right\} \\
&\leq 1
\end{aligned}$$

and therefore $r^\gamma \leq 2^\gamma \leq c_1$. As we conclude from Lemma 3.2 that h_D is an empirical risk minimizer over $\mathcal{F}$ for the loss L_B, we are able to use [7, Theorem 7.2], an improved oracle inequality for ERM. We obtain for all fixed $\tau \geq 1$ and $n \geq 1$ that

$$\mathcal{R}_{L_B,P}(h_D) - \mathcal{R}^*_{L_B,P} < 6(\mathcal{R}^*_{L_B,P,\mathcal{F}} - \mathcal{R}^*_{L_B,P}) + \frac{32c_1(\log(|\mathcal{F}|+1)+\tau)}{r^\gamma n}$$

holds with probability $P^n \geq 1 - e^{-\tau}$, where $\mathcal{R}^*_{L_B,P,\mathcal{F}} := \inf_{f\in\mathcal{F}} \mathcal{R}_{L_B,P}(f)$. Next, we refine the right-hand side of this oracle inequality. Obviously we have $|\mathcal{F}| \leq 2^{|J|}$. We bound the the cardinality $|J|$ by using a volume comparison argument. To this end, we define the set $\tilde{J} := \{\, j \geq 1 \,|\, A_j \cap 2\,[-1,1]^d \neq \emptyset \,\}$ and observe that $\bigcup_{j\in J} A_j \subset \bigcup_{j\in\tilde{J}} A_j \subset 2B_{\ell^d_\infty}$. Then,

$$|J|s^d = \lambda^d\left(\bigcup_{j\in J} A_j\right) \leq \lambda^d\left(\bigcup_{j\in \tilde{J}} A_j\right) \leq \lambda^d\left(4B_{\ell^d_\infty}\right) = 8^d,$$

such that we deduce with $|J| \leq 8^d s^{-d}$ that

$$\begin{aligned}
\log(|\mathcal{F}|+1) &\leq \log(2^{8^d s^{-d}} + 1) \\
&\leq \log(2 \cdot 2^{8^d s^{-d}}) \\
&= \log(2^{8^d s^{-d}+1}) \\
&= (8^d s^{-d} + 1)\log(2) \\
&\leq 8^d s^{-d} + 1 \\
&\leq 8^{d+1} s^{-d}.
\end{aligned}$$

Thus,

$$\mathcal{R}_{L_B,P}(h_D) - \mathcal{R}^*_{L_B,P} < 6(\mathcal{R}^*_{L_B,P,\mathcal{F}} - \mathcal{R}^*_{L_B,P}) + \frac{32c_1(8^{d+1}s^{-d}+\tau)}{r^\gamma n} \tag{20}$$

holds with probability $P^n \geq 1 - e^{-\tau}$.

Finally, we have to bound the *approximation error* $\mathcal{R}^*_{L_B,P,\mathcal{F}} - \mathcal{R}^*_{L_B,P} = \inf_{f\in\mathcal{F}} \mathcal{R}_{L_B,P}(f) - \mathcal{R}^*_{L_B,P}$. We find with $h_{P,s} \in \mathcal{F}$ and Lemma A.1 that

$$\begin{aligned}
\mathcal{R}^*_{L_B,P,\mathcal{F}} - \mathcal{R}^*_{L_B,P} &\leq \mathcal{R}_{L_B,P}(h_{P,s}) - \mathcal{R}^*_{L_B,P} \\
&= \int_{(X_1 \triangle \{h_{P,s}\geq 0\})\cap B} |2\eta - 1| \, dP_X \\
&= \sum_{j\in J^r_B} \int_{(X_1 \triangle \{h_{P,s}\geq 0\})\cap A_j} |2\eta - 1| \, dP_X \\
&= 0,
\end{aligned} \tag{21}$$

since $(X_1 \triangle \{h_{P,s} \geq 0\}) \cap A_j = \emptyset$ for each $j \in J^r_B$. To see the latter, we first remark that the latter set contains those $x \in A_j$ for that either $h_{P,s}(x) \geq 0$ and

$\eta(x) \leq 1/2$ or $h_{P,s}(x) < 0$ and $\eta(x) > 1/2$. Since we have $A_j \subset X_{-1} \cup X_1$ we can ignore the case $\eta(x) = 1/2$. Furthermore, we know by Lemma 3.1(ii) that either $A_j \cap X_{-1} = \emptyset$ or $A_j \cap X_1 = \emptyset$. Let us first consider the case $A_j \cap X_{-1} = \emptyset$ and thus $A_j \subset X_1$. According to the definition of the histogram rule, c.f. (1), we find for all $x \in A_j$ that $h_{P,s}(x) = 1$, since

$$\begin{aligned}
&f_{P,s}(x)\\
&= P(A_j(x) \times \{1\}) - P(A_j(x) \times \{-1\})\\
&= \int_{A_j} \int_Y \mathbf{1}_{A_j \times \{1\}}(x,y) P(dy|x)\, dP_X(x)\\
&\qquad - \int_{A_j} \int_Y \mathbf{1}_{A_j \times \{-1\}}(x,y) P(dy|x)\, dP_X(x)\\
&= \int_{A_j} \mathbf{1}_{A_j \times \{1\}}(x,1)\eta(x)\, dP_X(x) - \int_{A_j} \mathbf{1}_{A_j \times \{-1\}}(x,-1)(1-\eta(x))\, dP_X(x)\\
&= \int_{A_j} 2\eta(x) - 1\, dP_X(x)\\
&\geq 0.
\end{aligned}$$

Obviously we have $\eta(x) \geq 1/2$ and $h_{P,s}(x) = 1$ for all $x \in A_j$. Analogously we can show for cells with $A_j \cap X_1 = \emptyset$ for $j \in J_B^r$ that $\eta(x) \leq 1/2$ and $h_{P,s}(x) = -1$ for all $x \in A_j$. Hence, $(X_1 \triangle \{h_{P,s} \geq 0\}) \cap A_j = \emptyset$ for all $j \in J_B^r$ and the approximation error vanishes on the set B.

