Journal of Machine Learning Research 23 (2022) 1-59 Submitted 5/19; Revised 10/21; Published 7/22

Improved Classification Rates for Localized SVMs

Ingrid Blaschzyk INGRID.BLASCHZYK@MATHEMATIK.UNI-STUTTGART.DE
Ingo Steinwart INGO.STEINWART@MATHEMATIK.UNI-STUTTGART.DE
Institute for Stochastics and Applications
University of Stuttgart
70569 Stuttgart, Germany

Editor: Kenji Fukumizu

Abstract

Localized support vector machines solve SVMs on many spatially defined small chunks and besides their computational benefit compared to global SVMs one of their main characteristics is the freedom of choosing arbitrary kernel and regularization parameter on each cell. We take advantage of this observation to derive *global* learning rates for localized SVMs with Gaussian kernels and hinge loss. It turns out that our rates outperform under suitable sets of assumptions known classification rates for localized SVMs, for global SVMs, and other learning algorithms based on e.g., plug-in rules or trees. The localized SVM rates are achieved under a set of margin conditions, which describe the behavior of the data-generating distribution, and no assumption on the existence of a density is made. Moreover, we show that our rates are obtained adaptively, that is without knowing the margin parameters in advance. The statistical analysis of the excess risk relies on a simple partitioning based technique, which splits the input space into a subset that is close to the decision boundary and into a subset that is sufficiently far away. A crucial condition to derive then improved global rates is a margin condition that relates the distance to the decision boundary to the amount of noise.

Keywords: classification, margin conditions, hinge loss, support vector machines, spatial decomposition, Gaussian kernel

1. Introduction

Experimental results show that support vector machines (SVMs) handle small- and medium-sized datasets in supervised learning tasks (see Fernandez-Delgado et al., 2014; Meister and Steinwart, 2016; Thomann et al., 2017; Klambauer et al., 2017). Recently, it was shown that they even outperform self-normalizing neural-networks (SNNs) for such datasets (see Klambauer et al., 2017). However, many learning tasks, e.g., diagnostics of diseases on patient data, demand learning methods that handle large-scale datasets, where observations have high dimensions and/or the number of observations is large. At this point global SVMs and more generally kernel methods suffer from their computational complexity, which for SVMs is at least quadratically in space and time. To reduce this complexity Meister and Steinwart (2016) propose a data decomposition strategy, called localized SVMs, which solve SVMs on many *spatially* defined chunks and which leads to improved time and space complexities. Experimental results with `liquidSVM` (Steinwart and Thomann, 2017) show that localized

SVMs can even tackle datasets with 32 million of training samples (see Thomann et al., 2017).

One aspect of this paper is to show that localized SVMs do not only have computational advantages compared to global SVMs, as shown in (Thomann et al., 2017). Under reasonable assumptions we prove theoretically an advantage of those methods in terms of (faster) learning rates. Localized SVM *global* (classfication) rates have already been proven in (Thomann et al., 2017) but the results only showed that the learning rates match the global SVM rates and roughly speaking that one do not loose accuracy by localization. Our results in this paper instead verify a common intuition: Due to localization, classifiers should adapt better to the underlying target function compared to classifiers produced by global algorithms. Intuitively, this should reflect in theoretical results and this is exactly what we prove with our results in terms of *global* learning rates. The existing global convergence rates can be improved by assuming certain margin parameters, which describe the behaviour of the data-generating distribution and by applying localized training on decomposed input spaces. Further below, we describe these parameters and the applied technique in more detail. But first, we share a methodical overview.

In literature many approaches have been proposed that aim to reduce complexities for large-scale datasets. Approaches that as well base on data partitioning are for example random chunks (see Bottou and Vapnik, 1992), or distributed learning algorithms (see Zhang et al., 2015; Lin et al., 2017a; Mücke and Blanchard, 2018). Some other approaches apply strategies that are based on gradient approximation such as stochastic gradient decent algorithms (e.g., Lin and Rosasco, 2017; Pillaud-Vivien et al., 2018), or that are based on kernel matrices approximations such as Nyström Method (e.g., Williams and Seeger, 2001; Rudi et al., 2015), or Random Fourier Features (e.g., Rahimi and Recht, 2008; Rudi and Rosasco, 2017).

Most of the papers previously mentioned have in common that the authors usually consider least-squares regression in their theoretical analysis. Furthermore, they assume that the true regression function f^* is contained in some subspace of the chosen reproducing kernel Hilbert space H. Since for Gaussian kernels we have $[L_2, H]_{\beta,2} \subset C^\infty$ with $0 < \beta < 1$ the latter is a rather hard assumption for this type of kernel. Typically, the regularization parameters of these algorithms as well as the learning rates depend on unknown parameters, as the smoothness parameter or as a parameter that describes the decay of the eigenvalues of the associated kernel operator. Algorithms that are adaptive to these parameters are solely sketched if given and adaptive rates are not yet proven theoretically.

It is well known that least-squares regression results can be used to obtain learning rates for classification by applying a so-called calibration inequality in which the least-square regression classifier serves as a plug-in rule. In other words, to solve a classification problem a regression problem is solved in a first step. Fast rates can be achieved if a margin condition, namely the Tsybakov noise exponent (see Mammen and Tsybakov, 1999) is used. This exponent measures the amount of noise in the input space, where noise equals the probability of wrongly labelling some given input $x \in X$. Under the assumption of Tsybakov's noise exponent and some smoothness assumption on the regression function, fast rates for plug-in classifier are achieved in (Audibert and Tsybakov, 2007; Kohler and Krzyzak, 2007), and (Belkin et al., 2018) or for tree-based classifiers in (Binev et al., 2014). Some of the mentioned authors additionally make assumptions on the density of the marginal distribution

P_X to improve their rates that were achieved without density assumptions. However, it is well known that boundedness assumptions on the density of P_X together with smoothness and noise exponent assumptions substantially limit the class of considered distributions (see Audibert and Tsybakov, 2007; Kohler and Krzyzak, 2007; Binev et al., 2014).

Another possibility to solve the (binary) classification problem is to solve the optimization problem with respect to the hinge loss. This loss function has a natural connection to the classification loss, see Zhang's inequality (e.g., Steinwart and Christmann, 2008, Theorem 2.31). In fact, the Bayes classification function $f^*_{L_{\text{class}},P} := \text{sign}(2P(y = 1|\cdot) - 1)$ is a minimizer of the hinge excess risk. Clearly, this is not the case for the least-square excess risk. Following this path learning rates for SVMs are achieved in (Steinwart and Scovel, 2007; Steinwart and Christmann, 2008; Lin et al., 2017b; Thomann et al., 2017) without assumptions on the existence on the density of P_X or smoothness assumptions on the Bayes decision function, but with an additional margin condition that takes also the amount of mass around the decision boundary into consideration.

In this paper, we investigate the statistical properties of localized SVM classifiers with Gaussian kernels and hinge loss. Recently, it became more popular to analyze localized SVMs theoretically. For example, based on possible overlapping regions or decomposition with k-nearest neighbor universal consistency and/or robustness for these classifiers is shown in (Dumpert and Christmann, 2018) and (Hable, 2013). Meister and Steinwart (2016) achieve optimal learning rates for localized SVMs with Gaussian kernel and least-squares loss, whereas Mücke (2019) presents rates for localized SVMs with least-squares loss under assumptions on the eigenvalue decay of the kernel integral operator. Moreover, Thomann et al. (2017) derive learning rates for localized SVMs with Gaussian kernel and hinge loss. In contrast, there exist also several results that are rather experimentally investigated such as localized SVMs with a partition based on clusters (Cheng et al., 2007), decision trees (Bennett and Blue, 1998), or k-nearest-neighbors (Zhang et al., 2006).

Our aim is to derive global learning rates for localized SVMs with Gaussian kernel and hinge loss under a set of margin conditions. We show that these outperform or match under suitable assumptions the rates of several learning algorithms mentioned in the previous paragraphs. It turns out that the improvements result essentially from a margin condition that relates the distance to the decision boundary to the amount of noise. Descriptively, this condition restricts the location of noise, that means, if we have noise for some $x \in X$, this x has to be close to the decision boundary. Note that Blaschzyk and Steinwart (2018) showed recently under this condition together with a mild regularity assumption that rates for the simple histogram rule can be obtained, which under a suitable set of assumptions even outperform known rates for global SVMs.

To obtain classification rates for localized SVMs under margin conditions, we derive finite sample bounds on the excess classification risk by applying the splitting technique developed in (Blaschzyk and Steinwart, 2018). That means, we split the input space into two sets that depend on a splitting parameter $s > 0$, one that is close to the decision boundary and one that is sufficiently far away from the decision boundary, and analyze the excess risk separately on these sets. By a standard decomposition into a stochastic and an approximation error we derive in a first step local finite sample bounds. Compared to the technique in (Blaschzyk and Steinwart, 2018), which is applied on the histogram rule, we make the observation that it is necessary to refine the analysis of the approximation

error for localized SVMs. More precisely, on the set that has cells sufficiently close to the decision boundary we have to distinguish carefully between cells that intersect the decision boundary and those that do not. The result is an analysis on three rather than two sets. Based on local finite sample bounds on these sets we derive rates by taking advantage of the great flexibility local SVMs enable us by definition, that is, that kernel and regularization parameters can be chosen on each cell individually. Remarkably, our analysis shows that the individual regularization parameters $\lambda_{n,i}$ for $i \in \{1, \ldots, m\}$ have only a marginal effect on the stochastic error and that they may decay with an arbitrary polynomial rate without effecting the overall error rate. *Note that this effect also occurs in global SVM error bounds (e.g., Steinwart and Christmann, 2008, Theorem 7.23) but has been overlooked for several years.* By choosing in a final step the splitting parameter s appropriately, we then derive global learning rates that depend on the margin parameters. We emphasize that the splitting parameter s is no additional hyper-parameter and appears in the proof solely. Moreover, we show that training validation support vector machines (TV-SVMs) achieve the same learning rates adaptively, that is, without knowing the margin parameters in advance.