Altogether, for the oracle inequality on B we obtain with (20) and (21) that

$$\mathcal{R}_{L_B,P}(h_D) - \mathcal{R}^*_{L_B,P} < \frac{32c_1(8^{d+1}s^{-d} + \tau)}{r^\gamma n} \tag{22}$$

holds with probability $P^n \geq 1 - e^{-\tau}$.

Part 2: In the second part we establish an oracle inequality for $\mathcal{R}_{L_A,P}(h_D) - \mathcal{R}^*_{L_A,P}$, again by using [7, Theorem 7.2]. Analogously to Part 1 we define $h_f^A := L_A \circ f - L_A \circ f^*_{L_A,P}$ for $f \in \mathcal{F}$ and find $\|h_f^A\|_\infty \leq 1$. Since $(h_{f_0}^A)^2 = \mathbf{1}_A \frac{|f - f^*_{L,P}|}{2} =$

$\mathbf{1}_{(X_{-1}\triangle\{f<0\})\cap A}$ we find with [Appendix, Lemma A.1] that

$$\begin{aligned}
&\mathbb{E}_P(h_{f_0}^A)^2\\
&= \frac{1}{2}\int_A |f_0(x) - f^*_{L_A,P}(x)|\, dP_X(x)\\
&= \frac{1}{2}\int_{A\cap\{|2\eta-1|\geq t\}} |f_0(x) - f^*_{L_A,P}(x)|\, dP_X(x)\\
&\qquad + \frac{1}{2}\int_{A\cap\{|2\eta-1|< t\}} |f_0(x) - f^*_{L_A,P}(x)|\, dP_X(x)\\
&\leq \frac{1}{2t}\int_{A\cap\{|2\eta-1|\geq t\}} |2\eta(x)-1||f_0(x) - f^*_{L_A,P}(x)|\, dP_X(x)\\
&\qquad + P_X(\{x\in A : |2\eta(x)-1| < t\})\\
&\leq \frac{1}{2t}\int_{A} |2\eta(x)-1||f_0(x) - f^*_{L_A,P}(x)|\, dP_X(x)\\
&\qquad + P_X(\{x\in A : |2\eta(x)-1| < t\})\\
&\leq t^{-1}\mathbb{E}_P h_{f_0}^A + \min\{P_X(A), P_X(\{x\in X : |2\eta(x)-1| < t\})\}
\end{aligned} \tag{23}$$

for all $t > 0$. We turn our attention to the minimum and note, that by the definition of A we have

$$P_X(A) \leq P_X(\{\Delta_\eta(x) \leq 3r\}). \tag{24}$$

For $x \in X$ with $|2\eta(x)-1| < t$ by the definition of the lower control we conclude from

$$\frac{\Delta_\eta^\gamma(x)}{c_{\mathrm{LC}}} \leq |2\eta(x)-1| < t.$$

that

$$\Delta_\eta(x) \leq (c_{\mathrm{LC}}t)^{\frac{1}{\gamma}}$$

and consequently

$$\{x\in X : |2\eta(x)-1| < t\} \subset \{x \in X : \Delta_\eta(x) \leq (c_{\mathrm{LC}}t)^{\frac{1}{\gamma}}\}. \tag{25}$$

Then we find by (24), (25) and by the definition of the strong margin exponent that

$$\begin{aligned}
&\min\{P_X(A), P_X(\{x\in X : |2\eta(x)-1| < t\})\}\\
&\leq \min\{P_X(\{\Delta_\eta(x)\leq 3r\}), P_X(\{x\in X : \Delta_\eta(x) \leq (c_{\mathrm{LC}}t)^{\frac{1}{\gamma}}\})\}\\
&\leq \min\{(c_{\mathrm{SME}}3r)^\alpha, c_{\mathrm{SME}}^\alpha (c_{\mathrm{LC}}t)^{\frac{\alpha}{\gamma}}\}.
\end{aligned} \tag{26}$$

Combining (26) with (23) we obtain

$$\begin{aligned}
\mathbb{E}_P(h_{f_0}^A - \mathbb{E}_P h_{f_0}^A)^2 &\leq t^{-1}\mathbb{E}_P h_{f_0}^A + \min\{(c_{\mathrm{SME}}3r)^\alpha, c_{\mathrm{SME}}^\alpha (c_{\mathrm{LC}}t)^{\frac{\alpha}{\gamma}}\}\\
&\leq t^{-1}\mathbb{E}_P h_{f_0}^A + c_{\mathrm{SME}}^\alpha (c_{\mathrm{LC}}t)^{\frac{\alpha}{\gamma}}.
\end{aligned} \tag{27}$$

Minimizing the right-hand side of (27) yields

$$\min_{t>0}\left(t^{-1}\mathbb{E}_P h_{f_0}^A + c_{\mathrm{SME}}^{\alpha}(c_{\mathrm{LC}}t)^{\frac{\alpha}{\gamma}}\right) = c_2\left(\mathbb{E}_P h_{f_0}^A\right)^{\frac{\alpha}{\alpha+\gamma}},$$

where $c_2 := \frac{\alpha+\gamma}{\gamma} c_{\mathrm{SME}}^{\frac{\alpha\gamma}{\alpha+\gamma}}\left(\frac{\gamma c_{\mathrm{LC}}}{\alpha}\right)^{\frac{\alpha}{\alpha+\gamma}}$, such that with

$$V := \max\{1, c_2\} \tag{28}$$

and (18) we have

$$\mathbb{E}_P(h_{f_0}^A)^2 \le t^{-1}\mathbb{E}_P h_{f_0}^A + c_{\mathrm{SME}}(c_\gamma t^{\frac{1}{\gamma}})^\alpha = c_2\left(\mathbb{E}_P h_{f_0}^A\right)^{\frac{\alpha}{\alpha+\gamma}} \le V\left(\mathbb{E}_P h_{f_0}^A\right)^{\theta}. \tag{29}$$