The paper is organized as follows. In Section 2 we briefly describe the localized SVM ansatz, introduce notation and close with theoretical assumptions. Section 3 is divided up into several subsections: In Section 3.1 we present our main result followed by a detailed discussion on the choice of parameters and on some proof details in Section 3.2. We show that the presented learning rates can be achieved adaptively in Section 3.3. In Section 3.4 we compare our rates carefully with other known rates. The proofs of our main results are contained in Section 4. The results on individual sets, that is, bounds on the approximation error, oracle inequalities and learning rates, on predefined sets can be found in Subsection 4.4.1 up to Subsection 4.4.3. Some results on margin conditions and some technical results can be found in the appendix.

2. Preliminaries

Given a dataset $D := ((x_1, y_1), \ldots, (x_n, y_n))$ of observations, where $y_i \in Y := \{-1, 1\}$, the learning target in classification is to find a decision function $f_D \colon X \to Y$ such that for new data (x, y) we have $f_D(x) = y$ with high probability. We assume that $x_i \in B_{\ell_2^d}$, where $B_{\ell_2^d}$ denotes the closed unit ball of the d-dimensional Euclidean space ℓ_2^d and assume that our data D is generated independently and identically by a probability measure P on $\mathbb{R}^d \times Y$. We denote by P_X the marginal distribution on $\mathbb{R}^d$, write $X := \operatorname{supp}(P_X)$, and assume $X \subset B_{\ell_2^d}$ and $P_X(\partial X) = 0$.

We briefly describe the localized SVM approach in a generalized manner. Given a dataset D local SVMs construct a function f_D by solving SVMs on spatially defined small chunks of D. To be more precise, let $\mathcal{A} := (A_j)_{j=1,\ldots,m}$ be an arbitrary partition of $B_{\ell_2^d}$. We define for every $j \in \{1, \ldots, m\}$ the index set

$$I_j := \{i \in \{1, \ldots, n\} : x_i \in A_j\}$$

with $\sum_{j=1}^m |I_j| = n$, that indicates the samples of D contained in A_j and we define the corresponding local data set D_j by

$$D_j := ((x_i, y_i) \in D : i \in I_j)\,.$$

Then, one learns an individual SVM on *each* cell by solving the optimization problem

$$f_{D_j,\lambda_j} = \operatorname*{arg\,min}_{f\in H_j} \lambda_j \|f\|_{H_j}^2 + \frac{1}{n} \sum_{x_i,y_i\in D_j} L(x_i, y_i, f(x_i)) \tag{1}$$

for every $j \in \{1,\dots,m\}$, where $\lambda_j > 0$ is a regularization parameter, where H_j is a reproducing kernel Hilbert space (RKHS) over A_j with arbitrary reproducing kernel $k_j : A_j \times A_j \to \mathbb{R}$ (see Steinwart and Christmann, 2008, Chap. 4), and where $L : X \times Y \times \mathbb{R} \to [0,\infty)$ is a measurable function, called loss function, describing our learning goal. The final decision function $f_{D,\boldsymbol{\lambda}} : X \to \mathbb{R}$ is then defined by

$$f_{D,\boldsymbol{\lambda}}(x) := \sum_{j=1}^{m} \mathbf{1}_{A_j}(x) f_{D_j,\lambda_j}(x), \tag{2}$$

where $\boldsymbol{\lambda} := (\lambda_1,\dots,\lambda_m) \in (0,\infty)^m$. We make the following assumptions.

(H) For every $j \in \{1,\dots,m\}$ let $k_j : A_j \times A_j \to \mathbb{R}$ be the Gaussian kernel with width $\gamma_j > 0$, defined by

$$k_{\gamma_j}(x,x') := \exp\left(-\gamma_j^{-2}\|x-x'\|_2^2\right), \tag{3}$$

with corresponding RKHS H_{γ_j} over A_j and denote by $\hat{H}_j := \{\mathbf{1}_{A_j} f : f \in H_j\}$ the extended RKHS over $B_{\ell_2^d}$. For some $J \subset \{1,\dots,m\}$ define the joint RKHS H_J over $B_{\ell_2^d}$ by $H_J := \bigoplus_{j\in J} \hat{H}_j$, see (Meister and Steinwart, 2016, Sec. 3).

We write $f_{D_j,\lambda_j,\gamma_j}$ for the local SVM predictor in (1) to remember its local dependency on the kernel parameter γ_j and the regularization parameter λ_j on each cell A_j for $j \in \{1,\dots,m\}$. Clearly, we are free to choose different kernel and regularization parameters on each cell, since the predictors in (1) are computed *independently* on each cell. Moreover, we write $f_{D,\boldsymbol{\lambda},\boldsymbol{\gamma}}$ for the final decision function in (2), where $\boldsymbol{\gamma} := (\gamma_1,\dots,\gamma_m) \in (0,\infty)^m$. Note that we have immediately $f_{D,\boldsymbol{\lambda},\boldsymbol{\gamma}} \in H_J$ for $J = \{1,\dots,m\}$ since $\mathbf{1}_{A_j} f_{D_j,\lambda_j,\gamma_j} \in \hat{H}_j$ for every $j \in J$. To measure the quality of the predictor locally, we define a (local) loss $L_j : X \times Y \times \mathbb{R} \to [0,\infty)$ by

$$L_j(x,y,t) := \mathbf{1}_{A_j}(x) L(x,y,t).$$

Moreover, we define for an arbitrary index set $J \subset \{1,\dots,m\}$ the set $T := \bigcup_{j\in J} A_j$ and the associated loss $L_{J_T} : X \times Y \times \mathbb{R} \to [0,\infty)$ by

$$L_{J_T}(x,y,t) := \mathbf{1}_T(x) L(x,y,t),$$

where we sometimes use the abbreviation $L_T := L_{J_T}$ to avoid multiple subscripts. A typical loss function is the classification loss $L_{\text{class}} : Y \times \mathbb{R} \to [0,\infty)$, defined by

$$L_{\text{class}}(y,t) := \mathbf{1}_{(-\infty,0]}(y \operatorname{sign} t),$$

where sign $0 := 1$. For (local) SVMs the optimization problem is not solvable for the classification loss. A suitable convex surrogate is for example the hinge loss $L_{\text{hinge}} : Y \times \mathbb{R} \to [0, \infty)$, defined by

$$L_{\text{hinge}}(y, t) := \max\{0, 1 - yt\}$$

for $y = \pm 1$, $t \in \mathbb{R}$. Note that for convex losses the existence and uniqueness of (1) are secured, see e.g. (Steinwart and Christmann, 2008, Chap. 5.1) or (Meister and Steinwart, 2016). Since we are not interested in the loss of single labels, we consider the expected loss and define for a loss function L the L-risk of a measurable function $f : X \to \mathbb{R}$ by

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

Moreover, we define the optimal L-risk, called Bayes risk, with respect to P and L, by

$$\mathcal{R}^*_{L,P} := \inf \{\mathcal{R}_{L,P}(f) \mid f : X \to \mathbb{R} \text{ measurable}\}$$

and call a function $f^*_{L,P} : X \to \mathbb{R}$ attaining the infimum, Bayes decision function. For the classification loss, a Bayes decision function is given by $f^*_{L_{\text{class}},P}(x) := \text{sign}(2P(y = 1|x) - 1), x \in X$. A well-known result by Zhang (see Steinwart and Christmann, 2008, Theorem 2.31) shows that the excess classification-risk is bounded by the excess hinge-risk, that is,

$$\mathcal{R}_{L_{\text{class}},P}(f) - \mathcal{R}^*_{L_{\text{class}},P} \leq \mathcal{R}_{L_{\text{hinge}},P}(f) - \mathcal{R}^*_{L_{\text{hinge}},P}$$

for all functions $f : X \to \mathbb{R}$. Hence, we restrict our analysis to the hinge loss and we write in the following $L := L_{\text{hinge}}$. Since a short calculation in (Steinwart and Christmann, 2008, Example 2.27) shows that

$$L(y, \max\{-1, \min\{f(x), 1\}) \leq L(y, f(x))$$

for all $f : X \to \mathbb{R}$ and $y \in \{-1, 1\}$, it suffices to consider the loss and thus the risk for functions values restricted to the interval $[-1, 1]$. Thus, we define the clipping operator by

$$\widehat{t} := \max\{-1, \min\{t, 1\}\}$$

for $t \in \mathbb{R}$, which restricts values of t to $[-1, 1]$ (see Steinwart and Christmann, 2008, Chap. 2.2). For our decision function in (2) this means that the clipped decision function $\widehat{f}_{D,\boldsymbol{\lambda},\boldsymbol{\gamma}} : X \to [-1, 1]$ is then defined by the sum of the clipped empirical solutions $\widehat{f}_{D_j,\lambda_j,\gamma_j}$ since for all $x \in X$ there is exactly one $f_{D_j,\lambda_j,\gamma_j}$ with $f_{D_j,\lambda_j,\gamma_j}(x) \neq 0$.

In order to derive learning rates for the localized SVM predictor in (2) that measure the speed of convergence of the excess risk $\mathcal{R}_{L,P}(f_{D,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L,P}$ it is necessary to specify our partition $\mathcal{A}$. To this end, we denote the ball with radius $r > 0$ and center $s \in B_{\ell_2^d}$ by $B_r(s) := \{\, t \in \mathbb{R}^d \,|\, \|t - s\|_2 \leq r \,\}$ with Euclidean norm $\|\cdot\|_2$ in $\mathbb{R}^d$ and we define the radius r_A of a set $A \subset B_{\ell_2^d}$ by

$$r_A = \inf\{r > 0 : \exists s \in B_{\ell_2^d} \text{ such that } A \subset B_r(s)\}.$$

(A) Let $\mathcal{A} := (A_j)_{j=1,\ldots,m}$ be a partition of $B_{\ell_2^d}$ and $r > 0$ such that we have $\mathring{A}_j \neq \emptyset$ for every $j \in \{1,\ldots,m\}$, and such that there exist $z_1,\ldots,z_m \in B_{\ell_2^d}$ such that $A_j \subset B_r(z_j)$, and $\|z_i - z_j\|_2 \geq \frac{r}{2}$, $i \neq j$, and

$$r_{A_j} < r \leq 16m^{-\frac{1}{d}}, \qquad \text{f.a. } j \in \{1,\ldots,m\} \tag{4}$$

are satisfied.

Note that if one considers a Voronoi partition $(A_j)_{j=1,\ldots,m}$ of $B_{\ell_2^d}$ based on a r-net $z_1,\ldots,z_m \in B_{\ell_2^d}$ with $r \leq 16m^{-\frac{1}{d}}$ and $\|z_i - z_j\|_2 \geq \frac{r}{2}$, $i \neq j$, the assumptions above are immediately satisfied (see Meister and Steinwart, 2016).

Besides the assumption on the partition above, we need some assumptions on the probability measure P itself. To this end, we recall some notions from (Steinwart and Christmann, 2008, Chap. 8). 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, which means that the probability measures $P(\,\cdot\,|x)$ form a regular conditional probability of P. Clearly, if we have $\eta(x) = 0$ resp. $\eta(x) = 1$ for $x \in X$ we observe the label $y = -1$ resp. $y = 1$ with probability 1. Otherwise, if e.g., $\eta(x) \in (1/2,1)$ we observe the label $y = -1$ with the probability $1 - \eta(x) \in (0,1/2)$ and we call the latter probability noise. In the worst case we observe both labels with equal probability $1/2$ and we define the set containing the corresponding $x \in X$ by $X_0 := \{\, x \in X\colon \eta(x) = 1/2 \,\}$. Furthermore, we write

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

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

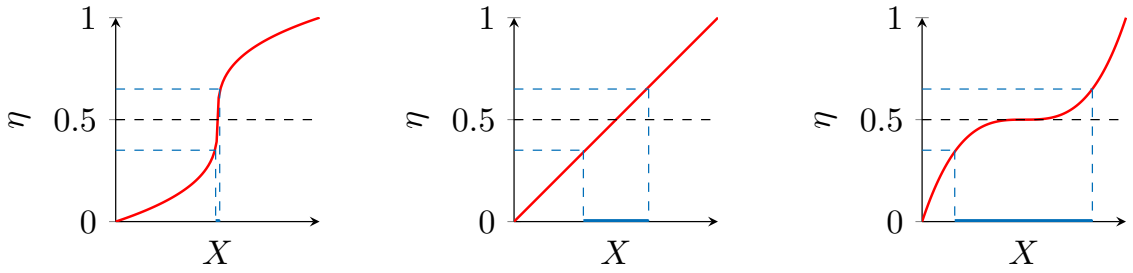
$$\Delta_\eta(x) := \begin{cases} d(x,X_1) & \text{if } x \in X_{-1}, \\ d(x,X_{-1}) & \text{if } x \in X_1, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $d(x,A) := \inf_{x' \in A} d(x,x')$, is called distance to the decision boundary.

In the following, we introduce various exponents that describe the behavior of the data-generating distribution P and that are typically used to derive learning rates in classification. The probably most known exponent is the (Tsybakov) noise exponent introduced in (Mammen and Tsybakov, 1999). We say that P has (Tsybakov) noise exponent (NE) $q \in [0,\infty]$ if there exist a constant $c_{\mathrm{NE}} > 0$ such that

$$P_X(\{x \in X : |2\eta(x) - 1| < \varepsilon\}) \leq (c_{\mathrm{NE}}\varepsilon)^q \tag{6}$$

for all $\varepsilon > 0$ (c.f. Steinwart and Christmann, 2008, Def. 8.22). Note that a common name for (6) is margin exponent but we call it noise exponent since it measures the amount of critical noise and does *not* locate the noise. Obviously, in the best case P has NE $q = \infty$ and hence, η is bounded away from $1/2$.



(a) $\zeta < 1$ (b) $\zeta = 1$ (c) $\zeta > 1$

Figure 1: Three examples of η (red line) with different values of LC ζ. The blue bar on the x-axis shows the set containing x of X with noise close to the critical level $1/2$. Figure from (Blaschzyk, 2020, Figure 2.3).

Moreover, we introduce an exponent that describes the existence of noise around the decision boundary. We say that P has margin-noise exponent (MNE) $\beta \in (0, \infty]$ if there exists a version η and a constant $c_{\text{MNE}} > 0$ such that

$$\int_{\{\Delta_\eta(x)<t\}} |2\eta(x) - 1| \, dP_X(x) \leq (c_{\text{MNE}} t)^\beta \tag{7}$$

for all $t > 0$. That is, we have a large margin-noise exponent, if we have low mass and/or a large amount of noise around the decision boundary. In Section 3.4, we introduce another exponent that is closely related to the margin-noise exponent, the so-called margin exponent. Since we do not apply this exponent in our analysis and since we need it only for comparison reasons we skip its definition here.

Next, we introduce an exponent that explicitly relates the amount of noise to its location. We say that the distance to the decision boundary Δ_η controls the noise from below if there exist a $\zeta \in [0, \infty)$, a version η, and a constant $c_{\text{LC}} > 0$ such that

$$\Delta_\eta^\zeta(x) \leq c_{\text{LC}} |2\eta(x) - 1| \tag{8}$$

for P_X-almost all $x \in X$. Descriptively, if $\eta(x)$ is close to $1/2$ for some $x \in X$, then this x is close to the decision boundary. In Figure 1, we describe the dependence of different values of $\eta(x)$ and the location of large noise which is close to the critical level $1/2$.

Moreover, a simple calculation in (8) shows that η is bounded away from $1/2$ if $\zeta = 0$ (see Blaschzyk, 2020, Section 2.3.1). This indicates that small values of ζ are maybe preferable for learning and in Section 3.2 we will see that this is indeed the case. In Appendix A, we show that the lower control condition can be related to the reverse Hölder continuity of η and that it can be related to the noise exponent if P has some margin exponent, see (26). For a more detailed description of the lower control condition (8) we refer the reader to (Blaschzyk, 2020, Section 2.3.1) and to (Blaschzyk and Steinwart, 2018, Fig. 1).

For examples of typical values of the exponents above and relations between them we refer the reader to (Steinwart and Christmann, 2008, Chap. 8 and Exercises), (Blaschzyk and Steinwart, 2018, Section 4), (Hamm and Steinwart, 2021, Ex. 4.5) and to Section 3.4.

Finally, we define some mild geometrical assumption on the decision boundary. To this end, we say according to (Federer, 1969, Sec. 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. Furthermore, we denote by $\partial_X T$ the relative boundary of T in X and we denote by $\mathcal{H}^{d-1}$ the $(d-1)$-dimensional Hausdorff measure on $\mathbb{R}^d$, see (Federer, 1969, Introduction). Then, we state the following assumptions on the decision boundary.

(G) Let $\eta : X \to [0,1]$ be a fixed version of the posterior probability of P. Let $X_0 = \partial_X X_1 = \partial_X X_{-1}$ and let X_0 be $(d-1)$-rectifiable with $\mathcal{H}^{d-1}(X_0) > 0$.

Remember that under assumption **(G)** we have $\mathcal{H}^{d-1}(X_0) < \infty$. In particular, Blaschzyk and Steinwart (2018) show under assumption **(G)** how to measure the d-dimensional Lebesgue measure λ^d of a certain set in the vicinity of the decision boundary. More precisely, Blaschzyk and Steinwart (2018, Lemma 2.1) show that there exists a $\delta^* > 0$ and a constant $c_d > 0$ such that

$$\lambda^d(\{\Delta_\eta(x) \leq \delta\}) \leq c_d \cdot \delta, \qquad \text{f.a. } \delta \in (0, \delta^*]. \tag{9}$$

We remark that for some sequences $a_n, b_n \in \mathbb{R}$ we write $a_n \simeq b_n$ if there exists constants $c_1, c_2 > 0$ such that $a_n \leq c_1 b_n$ and $a_n \geq c_2 b_n$ for sufficiently large n.

3. Classification Rates

3.1 Learning Rates for localized SVMs

In this section, we derive *global* learning rates for local SVMs with Gaussian kernel and hinge loss. We apply the splitting technique developed in (Blaschzyk and Steinwart, 2018), that is, we analyze the excess risk separately on overlapping sets that consists of cells that are close to and sufficiently far away from the decision boundary. By choosing *individual* kernel parameters on these sets we obtain local learning rates that we balance out in a last step to derive global learning rates. To this end, we define for $s > 0$ and a fixed version η of the posterior probability of P the set of indices of cells *near* the decision boundary by

$$J_N^s := \{\, j \in \{1, \ldots, m\} \,|\, \forall\, x \in A_j : \Delta_\eta(x) \leq 3s \,\}$$

and the set of indices of cells that are sufficiently *far* away by

$$J_F^s := \{\, j \in \{1, \ldots, m\} \,|\, \forall\, x \in A_j : \Delta_\eta(x) \geq s \,\}.$$

Moreover, we write

$$N^s := \bigcup_{j \in J_N^s} A_j \quad \text{and} \quad F^s := \bigcup_{j \in J_F^s} A_j. \tag{10}$$

Clearly, by dividing our input space into the two overlapping sets defined above we have to be sure to capture all cells in the input space and to assign the cells in F^s either to the class X_{-1} or to X_1. The following lemma gives a sufficient condition on our separation parameter s. Since the proof is almost identical to the one in (Blaschzyk and Steinwart, 2018, Lemma 3.1) we skip it here.

Lemma 1 *Let $(A_j)_{j=1,\ldots,m}$ be a partition of $B_{\ell_d^2}$ such that for every $j \in \{1, \ldots, m\}$ we have $\mathring{A}_j \neq \emptyset$ and (4) is satisfied for some $r > 0$. For $s \geq r$ define the sets N^s and F^s by (10). Moreover, let $X_0 = \partial_X X_1 = \partial_X X_{-1}$. Then, we have*

i) $X \subset N^s \cup F^s$,

ii) either $A_j \cap X_1 = \emptyset$ *or* $A_j \cap X_{-1} = \emptyset$ *for all* $j \in J_F^s$.

To prevent notational overload, we omit in the sets (of indices) defined above the dependence on s for the rest of this paper, while keeping in mind that all sets depend on this separation parameter.

Based on an analysis on the sets defined above, we present in the subsequent theorem our main result that yields global learning rates for localized SVMs under margin conditions. The result is remarkable in three ways. First, it shows that the regularization parameters $\lambda_{n,i}$ can decay arbitrarily fast in a polynomial way to achieve the presented rate, see also Section 3.2 for more details. Second, we will see in Section 3.3 that the theoretical rate presented below can be achieved adaptively for a subclass of the considered P. Third, we will show that localized SVMs do not only have computational advantages compared to global SVMs, but the theoretical rates outperform rates of SVMs and of other algorithms under suitable sets of assumptions, see Section 3.4.