Note, that the definition of V yields $V^{\frac{1}{2-\theta}} \ge 1$. Since h_D is an ERM over $\mathcal{F}$ for the loss L_A due to Lemma 3.2, by using [7, Theorem 7.2] we obtain for fixed $\tau \ge 1$ and $n \ge 1$ that

$$\begin{aligned}&\mathcal{R}_{L_A,P}(h_D) - \mathcal{R}^*_{L_A,P}\\ &< 6(\mathcal{R}^*_{L_A,P,\mathcal{F}} - \mathcal{R}^*_{L_A,P}) + 4\left(\frac{8V(\log(|\mathcal{F}|+1)+\tau)}{n}\right)^{\frac{1}{2-\theta}}\end{aligned} \tag{30}$$

holds with probability $P^n \ge 1-e^{-\tau}$. In order to refine the right-hand side in (30), we establish a bound on the cardinality $|\mathcal{F}| = 2^{|J_A|}$ and on the approximation error. To bound the mentioned cardinality we use the fact that A lies in a tube around the decision line, that is $\bigcup_{j\in J_A} A_j \subset \{\Delta_\eta(x) \le 3r\}$, see (6). We remark that $3r \le \delta^*$ holds, where δ^* is the constant from Lemma 2.1, since with assumption (10) we have

$$3r = 3\tilde{c}_{\alpha,\gamma,d}\left(\frac{\tau}{s^d n}\right)^{\frac{1-\theta}{1+\gamma(2-\theta)}} \le 3\min\left\{\frac{\delta^*}{3}, 1\right\} \le \delta^*.$$

Then, with Lemma 2.1 we find that

$$\lambda^d(\{\Delta_\eta(x) \le 3r\}) \le 12\mathcal{H}^{d-1}(\{\eta = 1/2\})r$$

and we obtain

$$|J_A|s^d = \lambda^d\left(\bigcup_{j\in J_A} A_j\right) \le \lambda^d(\{\Delta_\eta(x) \le 3r\}) \le 12\mathcal{H}^{d-1}(\{\eta = 1/2\})r.$$

This yields to

$$|J_A| \le 12\mathcal{H}^{d-1}(\{\eta = 1/2\})rs^{-d} = c_3 rs^{-d},$$

where $c_3 := 12\mathcal{H}^{d-1}(\{\eta = 1/2\})$. By $r \geq s \geq s^d$ we hence conclude that

$$\begin{aligned}
\log(|\mathcal{F}|+1) &= \log(2^{c_3 r s^{-d}} + 1) \\
&\leq \log(2 \cdot 2^{c_3 r s^{-d}}) \\
&= \log(2^{c_3 r s^{-d}+1}) \\
&= (c_3 r s^{-d} + 1)\log(2) \\
&\leq c_3 r s^{-d} + r s^{-d} \\
&\leq c_4 r s^{-d},
\end{aligned} \tag{31}$$

where $c_4 := 2\max\{12\mathcal{H}^{d-1}(\{\eta = 1/2\}), 1\}$. Thus (30) changes to

$$\mathcal{R}_{L_A,P}(h_D) - \mathcal{R}^*_{L_A,P} \leq 6(\mathcal{R}^*_{L_A,P,\mathcal{F}} - \mathcal{R}^*_{L_A,P}) + 4\left(\frac{8V(c_4 r s^{-d} + \tau)}{n}\right)^{\frac{1}{2-\theta}} \tag{32}$$

with probability $P^n \geq 1 - e^{-\tau}$.

Finally, we have to bound the *approximation error* $\mathcal{R}^*_{L_A,P,\mathcal{F}} - \mathcal{R}^*_{L_A,P}$ in (32). For $f_0 = h_{P,s}$ we have with Lemma A.1 that

$$\begin{aligned}
\mathcal{R}_{L_A,P}(h_{P,s}) - \mathcal{R}^*_{L_A,P} &= \int_{(X_1 \triangle \{h_{P,s} \geq 0\}) \cap A} |2\eta - 1| \, dP_X \\
&= \sum_{j \in J^r_A} \int_{(X_1 \triangle \{h_{P,s} \geq 0\}) \cap A_j} |2\eta - 1| \, dP_X.
\end{aligned}$$

We split J^r_A in indices where cells do not intersect the decision line and those which do by

$$\begin{aligned}
J^r_{A_1} &:= \{\, j \in J^r_A \mid P_X(A_j \cap X_1) = 0 \vee P_X(A_j \cap X_{-1}) = 0 \,\} \\
J^r_{A_2} &:= \{\, j \in J^r_A \mid P_X(A_j \cap X_1) > 0 \wedge P_X(A_j \cap X_{-1}) > 0 \,\}.
\end{aligned}$$

such that

$$\begin{aligned}
&\sum_{j \in J^r_A} \int_{(X_1 \triangle \{h_{P,s} \geq 0\}) \cap A_j} |2\eta - 1| \, dP_X \\
&= \sum_{j \in J^r_{A_1}} \int_{(X_1 \triangle \{h_{P,s} \geq 0\}) \cap A_j} |2\eta - 1| \, dP_X \\
&\quad + \sum_{j \in J^r_{A_2}} \int_{(X_1 \triangle \{h_{P,s} \geq 0\}) \cap A_j} |2\eta - 1| \, dP_X.
\end{aligned}$$