Theorem 2 *Let P be a probability measure on $\mathbb{R}^d \times \{-1,1\}$ for which P has MNE $\beta \in (0,\infty]$, NE $q \in [0,\infty]$ and LC $\zeta \in [0,\infty)$ and let **(G)** be satisfied for one η. Define $\kappa := \frac{q+1}{\beta(q+2)+d(q+1)}$. Let assumption **(A)** be satisfied for m_n and define*

$$r_n := n^{-\nu},$$

where ν satisfies

$$\nu \leq \begin{cases} \frac{\kappa}{1-\kappa} & \text{if } \beta \geq (q+1)(1+\max\{d,\zeta\}-d), \\ \frac{1-\beta\kappa}{\beta\kappa+\max\{d,\zeta\}} & \text{else,} \end{cases} \tag{11}$$

*and assume that **(H)** holds. Define for $J = \{1,\dots,m_n\}$ the set of indices*

$$J_{N_1} := \{\, j \in J \,|\, \forall x \in A_j : \Delta_\eta(x) \leq 3r_n \text{ and } P_X(A_j \cap X_1) > 0 \text{ and } P_X(A_j \cap X_{-1}) > 0 \,\},$$

as well as

$$\begin{aligned} \gamma_{n,j} &\simeq \begin{cases} r_n^\kappa n^{-\kappa} & \text{for } j \in J_{N_1}, \\ r_n & \text{else,} \end{cases} \\ \lambda_{n,j} &\simeq n^{-\sigma} \end{aligned} \tag{12}$$

for some $\sigma \geq 1$ and for every $j \in J$. Moreover, let $\tau \geq 1$ be fixed and define for δ^* considered in (9), $n^* := \max\{4, \left(4^{-1}\delta^*\right)^{-\frac{1}{\nu}}, \left(4^{-1}\delta^*\right)^{-\frac{1}{\alpha}}\}$. Then, for all $\varepsilon > 0$ there exists a constant $c_{\beta,d,\varepsilon,\sigma,q} > 0$ such that for all $n \geq n^*$ the localized SVM classifier satisfies

$$\mathcal{R}_{L_J,P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_J,P} \leq c_{\beta,d,\varepsilon,\sigma,q}\tau \cdot n^{-\beta\kappa(\nu+1)+\varepsilon} \tag{13}$$

with probability P^n not less than $1 - 9e^{-\tau}$.

Remark 3 *The exponents of r_n in* (11) *match for $\beta = (q+1)(1+\max\{d,\zeta\}-d)$. Moreover, a short calculation shows that in the case $\beta > (q+1)(1+\max\{d,\zeta\}-d)$ the best possible rate is achieved for $\nu := \frac{\kappa}{1-\kappa}$ and equals*

$$n^{-\frac{\beta(q+1)}{\beta(q+2)+(d-1)(q+1)}+\varepsilon}$$

In the other case, $\beta < (q+1)(1+\max\{d,\zeta\}-d)$, the best possible rate is achieved for $\nu := \frac{1-\beta\kappa}{\beta\kappa+\max\{d,\zeta\}}$ and equals

$$n^{-\frac{\beta\kappa[1+\max\{d,\zeta\}]}{\beta\kappa+\max\{d,\zeta\}}+\varepsilon}.$$

Remark 4 *The rates in Theorem 2 are better the smaller we choose the cell sizes r_n. The smaller r_n the more cells m_n are considered and training localized SVMs is more efficient. To be more precise, the complexity of the kernel matrices or the time complexity of the solver are reduced (see Thomann et al., 2017). However,* (11) *gives a lower bound on $r_n = n^{-\nu}$. For smaller r_n we do achieve rates for localized SVMs, but, we cannot ensure that they learn with the rate* (13)*. Indeed, we achieve slower rates. We illustrate this for the case $\beta < (q+1)(1+\max\{d,\zeta\}-d)$. The proof of Theorem 2 shows that if*

$$\left(\frac{q+1}{q+2}\right)(1+\max\{d,\zeta\}-d) < \beta < (q+1)(1+\max\{d,\zeta\}-d)$$

we can choose some $\nu \in \left[\frac{1-\beta\kappa}{\beta\kappa+\max\{d,\zeta\}}, \frac{\kappa}{1-\kappa}\right]$ such that the localized SVM classifier learns for some $\varepsilon > 0$ with rate

$$n^{-(1-\nu\max\{d,\zeta\})+\varepsilon}.$$

A short calculation shows that this rate is indeed slower than (13) *for the given range of ν and matches the rate in* (13) *only for $\nu := \frac{1-\beta\kappa}{\beta\kappa+\max\{d,\zeta\}}$. In the worst case, that is, $\nu := \frac{\kappa}{1-\kappa}$ the rate equals*

$$n^{-(1-\nu\max\{d,\zeta\})} = n^{-\left(1-\frac{\kappa\max\{d,\zeta\}}{1-\kappa}\right)} = n^{-\left(1-\frac{(q+1)\max\{d,\zeta\}}{\beta(q+2)+(d-1)(q+1)}\right)} = n^{-\frac{\beta(q+2)+(q+1)(d-1-\max\{d,\zeta\})}{\beta(q+2)+(d-1)(q+1)}}$$

up to ε in the exponent, where the numerator is positive since $\beta > \left(\frac{q+1}{q+2}\right)(1+\max\{d,\zeta\}-d)$.

3.2 Proof Details and Choice of Kernel and Regularization Parameters

To follow the arguments that lead to the various choices of the parameters in (11) and (12) we give a brief overview of the main effects influencing the proof of the previous theorem and of the main steps we have taken. Learning rates are derived from finite sample bounds on the excess risk which follow a typical decomposition into a bound on the approximation error and on the stochastic error. A key property to bound the stochastic error is to have a variance bound, that is a bound of the form

$$\mathbb{E}_P(L\circ f - L\circ f^*_{L,P})^2 \le V\cdot(\mathbb{E}_P(L\circ f - L\circ f^*_{L,P}))^\theta \tag{14}$$

with exponent $\theta \in (0,1]$ and some constant $V > 0$. Descriptively, a function whose risk is close to $f^*_{L,P}$ has low variance. Clearly, the best exponent is $\theta = 1$ and is obtained e.g., for the least-squares loss (see Steinwart and Christmann, 2008, Example 7.3). Moreover, under the assumption that P has NE $q \in [0,\infty]$ it can be shown for the hinge loss that $\theta = \frac{q}{q+1}$ (see Steinwart and Christmann, 2008, Theorem 8.24). Thus, we obtain $\theta = 1$ only in the special case $q = \infty$. However, we show in the next lemma that it is still possible for the hinge loss to obtain the best possible variance bound $\theta = 1$ on sets that are sufficiently far away from the decision boundary by using the lower control condition in (8). We take advantage of this fact in the analysis over the sets in (10).

Lemma 5 *Let $\eta\colon X \to [0,1]$ be a fixed version of the posterior probability of P. Assume that the associated distance to the decision boundary Δ_η controls the noise from below by the exponent $\zeta \in [0,\infty)$ and define the set $F := F^s$ as in (10). Furthermore, let $L := L_{hinge}$ be the hinge loss and let $f^*_{L,P} : X \to [-1,1]$ be a fixed Bayes decision function. Then, there exists a constant $c_{\mathrm{LC}} > 0$ independent of s such that for all measurable $f\colon X \to \mathbb{R}$ we have*

$$\mathbb{E}_P(L_F \circ \widehat{f} - L_F \circ f^*_{L,P})^2 \leq \frac{2c_{\mathrm{LC}}}{s^\zeta}\mathbb{E}_P(L_F \circ \widehat{f} - L_F \circ f^*_{L,P}).$$

Besides the stochastic error, we have to bound the approximation error. More precisely, we aim to find an appropriate $f_0 \in H_J$ such that the bound on

$$\sum_{j\in J}\lambda_j \|\mathbf{1}_{A_j} f_0\|^2_{\hat{H}_j} + \mathcal{R}_{L_J,P}(f_0) - \mathcal{R}^*_{L_J,P}$$

is small. Obviously, we control the error above if we control both, the norm and the excess risk.

The excess risk is small if $f_0 \in H_J$ is close to a Bayes decision function since its risk is then close to the Bayes risk. Note that we cannot assume the Bayes decision function to be contained in the RKHS H_J, since H_J does not contain functions that are constant on an open ball (see Steinwart and Christmann, 2008, Corollary 4.44). Nonetheless, we find a function $f_0 \in H_J$ that is similar to a Bayes decision function. To this end, we define f_0 on every cell A as the convolution of functions $K_\gamma : \mathbb{R}^d \to \mathbb{R}$ and $f \in L_2(\mathbb{R}^d)$ so that

$$(K_\gamma * f)_{|A} \in H(A),$$

and choose f as a function that is similar to a Bayes decision function on a ball containing A. Doing this, we observe the following cases. If a cell A has no intersection with the decision boundary and e.g., $A \cap X_1 \neq \emptyset$, but $A \cap X_{-1} = \emptyset$, we have for all $x \in A \cap X_1$ that $f^*_{L_{\mathrm{class}},P}(x) = 1$. Otherwise, if the cell intersects the decision boundary we find for the decision function that $f^*_{L_{\mathrm{class}},P} := \operatorname{sign}(2\eta - 1)$. Note, that this discontinuous step function makes classification harder than regression, where usually smoothness of the Bayes decision function is assumed. In order to approximate $f^*_{L_{\mathrm{class}},P}$ by the convolution above, we choose f as constant if the considered cell has no intersection with the decision boundary and as $\operatorname{sign}(2\eta-1)$ otherwise. Both depicted cases can occur on the set N, see (10), and motivates to divide the set of indices J_N into

$$\begin{aligned} J_{N_1} &:= \{\, j \in J_N \,|\, P_X(A_j \cap X_1) > 0 \text{ and } P_X(A_j \cap X_{-1}) > 0 \,\}, \\ J_{N_2} &:= \{\, j \in J_N \,|\, P_X(A_j \cap X_1) = 0 \text{ or } P_X(A_j \cap X_{-1}) = 0 \,\}, \end{aligned} \tag{15}$$

leading to an overall analysis on the corresponding sets N_1, N_2, and on the set F.

Concerning the norm, we make the observation that we are able to control the norm by choosing the regularization parameters λ_n sufficiently small on each set N_1, N_2, and F, and down below, we illustrate this for N_1.