We notice that, as in the calculation of the approximation error in Part 1, the first sum vanishes, since $(X_1 \Delta \{h_{P,s} \geq 0\}) \cap A_j = \emptyset$ for all $j \in J^r_{A_1}$. Moreover, we remark that $J^r_{A_2}$ only contains cells of width s that intersect the decision boundary. Hence, by using the margin-noise assumption we find

$$\begin{aligned}\mathcal{R}_{L_A,P}(h_{P,s}) - \mathcal{R}^*_{L_A,P} &= \sum_{j\in J^r_{A_2}} \int_{(X_1\triangle\{h_{P,s}\geq 0\})\cap A_j} |2\eta-1|\,dP_X \\ &\leq \int_{\{\Delta(x)\leq s\}} |2\eta-1|\,dP_X \\ &\leq (c_{\mathrm{MNE}}s)^\beta\,.\end{aligned} \tag{33}$$

Altogether for the oracle inequality on A with (32) we find that

$$\mathcal{R}_{L_A,P}(h_D) - \mathcal{R}^*_{L_A,P} \leq 6\,(c_{\mathrm{MNE}}s)^\beta + 4\left(\frac{8V(c_4 r s^{-d}+\tau)}{n}\right)^{\frac{1}{2-\theta}} \tag{34}$$

holds with with probability $P^n \geq 1-e^{-\tau}$.

Part 3: In the last part we combine the results obtained in Part 1, the oracle inequality on B and Part 2, the oracle inequality on A. That means, with the separation in (8) we obtain with (22) and (34) for the oracle inequality on X that

$$\begin{aligned}&\mathcal{R}_{L,P}(h_{D,s}) - \mathcal{R}^*_{L,P} \\ &\leq \left(\mathcal{R}_{L_A,P}(h_{D,s}) - \mathcal{R}^*_{L_A,P}\right) + \left(\mathcal{R}_{L_B,P}(h_{D,s}) - \mathcal{R}^*_{L_B,P}\right) \\ &\leq 6\,(c_{\mathrm{MNE}}s)^\beta + 4\left(\frac{8V(c_4 r s^{-d}+\tau)}{n}\right)^{\frac{1}{2-\theta}} + \frac{32c_1(8^{d+1}s^{-d}+\tau)}{r^\gamma n}\end{aligned} \tag{35}$$

holds with probability $P^n \geq 1-2e^{-\tau}$. Since $s\in(0,1]$ and $r\geq s$, we find that $rs^{-d}\geq 1$. Together with the fact $s^{-d},\tau\geq 1$ and $c_4\geq 1$ it follows that

$$\begin{aligned}&\mathcal{R}_{L,P}(h_{D,s}) - \mathcal{R}^*_{L,P} \\ &\leq 6\,(c_{\mathrm{MNE}}s)^\beta + 4\left(\frac{8V(c_4 r s^{-d}+\tau)}{n}\right)^{\frac{1}{2-\theta}} + \frac{32c_1(8^{d+1}s^{-d}+\tau)}{r^\gamma n} \\ &\leq 6\,(c_{\mathrm{MNE}}s)^\beta + 4\left(\frac{8V(c_4\tau r s^{-d}+c_4\tau r s^{-d})}{n}\right)^{\frac{1}{2-\theta}} + \frac{32c_1(8^{d+1}\tau s^{-d}+\tau s^{-d})}{r^\gamma n} \\ &\leq 6\,(c_{\mathrm{MNE}}s)^\beta + 4\left(\frac{c_5\tau r s^{-d}}{n}\right)^{\frac{1}{2-\theta}} + \frac{c_6\tau s^{-d}}{r^\gamma n} \\ &\leq 6\,(c_{\mathrm{MNE}}s)^\beta + r^{\frac{1}{2-\theta}}4\left(\frac{c_5\tau}{s^d n}\right)^{\frac{1}{2-\theta}} + \frac{c_6\tau}{r^\gamma s^d n},\end{aligned}$$

where $c_5 := 24V\max\{12\mathcal{H}^{d-1}(\{\eta=1/2\}),1\}$ and $c_6 := 64\cdot 8^{d+1}\max\{c_{LC},2^\gamma\}$. Thus, inserting r, defined in (17), with the choice of $\tilde{c}_{\alpha,\gamma,d} :=$

$\left(\frac{(\gamma(2-\theta)c_6)^{2-\theta}}{4^{2-\theta}c_5}\right)^{\frac{1}{1+\gamma(2-\theta)}}$ minimizes the right-hand side and yields to

$$\begin{aligned}
&\mathcal{R}_{L,P}(h_{D,s}) - \mathcal{R}^*_{L,P}\\
&\leq 6\,(c_{\mathrm{MNE}}s)^{\beta} + r^{\frac{1}{2-\theta}}4\left(\frac{c_5\tau}{s^d n}\right)^{\frac{1}{2-\theta}} + \frac{c_6\tau}{r^{\gamma}s^d n}\\
&= 6\,(c_{\mathrm{MNE}}s)^{\beta} + 4(\tilde{c}_{\alpha,\gamma,d}c_5)^{\frac{1}{2-\theta}}\left(\frac{\tau}{s^d n}\right)^{\frac{2-\theta+\gamma(2-\theta)}{(1+\gamma(2-\theta))(2-\theta)}} + \frac{c_6}{\tilde{c}^{\gamma}_{\alpha,\gamma,d}}\left(\frac{\tau}{s^d n}\right)^{\frac{1+\gamma}{1+\gamma(2-\theta)}}\\
&= 6\,(c_{\mathrm{MNE}}s)^{\beta} + \left(\frac{\tilde{c}_{\alpha,\gamma,d}^{\frac{1+\gamma(2-\theta)}{2-\theta}}4c_5^{\frac{1}{2-\theta}} + c_6}{\tilde{c}^{\gamma}_{\alpha,\gamma,d}}\right)\left(\frac{\tau}{s^d n}\right)^{\frac{1+\gamma}{1+\gamma(2-\theta)}}\\
&= 6\,(c_{\mathrm{MNE}}s)^{\beta} + \left(\frac{\gamma(2-\theta)c_6 + c_6}{\tilde{c}^{\gamma}_{\alpha,\gamma,d}}\right)\left(\frac{\tau}{s^d n}\right)^{\frac{1+\gamma}{1+\gamma(2-\theta)}}\\
&\leq 6\,(c_{\mathrm{MNE}}s)^{\beta} + \left(\frac{2c_6\max\{\gamma(2-\theta),1\}}{\tilde{c}^{\gamma}_{\alpha,\gamma,d}}\right)\left(\frac{\tau}{s^d n}\right)^{\frac{1+\gamma}{1+\gamma(2-\theta)}}
\end{aligned}$$

and we find again by inserting θ that

$$\mathcal{R}_{L,P}(h_{D,s}) - \mathcal{R}^*_{L,P} \leq 6\,(c_{\mathrm{MNE}}s)^{\beta} + c_{\alpha,\gamma,d}\left(\frac{\tau}{s^d n}\right)^{\frac{(1+\gamma)(\alpha+\gamma)}{(1+\gamma)(\alpha+\gamma)+\gamma^2}} \tag{36}$$

holds with probability $P^n \geq 1-2e^{-\tau}$, where $c_{\alpha,\gamma,d} := \frac{2c_6\max\{\gamma(2-\theta),1\}}{\tilde{c}^{\gamma}_{\alpha,\gamma,d}} = \frac{2c_6\max\left\{\frac{\gamma(\alpha+2\gamma)}{\alpha+\gamma},1\right\}}{\tilde{c}^{\gamma}_{\alpha,\gamma,d}}$. □

Proof of Theorem 3.5: We begin by proving that the chosen sequence s_n satisfies assumptions (9) and (10). To this end, we define $n_{\tau,\alpha,\beta,\gamma,d} := \left(\frac{\tilde{c}_{\alpha,\beta,\gamma,\tau,d}}{c_1}\right)^{\frac{1}{\zeta_1}}$ with $c_1 := \tilde{c}_{\alpha,\gamma,d}^{\frac{\kappa+\gamma^2}{\kappa+\gamma^2+d\gamma}}\tau^{\frac{\gamma}{\kappa+\gamma^2+d\gamma}}$, where $\tilde{c}_{\alpha,\gamma,d}$ is the constant from Theorem 3.4, and $\zeta_1 := \frac{\kappa(\kappa+\gamma^2+d\gamma)-\gamma(\beta(\kappa+\gamma^2)+d\kappa)}{(\beta(\kappa+\gamma^2)+d\kappa)(\kappa+\gamma^2+d\gamma)}$. We remark that $\zeta_1 \geq 0$ since we find by $\beta \leq \gamma^{-1}(1+\gamma)(\alpha+\gamma)$ that

$$\begin{aligned}
\kappa(\kappa+\gamma^2+d\gamma) - \gamma(\beta(\kappa+\gamma^2)+d\kappa) &= \kappa^2 + k\gamma^2 - \gamma\beta\kappa - \beta\gamma^3\\
&\geq \kappa^2 + \kappa\gamma^2 - \kappa^2 - \kappa\gamma^2\\
&= 0.
\end{aligned}$$

Then, for $n \geq n_{\tau,\alpha,\beta,\gamma,d}$ a simple calculation shows that the latter is equivalent to

$$c_1 n^{\frac{-\gamma}{\kappa+\gamma^2+d\gamma}} \geq \tilde{c}_{\alpha,\beta,\gamma,\tau,d}n^{-\frac{\kappa}{\beta(\kappa+\gamma^2)+d\kappa}},$$

which equals assumption (9) with $s_n := \tilde{c}_{\alpha,\beta,\gamma,\tau,d}n^{-\frac{\kappa}{\beta(\kappa+\gamma^2)+d\kappa}}$. To see that assumption (10) is satisfied we define $\tilde{n}_{\tau,\alpha,\beta,\gamma,d} := \left(\frac{c_2}{\tilde{c}_{\alpha,\beta,\gamma,\tau,d}}\right)^{\frac{1}{\zeta_2}}$ with $c_2 :=$

$\tau^{\frac{1}{d}} \left(\frac{\tilde{c}_{\alpha,\gamma,d}}{\min\{\frac{\delta^*}{3},1\}} \right)^{\frac{\kappa+\gamma^2}{d\gamma}}$, where $\tilde{c}_{\alpha,\gamma,d}$ is the constant from Theorem 3.4, δ^* the one from Lemma 2.1 and where $\zeta_2 := \frac{\beta(\kappa+\gamma^2)}{d(\beta(\kappa+\gamma^2)+d\kappa)}$. Then, a simple transformation shows again that for all $n \geq \tilde{n}_{\tau,\alpha,\beta,\gamma,d}$ we find

$$\tilde{c}_{\alpha,\beta,\gamma,\tau,d} n^{-\frac{\kappa}{(\beta(\kappa+\gamma^2)+d\kappa)}} \geq c_2 n^{-1/d},$$

which equals assumption (10) with $s_n := \tilde{c}_{\alpha,\beta,\gamma,\tau,d} n^{-\frac{\kappa}{(\beta(\kappa+\gamma^2)+d\kappa)}}$.