We refer the reader for a more detailed analysis on the approximation error bounds on the sets N_1, N_2 and F to Section 4.4.1. In Sections 4.4.2 and 4.4.3, we combine the bounds on the approximation and stochastic error and present bounds on the excess risks on N_1, N_2 and F. These sections also include the learning rates for the mentioned sets and in the following, we take a closer look at those learning rates and start with the set N_1.

By applying the tools above and under the assumption that P has MNE β and NE q we derive learning rates on the set N_1 on the basis of the excess risk bound in Theorem 17, which has for some $p \in (0, \frac{1}{2})$ with high probability the form

$$\mathcal{R}_{L_{N_1},P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_{N_1},P} \preceq \frac{\lambda_n r_n}{\gamma_n^d} + \gamma_n^\beta + \left(\frac{r_n \lambda_n^{-p} \gamma_n^{-d}}{n}\right)^{\frac{q+1}{q+2-p}} + \left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}}, \tag{16}$$

where we already made some slight modifications, e.g. we set $\lambda_n := \lambda_{n,j}$ and $\gamma_n := \gamma_{n,j}$ for all $j \in J_{N_1}$ as at beginning of the proof of Theorem 18. Note that the fist two terms on the right-hand side result from bounding the approximation error, norm and excess risk, while the other terms result from bounding the stochastic error.

First, we observe that the regularization parameters λ_n influence the first and third term in the right-hand side of (16). We set $\lambda_n \simeq n^{-\sigma}$ for some $\sigma \geq 1$ and observe that λ_n in the third term can be bounded by

$$\lambda_n^{-p\left(\frac{q+1}{q+2-p}\right)} \simeq n^{p\sigma\left(\frac{q+1}{q+2-p}\right)} \leq n^{\tilde{\varepsilon}},$$

where $\tilde{\varepsilon} \geq \frac{p\sigma(q+1)}{q+2}$. This means we can choose σ arbitrary large since we are able to choose the parameter p sufficiently small. In other words, we are able to choose the decay of λ_n in an arbitrarily fast polynomial way, see (12), and the regularization parameters λ_n influence the stochastic error only marginally. However, in practice λ_n should be chosen carefully since a larger σ makes the constant in the excess risk bounds larger.

Remark 6 *The effect that the regularization parameters λ_n may decay with an arbitrarily fast polynomial rate without affecting the global error rate also occurs for global SVMs. To see that, we remark that the general oracle inequality for global SVMs given in (Steinwart and Christmann, 2008, Theorem 7.23) has exactly the form of* (16) *for $r_n = 1$ and hence, our considerations for the regularization parameters λ_n in the previous paragraph are still valid. In particular, this means that this effect even occurs for multiple types of SVM regression (with $r_n = 1$ and $q = \infty$) such as least-squares or quantile regression (see Eberts and Steinwart, 2013), or as expectile regression (Farooq and Steinwart, 2019).*

Second, in (16) we observe a different behavior in terms of γ_n. While the second term on the right-hand side tends to zero for $\gamma_n \to 0$, the bound on the stochastic error behaves in γ_n exactly the opposite way. Motivated by the approximation of the Bayes decision function discussed above we choose appropriately small kernel parameters γ_n on N_1, see (12), leading to convolutions with steep kernels.

Both explained choices for the regularization parameters λ_n and kernel parameters γ_n lead on the set N_1 with high probability to the bound

$$\mathcal{R}_{L_{N_1},P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_{N_1},P} \preceq r_n^{\beta k} n^{-\beta k}$$

for lower bounded r_n, see Theorem 18. Note that the restriction on r_n guarantees that the kernel parameters γ_n satisfy the condition $\gamma_n \leq r_n$, which is required to measure the capacity of the underlying Gaussian RKHSs by entropy numbers, see Section 4.4.

For the sets N_2 and F, which have no intersection with the decision boundary, we derive learning rates on the basis of the corresponding excess risk bounds, see Theorems 19 and 21 as described for the set N_1. Since the considerations on the behavior of λ_n we made above are also true for the bounds on the sets N_2 and F, we skip this discussion here but we remark at this point that $\lambda_n \simeq n^{-\sigma}$ with $\sigma \geq 1$ is an appropriate choice for all three sets and justifies the choice of the regularization parameters in (12). Concerning the kernel parameters γ_n, the bounds in Theorems 19 and 21 again show a trade-off in γ_n. Motivated by the approximation of the Bayes decision function discussed above we choose for the sets N_2 and F appropriately large kernel parameters γ_n leading to convolutions with flat kernels. Note that this means to choose λ_n equally to r_n, see (12).

Both explained choices for the regularization parameters λ_n and kernel parameters γ_n lead on the set N_2 and F with high probability to bounds of the form

$$\mathcal{R}_{L_{N_2},P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_{N_2},P} \preceq \left(\frac{s_n}{r_n^d}\right)^{\frac{q+1}{q+2}} n^{-\frac{q+1}{q+2}}$$

and

$$\mathcal{R}_{L_F,P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_F,P} \preceq \max\{r_n^{-d}, s_n^{-\zeta}\} \cdot n^{-1+\varepsilon},$$

see Theorems 20 and 22. We observe that these bounds do not tend to zero for $r_n \to 0$, as the bound on N_1. Moreover, both bounds depend in an opposite way on the separation parameter s_n. In (Blaschzyk and Steinwart, 2018) this is handled by a straightforward optimization over the parameter s_n. Unfortunately, in our case the optimal s^* does not fulfil the basic requirement $r_n \leq s^*$ that results from Lemma 1. We bypass this difficulty by choosing $s_n = r_n$ in the proof of our main Theorem 2. This choice has two effects. First, the rates on N_2 are always better than the rates on N_1. Second, for the rates on N_1 and F the combination of our considered margin parameters and the dimension d affects the speed of the rates. This leads to the differentiation of r_n in (11). If $\beta \geq (q+1)(1+\max\{d,\zeta\}-d)$ the rate on N_1 dominates the one on F and ν has to fulfil $\nu \leq \frac{\kappa}{1-\kappa}$. In the other case, if $\beta \leq (q+1)(1+\max\{d,\zeta\}-d)$, the rate on F dominates N_1, but only if $\nu \leq \frac{1-\beta\kappa}{\beta k+\max\{d,\zeta\}}$. Unfortunately, we find in the latter case $\frac{1-\beta\kappa}{\beta k+\max\{d,\zeta\}} \leq \frac{\kappa}{1-\kappa}$ such that r_n cannot be chosen that small as in the other case in order to learn with rate $n^{-\beta\kappa(\nu+1)}$. Larger r_n would lead to a worse learning rate. In summary, the interplay of the considered margin conditions together with the dimension d affects the rate presented in Theorem 2.

3.3 Adaptive Learning Rates for Localized SVMs

Before comparing our rates in (13) with rates obtained by other algorithms in the next section, we show that our rates are achieved adaptively by a training validation approach.

That means, without knowing the MNE β, the NE q and LC ζ in advance. To this end, we briefly describe the training validation support vector machine ansatz given in (Meister and Steinwart, 2016). We define $\Lambda := (\Lambda_n)$ and $\Gamma := (\Gamma_n)$ as sequences of finite subsets $\Lambda_n \subset (0, n^{-1}]$ and $\Gamma_n \subset (0, r_n]$. For a dataset $D := ((x_1, y_1), \ldots, (x_n, y_n))$ we define

$$D_1 := ((x_1, y_1), \ldots, (x_l, y_l)),$$
$$D_2 := ((x_{l+1}, y_{l+1}), \ldots, (x_n, y_n)),$$

where $l := \lfloor \frac{n}{2} \rfloor + 1$ and $n \geq 4$. Moreover, we split these sets into

$$D_j^{(1)} := \left((x_i, y_i)_{i \in \{1,\ldots,l\}} : x_i \in A_j\right), \qquad j \in \{1, \ldots, m_n\},$$
$$D_j^{(2)} := \left((x_i, y_i)_{i \in \{l+1,\ldots,n\}} : x_i \in A_j\right), \qquad j \in \{1, \ldots, m_n\},$$

and define $l_j := |D_j^{(1)}|$ for all $j \in \{1, \ldots, m_n\}$ such that $\sum_{j=1}^{m_n} l_j = l$. We use $D_j^{(1)}$ as a training set by computing a local SVM predictor

$$f_{D_j^{(1)}, \lambda_j, \gamma_j} := \operatorname*{arg\,min}_{f \in \hat{H}_{\gamma_j}(A_j)} \lambda_j \|f\|^2_{\hat{H}_{\gamma_j}(A_j)} + \mathcal{R}_{L_j, D_j^{(1)}}(f)$$

for every $j \in \{1, \ldots, m_n\}$. Then, we use $D_j^{(2)}$ to determine (λ_j, γ_j) by choosing a pair $(\lambda_{D_2,j}, \gamma_{D_2,j}) \in \Lambda_n \times \Gamma_n$ such that

$$\mathcal{R}_{L_j, D_2}(\widehat{f}_{D_j^{(1)}, \lambda_{D_2,j}, \gamma_{D_2,j}}) = \min_{(\lambda_j, \gamma_j) \in \Lambda_n \times \Gamma_n} \mathcal{R}_{L_j, D_j^{(2)}}(\widehat{f}_{D_j^{(1)}, \lambda_j, \gamma_j}).$$

Finally, we call the function $f_{D_1, \boldsymbol{\lambda}_{D_2}, \boldsymbol{\gamma}_{D_2}}$, defined by

$$f_{D_1, \boldsymbol{\lambda}_{D_2}, \boldsymbol{\gamma}_{D_2}} := \sum_{j=1}^{m_n} \mathbf{1}_{A_j} f_{D_j^{(1)}, \lambda_{D_2,j}, \gamma_{D_2,j}}, \tag{17}$$

training validation support vector machine (TV-SVM) w.r.t Λ and Γ. We remark that the parameter selection is performed *independently* on each cell and leads to $m \cdot |\Lambda| \cdot |\Gamma|$ many candidates. For more details we refer the reader to (Meister and Steinwart, 2016, Sec. 4.2).

The subsequent theorem shows that the TV-SVM, defined in (17), achieves the same rates as the local SVM predictor in (2).

Theorem 7 *Let the assumptions of Theorem 2 be satisfied with*

$$r_n \simeq n^{-\nu}$$

for some $\nu > 0$. Furthermore, fix an ρ_n-net $\Lambda_n \subset (0, n^{-1}]$ and an $\delta_n r_n$-net $\Gamma_n \subset (0, r_n]$ with $\rho_n \leq n^{-2}$ and $\delta_n \leq n^{-1}$. Assume that the cardinalities $|\Lambda_n|$ and $|\Gamma_n|$ grow polynomially in n. Let $\tau \geq 1$. If

$$\nu \leq \begin{cases} \frac{\kappa}{1-\kappa} & \text{if } \beta \geq (q+1)(1 + \max\{d, \zeta\} - d), \\ \frac{1-\beta\kappa}{\beta\kappa + \max\{d,\zeta\}} & \text{else.} \end{cases} \tag{18}$$

then, for all $\varepsilon > 0$ there exists a constant $c_{d,\beta,q,\varepsilon} > 0$ such that the TV-SVM, defined in (17)*, satisfies*

$$\mathcal{R}_{L_J,P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_J,P} \leq c_{d,\beta,q,\varepsilon}\tau \cdot n^{-\beta\kappa(\nu+1)+\varepsilon}$$

with probability P^n not less than $1 - e^{-\tau}$.

Remark 8 *The previous result shows two aspects. First, it reveals a trade-off between the "range of adaptivity" and "computational complexity": On the one hand, the smaller we choose ν the bigger the set of P for which we learn with the correct learning rate without knowing the margin parameters β, q and ζ. On the other hand, smaller values of ν lead to bigger cells and hence, training is more costly. This trade-off also appears in the result for Training-Validation-Voronoi-Partition-SVMs using least squares loss and Gaussian kernels in (Meister and Steinwart, 2016, Thm. 7).*

Second, it shows that by choosing ν as large as it is computationally feasible in practice, we achieve adaptivity for a large class of various distributions described by the margin parameters β, q and ζ but unfortunately not for all combinations of these margin parameters as in the global training-validation approach (see Steinwart and Christmann, 2008, Thm. 8.26). At a first glance this might seem restrictive. However, adaptivity results for common approaches that also aim to reduce complexities of certain algorithms, also need assumptions on P. For example, the eigenvalue decay of kernel operators has to be known (see Zhang et al., 2015).

3.4 Comparison of Rates

In this section, we compare the results for localized SVMs with Gaussian kernel and hinge loss from Theorem 2 to the results from various classifiers, we mentioned in the introduction. We compare the rates to the ones obtained by global and local SVMs with Gaussian kernel and hinge loss in (Thomann et al., 2017, Theorem 3.2), (Steinwart and Christmann, 2008, (8.18)) and (Lin et al., 2017b). Moreover, we make comparisons with the rates achieved by various plug-in classifier in (Kohler and Krzyzak, 2007; Audibert and Tsybakov, 2007; Binev et al., 2014; Belkin et al., 2018), and by the histogram rule in (Blaschzyk and Steinwart, 2018). We remark that in all comparisons we try to find reasonable sets of assumptions such that both, our conditions and the conditions of the compared methods are satisfied. This means in particular that our rates as well as the other rates are achieved under less assumptions. We emphasize that the rates for localized SVMs in Theorem 2 do not need an assumption on the existence of a density of the marginal distributions.

Throughout this section we assume **(A)** for some $r_n := n^{-\nu}$, **(G)** for some η, and **(H)** to be satisfied. Moreover, we denote by (i), (ii) and (iii) the following assumptions on P:

(i) P has MNE $\beta \in (0, \infty]$,

(ii) P has NE $q \in [0, \infty]$,

(iii) P has LC $\zeta \in [0, \infty)$.

Note that under the just mentioned assumptions the assumptions of Theorem 2 for localized SVMs using hinge loss are satisfied. First, we compare the rates to the known ones for local and global SVMs.

Local and global SVM. Under assumptions (i) and (ii), (Steinwart and Christmann, 2008, (8.18)) show that global SVMs using hinge loss and Gaussian kernels learn with the rate

$$n^{-\beta\kappa} = n^{-\frac{\beta(q+1)}{\beta(q+2)+d(q+1)}}. \tag{19}$$

We remark, that in the special case that (i) is satisfied for $\beta = \infty$ this rate is also achieved for the same method in (Lin et al., 2017b). The rate is also matched by localized SVMs in (Thomann et al., 2017) using hinge loss and Gaussian kernel as well as cell sizes $r_n = n^{-\nu}$ for some $\nu \leq \kappa$. We show now that under a mild additional assumption our derived rates for localized SVMs outperform the one above. To this end, we assume (iii) in addition to (i) and (ii). Then, the rate in (13) is satisfied and better by $\nu\beta\kappa$ for all ν our analysis is applied to. According to Remark 3 the improvement is at most $q+1$ in the denominator if $\beta \geq (q+1)(1+\max\{d,\zeta\}-d)$. In the other case, we obtain the fastest rate with $\nu := \frac{1-\beta\kappa}{\beta\kappa+\max\{d,\zeta\}}$ such that the exponent of our rate in (13) equals

$$\frac{\beta\kappa(1+\max\{d,\zeta\})}{\beta\kappa+\max\{d,\zeta\}} = \frac{\beta(q+1)(1+\max\{d,\zeta\})}{\beta(q+1)+\max\{d,\zeta\}(\beta(q+2)+d(q+1))} = \frac{\beta(q+1)}{\beta(q+2)+d(q+1)-\frac{\beta+d(q+1)}{1+\max\{d,\zeta\}}}. \tag{20}$$

Compared to (19) we then have at most an improvement of $\frac{\beta+d(q+1)}{1+\max\{d,\zeta\}}$ in the denominator.◀

The main improvement in the comparison above results from the strong effect of the lower-control condition (iii). Descriptively, (iii) restricts the location of noise in the sense that if we have high noise for some $x \in X$, that is $\eta(x) \approx 1/2$, then, (iii) forces this x to be located close to the decision boundary. Note that this does not mean that we have no noise far away from the decision boundary. It is still allowed to have noise $\eta(x) \in (0, 1/2-\varepsilon] \cup [1/2+\varepsilon, 1)$ for $x \in X$ and some $\varepsilon > 0$, only the case that $\eta(x) = 1/2$ is prohibited. We refer the interested reader to a more precise description of this effect to (Blaschzyk and Steinwart, 2018) and proceed with our next comparison.

In the following, we compare our result with results that make besides assumption (ii) some smoothness condition on η, namely that

(iv) η is Hölder-continuous for some $\rho \in (0, 1]$.

This assumption can be seen as a strong reverse assumption to (iii) since it implies that the distance to the decision boundary controls the noise from above, which means that there exists a ρ and a constant $\tilde{c} > 0$ such that $\tilde{c}|2\eta(x)-1| \leq \Delta_\eta^\rho(x)$ for all $x \in X$ (see Blaschzyk and Steinwart, 2018, Lemma A.2). In particular, if (iii) and (iv) are satisfied, then $\rho \leq \zeta$. Note that we observe vice versa that a reverse Hölder-continuity assumption implies (iii) if η is continuous, see Lemma 23.

If we assume (iii) in addition to (ii) and (iv) we satisfy the assumptions for localized SVMs in Theorem 2 since we find with (Blaschzyk and Steinwart, 2018, Lemma A.2) and (Steinwart and Christmann, 2008, Lemma 8.23) that the MNE equals $\beta = \rho(q+1)$. We observe that

$$\beta = \rho(q+1) \leq (q+1)(1+\max\{d,\zeta\}-d) \tag{21}$$

and according to Theorem 2 the localized SVMs learn with the rate

$$n^{-\beta\kappa(\nu+1)} \tag{22}$$

for arbitrary $\nu \leq \frac{1-\beta\kappa}{\beta\kappa+\max\{d,\zeta\}}$. In particular, this rate is upper bounded by

$$n^{-\beta\kappa(\nu+1)} < n^{-\beta\kappa} = n^{-\frac{\rho(q+1)}{\rho(q+2)+d}}. \tag{23}$$

Plug-in classifier I. Under assumption (ii), (iv) and the assumption that the support of the marginal distribution P_X is included in a compact set, the so called "Hybrid" plug-in classifiers in (Audibert and Tsybakov, 2007, Eq. (4.1)) learn with the optimal rate

$$n^{-\frac{\rho(q+1)}{\rho(q+2)+d}} \tag{24}$$

(see Audibert and Tsybakov, 2007, Theorem 4.3). If we assume in addition (iii), the localized SVM rate again equals (22) and satisfies (23) such that our rate is faster for arbitrary $\nu \leq \frac{1-\beta\kappa}{\beta\kappa+\max\{d,\zeta\}}$. For $\nu := \frac{1-\beta\kappa}{\beta\kappa+\max\{d,\zeta\}}$ we find for the exponent in (22) that

$$\beta\kappa(\nu+1) = \frac{\beta\kappa(1+\max\{d,\zeta\})}{\beta\kappa+\max\{d,\zeta\}} = \frac{\rho(q+1)}{\frac{\rho(q+1)+\max\{d,\zeta\}(\rho(q+2)+d)}{1+\max\{d,\zeta\}}} = \frac{\rho(q+1)}{\rho(q+2)+d-\frac{\rho+d}{1+\max\{d,\zeta\}}}$$

such that we have at most an improvement of $\frac{\rho+d}{1+\max\{d,\zeta\}}$ in the denominator.◀

In the comparison above the localized SVM rate outperforms the optimal rate by making the additional assumption (iii). This is not surprising, since the assumptions we made imply the assumptions of (Audibert and Tsybakov, 2007). We emphasize once again that our rates as well as the other rates are achieved under less assumptions.