Finally, we obtain for all $n \geq n_0 := \lceil \max\{n_{\tau,\alpha,\beta,\gamma,d}, \tilde{n}_{\tau,\alpha,\beta,\gamma,d}\} \rceil$ by inserting our chosen sequence s_n, satisfying (9) and (10), in (11) that

$$\begin{aligned}
&\mathcal{R}_{L,P}(h_{D,s_n}) - \mathcal{R}^*_{L,P} \\
&\leq 6(c_{\mathrm{MNE}} s)^\beta + c_{\alpha,\gamma,d} \left(\frac{\tau}{s^d n} \right)^{\frac{\kappa}{\kappa+\gamma^2}} \\
&= 6 c^\beta_{\mathrm{MNE}} \tilde{c}^\beta_{\alpha,\beta,\gamma,\tau,d} n^{-\frac{\beta\kappa}{\beta(\kappa+\gamma^2)+d\kappa}} + c_{\alpha,\gamma,d} \tau^{\frac{\kappa}{\kappa+\gamma^2}} \tilde{c}^{-\frac{d\kappa}{\kappa+\gamma^2}}_{\alpha,\beta,\gamma,\tau,d} n^{-\frac{\beta\kappa}{\beta(\kappa+\gamma^2)+d\kappa}} \\
&= \left(\frac{6 c^\beta_{\mathrm{MNE}} \tilde{c}^{\frac{\beta(\kappa+\gamma^2)+d\kappa}{\kappa+\gamma^2}}_{\alpha,\beta,\gamma,\tau,d} + c_{\alpha,\gamma,d} \tau^{\frac{\kappa}{\kappa+\gamma^2}}}{\tilde{c}^{\frac{d\kappa}{\kappa+\gamma^2}}_{\alpha,\beta,\gamma,\tau,d}} \right) n^{-\frac{\beta\kappa}{\beta(\kappa+\gamma^2)+d\kappa}} \\
&= \left(\frac{\frac{d\kappa}{\beta(\kappa+\gamma^2)} c_{\alpha,\gamma,d} \tau^{\frac{\kappa}{\kappa+\gamma^2}} + c_{\alpha,\gamma,d} \tau^{\frac{\kappa}{\kappa+\gamma^2}}}{\tilde{c}^{\frac{d\kappa}{\kappa+\gamma^2}}_{\alpha,\beta,\gamma,\tau,d}} \right) n^{-\frac{\beta\kappa}{\beta(\kappa+\gamma^2)+d\kappa}} \\
&\leq \left(\frac{2 \max\left\{ \frac{d\kappa}{\beta(\kappa+\gamma^2)}, 1 \right\} c_{\alpha,\gamma,\delta} \tau^{\frac{\kappa}{\kappa+\gamma^2}}}{\tilde{c}^{\frac{d\kappa}{\kappa+\gamma^2}}_{\alpha,\beta,\gamma,\tau,d}} \right) n^{-\frac{\beta\kappa}{\beta(\kappa+\gamma^2)+d\kappa}} \\
&= c_{\alpha,\beta,\gamma,\tau,d} n^{-\frac{\beta\kappa}{\beta(\kappa+\gamma^2)+d\kappa}}
\end{aligned}$$

holds with probability $P^n \geq 1 - 2e^{-\tau}$, where $c_{\alpha,\beta,\gamma,\tau,d} :=$ $2 \max\left\{ \frac{d\kappa}{\beta(\kappa+\gamma^2)}, 1 \right\} c_{\alpha,\gamma,\delta} \tau^{\frac{\kappa}{\kappa+\gamma^2}} \cdot \tilde{c}^{-\frac{d\kappa}{\kappa+\gamma^2}}_{\alpha,\beta,\gamma,\tau,d}$. □

Appendix A: Appendix

Lemma A.1. *Let $Y := \{-1, 1\}$ and P be a probability measure on $X \times Y$. For $\eta(x) := P(y = 1|x), x \in X$ define the set $X_1 := \{ x \in X \,|\, \eta(x) > 1/2 \}$. Let L be the classification loss and consider for $A \subset X$ the loss $L_A(x, y, t) := \mathbf{1}_A(x) L(y, t)$, where $y \in Y, t \in \mathbb{R}$. For a measurable $f\colon X \to \mathbb{R}$ we then have*

$$\mathcal{R}_{L_A,P}(f) - \mathcal{R}^*_{L_A,P} = \int_{(X_1 \triangle \{f \geq 0\}) \cap A} |2\eta(x) - 1| \, dP_X(x),$$

where $\triangle$ denotes the symmetric difference.

Proof of Lemma A.1: It is well known, e.g., [7, Example 3.8], that

$$\begin{aligned} &\mathcal{R}_{L_A,P}(f) - \mathcal{R}^*_{L_A,P} \\ &= \int_A |2\eta(x)-1| \cdot \mathbf{1}_{(-\infty,0)}((2\eta(x)-1)\mathrm{sign} f(x))\, dP_X(x). \end{aligned} \tag{37}$$

Next, for P_X-almost all $x \in A$ we have

$$\mathbf{1}_{(-\infty,0]}((2\eta(x)-1)\mathrm{sign} f(x)) = 1 \Leftrightarrow (2\eta(x)-1)\mathrm{sign} f(x) \le 0.$$

The latter is true if for $x \in A$ holds that $f(x) < 0$ and $\eta(x) > 1/2$ or that $f(x) \ge 0$ and $\eta(x) \le 1/2$ or that $\eta(x) = 1/2$. However, for $\eta(x) = 1/2$ we have $|2\eta(x) - 1| = 0$ and hence this case can be ignored. Then, the latter obviously equals the set $(X_1 \triangle \{f \ge 0\}) \cap A$ and we obtain in (37)

$$\mathcal{R}_{L_A,P}(f) - \mathcal{R}^*_{L_A,P} = \int_{(X_1 \triangle \{f\ge 0\})\cap A} |2\eta(x)-1|\, dP_X(x). \qquad \square$$