Tree-based and Plug-in classifier. Assume that (ii) and (iv) are satisfied. Then, the classifiers resulting from the tree-based adaptive partitioning methods in (Binev et al., 2014, Sec. 6) yield under assumptions (ii) and (iv) the rate

$$n^{-\frac{\rho(q+1)}{\rho(q+2)+d}}$$

(see Binev et al., 2014, Theorems 6.1(i) and 6.3(i)). In fact the rate is achieved under milder assumptions, namely (ii) and some condition on the behavior of the approximation error w.r.t. P, however, by (Binev et al., 2014, Prop. 4.1) the latter is immediately satisfied under (ii) and (iv). Moreover, (Kohler and Krzyzak, 2007, Theorems 1, 3, and 5) showed that plug-in-classifiers based on kernel, partitioning and nearest neighbor regression estimates learn with rate

$$n^{-\frac{\rho(q+1)}{\rho(q+3)+d}}. \tag{25}$$

Actually, this rate holds under a slightly weaker assumption than (ii), namely that there exists a $\bar{c} > 0$ and some $\alpha > 0$ such that for all $\delta > 0$ the inequality

$$\mathbb{E}(|\eta - 1/2| \cdot \mathbf{1}_{\{|\eta-1/2|\leq\delta\}}) \leq \bar{c} \cdot \delta^{1+\alpha}$$

is satisfied, but this is implied by (ii) (see Döring et al., 2015, Sec. 5). To compare our rates we add (iii) to (ii) and (iv). Then, the localized SVM rate again equals (22) and is faster for all ν our analysis is applied to. The improvement to the rate from (Binev et al., 2014) is equal to the improvement in the previous comparison, whereas compared to the rate from (Kohler and Krzyzak, 2007) the improvement is at least better by ρ in the denominator.◀

The three comparisons above have in common that rates are solely improved by assumption (iii). This condition was even sufficient enough to improve the optimal rate in (24). It is to emphasize that neither for the rates from Theorem 2 or the rates from the mentioned authors above nor in our comparisons assumptions on the existence of a density of the marginal distribution P_X have to be made. As mentioned in the introduction assumptions without conditions on the density of distributions are preferable, however, to compare our rates we find subsequently assumption sets that do contain those.

Plug-in classifier II. Let us assume that (ii) and (iv) are satisfied and that P_X has a uniformly bounded density w.r.t. the Lebesgue measure. Then, Audibert and Tsybakov (2007, Theorem 4.1) show that plug-in classifiers learns with the optimal rate

$$n^{-\frac{\rho(q+1)}{\rho(q+2)+d}}.$$

If we assume in addition (iii), the localized SVM rate again equals (22) and satisfies (23) such that our rate is faster for arbitrary $\nu \leq \frac{1-\beta\kappa}{\beta\kappa+\max\{d,\zeta\}}$. ◀

Before we proceed, we define another margin condition that measures the amount of mass close to the decision boundary and we say according to (Steinwart and Christmann, 2008, Definition 8.6) that P has margin exponent (ME) $\alpha \in (0,\infty]$, if there exists a constant $c_{\mathrm{ME}} > 0$ such that

$$P_X(\{\Delta_\eta(x) < t\}) \leq (c_{\mathrm{ME}} t)^\alpha \tag{26}$$

for all $t > 0$. Descriptively, large values of α reflect a low concentration of mass in the vicinity of the decision boundary.

Plug-in classifier III. Let us assume that (ii), (iv) are satisfied and that that P_X has a density with respect to the Lebesgue measure that is bounded away from zero. Then, the authors in Belkin et al. (2018) show that plug-in classifiers based on a weighted and interpolated nearest neighbor scheme obtain the rate

$$n^{-\frac{\rho q}{p(q+2)+d}}. \tag{27}$$

Under the same conditions, Kohler and Krzyzak (2007) improved for plug-in-classifier based on kernel, partitioning, and nearest neighbor regression estimates the rate in (25) to

$$n^{-\frac{\rho(q+1)}{2\rho+d}}. \tag{28}$$

To compare results we add (iii) to (ii) and (iv). Then, the localized SVM rate equals (22) and satisfies (23) such that our rate is obviously faster than the rate in (27) for all possible choices of ν. The improvement compared to (27) is at least $\frac{\rho}{\rho(q+2)+d}$. In order to compare our rate with (28) we take a closer look on the rate and the margin parameters under the stated conditions. A short calculation shows for the exponent of the rate in (22) that

$$\beta\kappa(\nu+1) = \frac{\rho(q+1)}{\frac{\rho(q+2)+d}{(\nu+1)}} = \frac{\rho(q+1)}{2\rho+d+\frac{\rho(q+2)-2\rho(\nu+1)+d-d(\nu+1)}{(\nu+1)}}$$

and its easy to derive that our exponent is only larger than the one in (28) or equals it if $\nu \geq \frac{pq}{2\rho+d}$. We show that the largest ν we can choose satisfies this bound if $\rho = 1$ and derive

a rate for this case. Since P_X has a density with respect to the Lebesgue measure that is bounded away from zero, we restrict ourselves to the case that $\rho q \leq 1$ and hence $q \leq 1$, see Remark 25. Moreover, Blaschzyk and Steinwart (2018, Lemma A.2) and Steinwart and Christmann (2008, Lemma 8.23) yield $\alpha = \rho q = q$. Furthermore, we find by Lemma 24 that $q = \frac{\alpha}{\zeta}$ and we follow $\zeta = \rho = 1$. Thus, a short calculation shows that

$$\nu := \tfrac{1-\beta\kappa}{\beta\kappa+\max\{d,\zeta\}} = \tfrac{1-\beta\kappa}{\beta\kappa+d} = \tfrac{\rho+d}{\rho(q+1)+d(\rho(q+2)+d)} = \tfrac{1+d}{q+1+d(q+2+d)} \geq \tfrac{q}{2+d}$$

is satisfied for all $q \leq 1$. By inserting this ν into the exponent of the localized SVM rate in (22) we find

$$\beta\kappa(\nu+1) = \tfrac{\beta\kappa(1+d)}{\beta\kappa+d} = \tfrac{\rho(q+1)}{\rho(q+2)+d+\frac{\rho(q+1)-(\rho(q+2)+d)}{1+d}} = \tfrac{q+1}{q+2+d+\frac{q+1-(q+2+d)}{1+d}} = \tfrac{q+1}{q+1+d}.$$

Hence, the localized SVM rate is faster than the rate in (28) for all $q < 1$ and matches it if $q = 1$. ◀

Under assumptions that contained that P_X has a density w.r.t. Lebesgue measure that is bounded away from zero, we improved in the previous comparison the rates from (Belkin et al., 2018) and in the case that η is Lipschitz, the rates from (Kohler and Krzyzak, 2007). We remark that under a slight stronger density assumption Audibert and Tsybakov (2007) showed that certain plug-in classifier achieve the optimal rate in (28).

Finally, we compare our rates to the ones derived for the histogram rule in (Blaschzyk and Steinwart, 2018), where we also considered a set of margin conditions and a similar strategy to derive their rates. Note that under a certain assumption set the authors showed that the histogram rule outperformed the global SVM rates from (Steinwart and Christmann, 2008, (8.18)) and the localized SVM rates from (Thomann et al., 2017).

Histogram rule. Let us assume that (i) and (iii) are satisfied and that

(v) P has ME $\alpha \in (0, \infty]$,

see (26). Then, we find by Lemma 24 that we have NE $q = \frac{\alpha}{\zeta}$ and according to (Blaschzyk and Steinwart, 2018, Theorem 3.5) the histogram rule then learns with rate

$$n^{-\frac{\beta(q+1)}{\beta(q+1)+d(q+1)+\frac{\beta\zeta}{1+\zeta}}} \tag{29}$$

as long as $\beta \leq (1+\zeta)(q+1)$. Under these assumptions the localized SVM learns with the rate from Theorem 2 that is

$$n^{-\beta\kappa(\nu+1)},$$

where our rate depends on ν. To compare our rates we have to pay attention to the range of β that provides a suitable ν, see (11). If we have that $(q+1)(1+\max\{d,\zeta\}-d) \leq \beta \leq (q+1)(1+\zeta)$, then a short calculation shows that our local SVM rate in (13) is faster if ν is not too small, that is if ν satisfies

$$\beta\left((\beta+d)(q+1)(\zeta+1)+\beta\zeta\right)^{-1} \leq \nu \leq \tfrac{\kappa}{1-\kappa}$$

According to Remark 3 the best possible rate is achieved for $\nu := \frac{\kappa}{1-\kappa}$ and has then exponent

$$\frac{\beta(q+1)}{\beta(q+2)+(d-1)(q+1)} = \frac{\beta(q+1)}{\beta(q+1)+d(q+1)+\frac{\beta\zeta}{(1+\zeta)}+\beta-(q+1)-\frac{\beta\zeta}{1+\zeta}} = \frac{\beta(q+1)}{\beta(q+1)+d(q+1)-\left(q+\frac{\zeta}{1+\zeta}\right)}$$

such that compared to (29) we have an improvement of $q+\frac{\zeta}{1+\zeta}$ in the denominator. In the other case, that is, if $\beta \leq (q+1)(1+\max\{d,\zeta\}-d)$ a short calculation shows that our local SVM rate is better for all choices

$$\beta\left((\beta+d)(q+1)(\zeta+1)+\beta\zeta\right)^{-1} \leq \nu \leq \frac{1-\beta\kappa}{\beta k+\max\{d,\zeta\}},$$

In this case we find due to Remark 3 that the best possible rate is achieved for $\nu := \frac{1-\beta\kappa}{\beta k+\max\{d,\zeta\}}$ and has exponent

$$\frac{\beta\kappa(1+\max\{d,\zeta\})}{\beta k+\max\{d,\zeta\}} = \frac{\beta(q+1)}{\beta(q+1)+d(q+1)+\frac{\beta\max\{d,\zeta\}-d(q+1)}{1+\max\{d,\zeta\}}} = \frac{\beta(q+1)}{\beta(q+1)+d(q+1)+\frac{\beta\zeta}{1+\zeta}-\left(\frac{d(q+1)-\beta\max\{d,\zeta\}}{1+\max\{d,\zeta\}}+\frac{\beta\zeta}{1+\zeta}\right)}.$$

Compared to (29) the rate is better by $\frac{d(q+1)-\beta\max\{d,\zeta\}}{1+\max\{d,\zeta\}}+\frac{\beta\zeta}{1+\zeta} > 0$ in the denominator. We remark that the lower bound on ν is not surprising since if $\nu \to 0$ our rate matches the global rate in (19) and Blaschzyk and Steinwart (2018) showed that under a certain assumption set the rate of the histogram classifier is faster than the one of the global SVM. Moreover, we remark that our rates in Theorem 2 hold for all values of β and not only for a certain range of β. ◀

Acknowledgments

Reprinted/adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Spectrum Improved Classification Rates for Localized Algorithms under Margin Conditions by Ingrid Blaschzyk © (2020)

4. Proofs

In this section we state the proofs of Section 3. We define $\gamma_{\max} := \max_{j\in J}\gamma_j$ and $\gamma_{\min} := \min_{j\in J}\gamma_j$ w.r.t. some $J \subset \{1,\ldots,m\}$.