Lemma A.2. *Let $X := [-1,1]^d$ and P be a probability measure on $X \times \{-1,1\}$ with fixed version $\eta\colon X \to [0,1]$ of its posterior probability. Then, if η is Hoelder-continuous with exponent γ, we have that Δ_η controls the noise from below with exponent γ.*

Proof of Lemma A.2: Fix w.l.o.g. an $x \in X_1$. Then, $\eta(x) > 1/2$. Since η is Hoelder-continuous with exponent γ, there exists a constant $c > 0$ such that we have

$$|2\eta(x)-1| = 2|\eta(x) - 1/2| \le 2|\eta(x) - n(x')| \le 2c(d(x,x'))^\gamma$$

for all $x' \in X_{-1}$ and hence

$$|2\eta(x)-1| \le 2c \inf_{\tilde{x}\in X_{-1}} (d(x,\tilde{x}))^\gamma = 2c\Delta^\gamma_\eta(x).$$

Obviously the last inequality holds immediately for $x \in X$ with $\eta(x) = 1/2$. $\square$

References

[1] Audibert, J. Y. and Tsybakov, A. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.* **35** 608–633. MR2336861

[2] Binev, P., Cohen, A., Dahmen, W. and DeVore, R. (2014). Classification algorithms using adaptive partitioning. *Ann. Statist* **42** 2141–2163. MR3269976

[3] Breiman, L. (2001). Random Forests. *Machine Learning* **45** 5–32.

[4] Devroye, L., Györfi, L. and Lugosi, L. (1996). *A Probabilistc Theory of Pattern Recognition*. Springer. MR1383093

[5] Federer, H. (1969). *Geometric measure theory*. Springer. MR0257325

[6] Steinwart, I. (2015). Fully adaptive density-based clustering. *Ann. Statist.* **43** 2132–2167. MR3396981

[7] Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer. MR2450103

Ingrid Blaschzyk
Fachbereich Mathematik, Universität Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany
E-Mail: ingrid.blaschzyk@mathematik.uni-stuttgart.de

Ingo Steinwart
Fachbereich Mathematik, Universität Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany
E-Mail: steinwart@mathematik.uni-stuttgart.de
WWW: http://www.isa.uni-stuttgart.de/Steinwart/

Erschienene Preprints ab Nummer 2012-001

Komplette Liste: `http://www.mathematik.uni-stuttgart.de/preprints`

2016-004 *Blaschzyk, I.; Steinwart, I.:* Improved Classification Rates under Refined Margin Conditions

2016-003 *Feistauer, M.; Roskovec, F.; Sändig, AM.:* Discontinuous Galerkin Method for an Elliptic Problem with Nonlinear Newton Boundary Conditions in a Polygon

2016-002 *Steinwart, I.:* A Short Note on the Comparison of Interpolation Widths, Entropy Numbers, and Kolmogorov Widths

2016-001 *Köster, I.:* Sylow Numbers in Spectral Tables

2015-016 *Hang, H.; Steinwart, I.:* A Bernstein-type Inequality for Some Mixing Processes and Dynamical Systems with an Application to Learning

2015-015 *Steinwart, I.:* Representation of Quasi-Monotone Functionals by Families of Separating Hyperplanes

2015-014 *Muhammad, F.; Steinwart, I.:* An SVM-like Approach for Expectile Regression

2015-013 *Nava-Yazdani, E.:* Splines and geometric mean for data in geodesic spaces

2015-012 *Kimmerle, W.; Köster, I.:* Sylow Numbers from Character Tables and Group Rings

2015-011 *Györfi, L.; Walk, H.:* On the asymptotic normality of an estimate of a regression functional

2015-010 *Gorodski, C, Kollross, A.:* Some remarks on polar actions

2015-009 *Apprich, C.; Höllig, K.; Hörner, J.; Reif, U.:* Collocation with WEB–Splines

2015-008 *Kabil, B.; Rodrigues, M.:* Spectral Validation of the Whitham Equations for Periodic Waves of Lattice Dynamical Systems

2015-007 *Kollross, A.:* Hyperpolar actions on reducible symmetric spaces

2015-006 *Schmid, J.; Griesemer, M.:* Well-posedness of Non-autonomous Linear Evolution Equations in Uniformly Convex Spaces

2015-005 *Hinrichs, A.; Markhasin, L.; Oettershagen, J.; Ullrich, T.:* Optimal quasi-Monte Carlo rules on higher order digital nets for the numerical integration of multivariate periodic functions

2015-004 *Kutter, M.; Rohde, C.; Sändig, A.-M.:* Well-Posedness of a Two Scale Model for Liquid Phase Epitaxy with Elasticity

2015-003 *Rossi, E.; Schleper, V.:* Convergence of a numerical scheme for a mixed hyperbolic-parabolic system in two space dimensions

2015-002 *Döring, M.; Györfi, L.; Walk, H.:* Exact rate of convergence of kernel-based classification rule

2015-001 *Kohler, M.; Müller, F.; Walk, H.:* Estimation of a regression function corresponding to latent variables

2014-021 *Neusser, J.; Rohde, C.; Schleper, V.:* Relaxed Navier-Stokes-Korteweg Equations for Compressible Two-Phase Flow with Phase Transition

2014-020 *Kabil, B.; Rohde, C.:* Persistence of undercompressive phase boundaries for isothermal Euler equations including configurational forces and surface tension

2014-019 *Bilyk, D.; Markhasin, L.:* BMO and exponential Orlicz space estimates of the discrepancy function in arbitrary dimension

2014-018 *Schmid, J.:* Well-posedness of non-autonomous linear evolution equations for generators whose commutators are scalar

2014-017 *Margolis, L.:* A Sylow theorem for the integral group ring of $PSL(2, q)$

2014-016 *Rybak, I.; Magiera, J.; Helmig, R.; Rohde, C.:* Multirate time integration for coupled saturated/unsaturated porous medium and free flow systems

2014-015 *Gaspoz, F.D.; Heine, C.-J.; Siebert, K.G.:* Optimal Grading of the Newest Vertex Bisection and H^1-Stability of the L_2-Projection

2014-014 *Kohler, M.; Krzyżak, A.; Walk, H.:* Nonparametric recursive quantile estimation

2014-013 *Kohler, M.; Krzyżak, A.; Tent, R.; Walk, H.:* Nonparametric quantile estimation using importance sampling

2014-012 *Györfi, L.; Ottucsák, G.; Walk, H.:* The growth optimal investment strategy is secure, too.

2014-011 *Györfi, L.; Walk, H.:* Strongly consistent detection for nonparametric hypotheses

2014-010 *Köster, I.:* Finite Groups with Sylow numbers $\{q^x, a, b\}$

2014-009 *Kahnert, D.:* Hausdorff Dimension of Rings

2014-008 *Steinwart, I.:* Measuring the Capacity of Sets of Functions in the Analysis of ERM

2014-007 *Steinwart, I.:* Convergence Types and Rates in Generic Karhunen-Loève Expansions with Applications to Sample Path Properties

2014-006 *Steinwart, I.; Pasin, C.; Williamson, R.; Zhang, S.:* Elicitation and Identification of Properties

2014-005 *Schmid, J.; Griesemer, M.:* Integration of Non-Autonomous Linear Evolution Equations

2014-004 *Markhasin, L.:* L_2- and $S^r_{p,q}B$-discrepancy of (order 2) digital nets

2014-003 *Markhasin, L.:* Discrepancy and integration in function spaces with dominating mixed smoothness

2014-002 *Eberts, M.; Steinwart, I.:* Optimal Learning Rates for Localized SVMs

2014-001 *Giesselmann, J.:* A relative entropy approach to convergence of a low order approximation to a nonlinear elasticity model with viscosity and capillarity

2013-016 *Steinwart, I.:* Fully Adaptive Density-Based Clustering

2013-015 *Steinwart, I.:* Some Remarks on the Statistical Analysis of SVMs and Related Methods

2013-014 *Rohde, C.; Zeiler, C.:* A Relaxation Riemann Solver for Compressible Two-Phase Flow with Phase Transition and Surface Tension

2013-013 *Moroianu, A.; Semmelmann, U.:* Generalized Killling spinors on Einstein manifolds

2013-012 *Moroianu, A.; Semmelmann, U.:* Generalized Killing Spinors on Spheres

2013-011 *Kohls, K; Rösch, A.; Siebert, K.G.:* Convergence of Adaptive Finite Elements for Control Constrained Optimal Control Problems

2013-010 *Corli, A.; Rohde, C.; Schleper, V.:* Parabolic Approximations of Diffusive-Dispersive Equations

2013-009 *Nava-Yazdani, E.; Polthier, K.:* De Casteljau's Algorithm on Manifolds

2013-008 *Bächle, A.; Margolis, L.:* Rational conjugacy of torsion units in integral group rings of non-solvable groups

2013-007 *Knarr, N.; Stroppel, M.J.:* Heisenberg groups over composition algebras

2013-006 *Knarr, N.; Stroppel, M.J.:* Heisenberg groups, semifields, and translation planes

2013-005 *Eck, C.; Kutter, M.; Sändig, A.-M.; Rohde, C.:* A Two Scale Model for Liquid Phase Epitaxy with Elasticity: An Iterative Procedure

2013-004 *Griesemer, M.; Wellig, D.:* The Strong-Coupling Polaron in Electromagnetic Fields

2013-003 *Kabil, B.; Rohde, C.:* The Influence of Surface Tension and Configurational Forces on the Stability of Liquid-Vapor Interfaces

2013-002 *Devroye, L.; Ferrario, P.G.; Györfi, L.; Walk, H.:* Strong universal consistent estimate of the minimum mean squared error

2013-001 *Kohls, K.; Rösch, A.; Siebert, K.G.:* A Posteriori Error Analysis of Optimal Control Problems with Control Constraints

2012-013 *Diaz Ramos, J.C.; Dominguez Vázquez, M.; Kollross, A.:* Polar actions on complex hyperbolic spaces

2012-012 *Moroianu; A.; Semmelmann, U.:* Weakly complex homogeneous spaces

2012-011 *Moroianu; A.; Semmelmann, U.:* Invariant four-forms and symmetric pairs

2012-010 *Hamilton, M.J.D.:* The closure of the symplectic cone of elliptic surfaces

2012-009 *Hamilton, M.J.D.:* Iterated fibre sums of algebraic Lefschetz fibrations

2012-008 *Hamilton, M.J.D.:* The minimal genus problem for elliptic surfaces

2012-007 *Ferrario, P.:* Partitioning estimation of local variance based on nearest neighbors under censoring

2012-006 *Stroppel, M.:* Buttons, Holes and Loops of String: Lacing the Doily

2012-005 *Hantsch, F.:* Existence of Minimizers in Restricted Hartree-Fock Theory

2012-004 *Grundhöfer, T.; Stroppel, M.; Van Maldeghem, H.:* Unitals admitting all translations

2012-003 *Hamilton, M.J.D.:* Representing homology classes by symplectic surfaces

2012-002 *Hamilton, M.J.D.:* On certain exotic 4-manifolds of Akhmedov and Park

2012-001 *Jentsch, T.:* Parallel submanifolds of the real 2-Grassmannian