4.1 Proof of Theorem 2

Proof [Proof of Theorem 2] By Theorem 1 for $s_n := n^{-\nu}$ we find that

$$\begin{aligned}
&\mathcal{R}_{L_J,P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_J,P} \\
&\leq \mathcal{R}_{L_N,P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_N,P} + \mathcal{R}_{L_F,P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_F,P} \\
&\leq \mathcal{R}_{L_{N_1},P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_{N_1},P} + \mathcal{R}_{L_{N_2},P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_{N_2},P} + \mathcal{R}_{L_F,P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_F,P}.
\end{aligned} \tag{30}$$

In the subsequent steps we bound the excess risks above separately for both choices of ν by applying Theorems 18, 20 and 22 for $\alpha := \nu$. First, we consider the case $\beta \geq$

$(q+1)(1+\max\{d,\zeta\}-d)$ and check some requirements for the mentioned theorems. Since $\beta \geq \frac{q+1}{q+2}(1+\max\{d,\zeta\}-d)$ we have

$$\nu \leq \tfrac{\kappa}{1-\kappa} = \tfrac{q+1}{\beta(q+2)+(d-1)(q+1)} \leq \tfrac{1}{\max\{d,\zeta\}}.$$

Moreover,

$$1+\nu(1-d) \geq 1 - \tfrac{\kappa(d-1)}{1-\kappa} = 1 - \tfrac{(q+1)(d-1)}{\beta(q+2)+(d-1)(q+1)} = \tfrac{\beta(q+2)}{\beta(q+2)+(d-1)(q+1)} > 0.$$

Hence, we apply Theorem 18 and Theorems 20, 22 with $\alpha := \nu$. That means, together with

$$\beta\kappa(\nu+1) \leq \tfrac{\beta\kappa}{1-\kappa} = \tfrac{(q+1)\beta(q+2)}{(q+2)(\beta(q+2)+(d-1)(q+1))} = \tfrac{q+1}{q+2}\left(1 - \tfrac{\kappa(d-1)}{1-\kappa}\right) \leq \tfrac{(q+1)(1-\nu(d-1))}{q+2} \tag{31}$$

and

$$\beta\kappa(\nu+1) \leq \tfrac{\beta(q+1)}{\beta(q+2)+(d-1)(q+1)} = 1 - \tfrac{(d-1)(q+1)+\beta}{\beta(q+2)+(d-1)(q+1)} \leq 1 - \tfrac{(q+1)\max\{d,\zeta\}}{\beta(q+2)+(d-1)(q+1)} \leq 1 - \tfrac{\kappa\max\{d,\zeta\}}{1-\kappa} \tag{32}$$

so that $\beta\kappa(\nu+1) \leq 1-\nu\max\{d,\zeta\}$ for $\nu \leq \frac{\kappa}{1-\kappa}$, we obtain in (30) for $\varepsilon_1,\varepsilon_2,\varepsilon_3 > 0$ and with probability P^n not less than $1-9e^{-\tau}$ that

$$\begin{aligned}
&\mathcal{R}_{L_J,P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_J,P} \\
&\leq \mathcal{R}_{L_{N_1},P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_{N_1},P} + \mathcal{R}_{L_{N_2},P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_{N_2},P} + \mathcal{R}_{L_F,P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_F,P} \\
&\leq c_1\tau\left(n^{-\beta\kappa(\nu+1)}n^{\varepsilon_1} + n^{-\frac{(q+1)(1+\alpha-\nu d)}{q+2}}n^{\varepsilon_2} + n^{-(1-\max\{\nu d,\alpha\zeta\})}n^{\varepsilon_3}\right) \\
&\leq c_2\tau n^{\varepsilon}\left(2n^{-\beta\kappa(\nu+1)} + n^{-(1-\nu\max\{d,\zeta\})}\right) \\
&\leq c_3\tau n^{-\beta\kappa(\nu+1)+\varepsilon},
\end{aligned} \tag{33}$$

holds for some $\varepsilon := \max\{\varepsilon_1,\varepsilon_2,\varepsilon_3\}$ and some constants c_1 depending on $d,\beta,q,\xi,\varepsilon_1,\varepsilon_2,\varepsilon_3,\sigma$, and $c_2,c_3 > 0$ depending on $d,\beta,q,\xi,\varepsilon,\sigma$.

Second, we consider the case $\beta < (q+1)(1+\max\{d,\zeta\}-d)$ and check again the requirements on $\nu \leq \frac{1-\beta\kappa}{\beta k+\max\{d,\zeta\}}$ for the theorems applied above. We have

$$\nu \leq \tfrac{1-\beta\kappa}{\beta\kappa+\max\{d,\zeta\}} = \tfrac{\beta+d(q+1)}{\beta(q+1)+\max\{d,\zeta\}(\beta(q+2)+d(q+1))} \leq \tfrac{q+1}{\beta(q+2)+(d-1)(q+1)} = \tfrac{\kappa}{1-\kappa}, \tag{34}$$

and

$$\nu \leq \tfrac{1-\beta\kappa}{\beta\kappa+\max\{d,\zeta\}} \leq \tfrac{1-\beta\kappa}{\max\{d,\zeta\}} \leq \tfrac{1}{\max\{d,\zeta\}}.$$

Moreover,

$$1+\nu(1-d) \geq 1 - \tfrac{(1-\beta\kappa)(d-1)}{\beta\kappa+\max\{d,\zeta\}} = \tfrac{\max\{d,\zeta\}-d(1-\beta\kappa)+1}{\beta\kappa+\max\{d,\zeta\}} \geq \tfrac{\max\{d,\zeta\}-d+1}{\beta\kappa+\max\{d,\zeta\}} > 0.$$

Again, we apply Theorem 17 and Theorems 19, 21 for $\alpha := \nu$. Together with (34) we find similar to (31) and (32) that

$$\beta\kappa(\nu+1) \leq \tfrac{\beta\kappa}{1-\kappa} = \tfrac{q+1}{q+2}\left[1 - \tfrac{\kappa(d-1)}{1-\kappa}\right] \leq \tfrac{q+1}{q+2}\left[1 - \tfrac{(1-\beta\kappa)(d-1)}{\beta k+\max\{d,\zeta\}}\right] \leq \tfrac{(q+1)(1-\nu(d-1))}{q+2},$$

and

$$\beta\kappa(\nu+1) \leq \tfrac{\beta\kappa(1+\max\{d,\zeta\})}{\beta\kappa+\max\{d,\zeta\}} = 1 - \tfrac{\max\{d,\zeta\}(1-\beta\kappa)}{\beta\kappa+\max\{d,\zeta\}} \leq 1 - \nu\max\{d,\zeta\}$$

such that we obtain in (30) that

$$\begin{aligned}
&\mathcal{R}_{L_J,P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_J,P} \\
&\leq \mathcal{R}_{L_{N_1},P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_{N_1},P} + \mathcal{R}_{L_{N_2},P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_{N_2},P} + \mathcal{R}_{L_F,P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_F,P} \\
&\leq c_1\tau\left(n^{\varepsilon_1}n^{-\beta\kappa(\nu+1)} + n^{\varepsilon_2}n^{-\frac{(q+1)(1+\alpha-\nu d)}{q+2}} + n^{\varepsilon_3}n^{-(1-\nu\max\{d,\zeta\})}\right) \\
&\leq c_2\tau n^{\varepsilon}\left(2n^{-\beta\kappa(\nu+1)} + n^{-(1-\nu\max\{d,\zeta\})}\right) \\
&\leq c_3\tau n^{-\beta\kappa(\nu+1)+\varepsilon},
\end{aligned}$$

holds with probability P^n not less than $1 - 9e^{-\tau}$. ■

4.2 Proof of Lemma 5

Proof [Proof of Lemma 5] Since $\widehat{f} : X \to [-1,1]$ we consider functions $f : X \to [-1,1]$. Then, an analogous calculation as in the proof of (Steinwart and Christmann, 2008, Theorem 8.24) yields $(L_F \circ f - L_F \circ f^*_{L,P})^2 = f - f^*_{L,P}$. Following the same arguments as in (Blaschzyk and Steinwart, 2018, Lemma 3.4) we find for all $x \in F$ with the lower-control assumption that

$$1 \leq \frac{c_{\mathrm{LC}}}{s^\zeta}|2\eta(x) - 1|.$$

Then, we have

$$\begin{aligned}
\mathbb{E}_P(L_F \circ f - L_F \circ f^*_{L,P})^2 &= \int_F |f(x) - f^*_{L,P}(x)|^2 dP_X(x) \\
&\leq 2\int_F |f(x) - f^*_{L,P}(x)| dP_X(x) \\
&\leq \frac{2c_{\mathrm{LC}}}{s^\zeta}\int_F |f(x) - f^*_{L,P}(x)||2\eta(x) - 1| dP_X(x) \\
&\leq \frac{2c_{\mathrm{LC}}}{s^\zeta}\mathbb{E}_P(L_F \circ f - L_F \circ f^*_{L,P}).
\end{aligned}$$

■

4.3 Proof of Theorem 7

Proof [Proof of Theorem 7] We assume $n \geq n^* := \max\{4, \left(4^{-1}\delta^*\right)^{-\frac{1}{\nu}}, \left(4^{-1}\delta^*\right)^{-\frac{1}{\alpha}}\}$ without loss of generality. For $n < n^*$ the equation in (18) is immediately satisfied with constant $\tilde{c}_{d,\beta,q,\varepsilon} := (n^*)^{\beta\kappa(\nu+1)-\varepsilon}$ and with probability 1. We analyze the excess risk

$\mathcal{R}_{L_J,P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L,P}$ by applying the splitting technique described in Section 3.1 and by applying a generic oracle inequality for ERM given in (Steinwart and Christmann, 2008, Theorem 7.2) on each set. To this end, we define $s_n := r_n$ and find by Theorem 1 that

$$\begin{aligned}
&\mathcal{R}_{L_J,P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_J,P} \\
&\leq \mathcal{R}_{L_N,P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_N,P} + \mathcal{R}_{L_F,P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_F,P} \\
&\leq \mathcal{R}_{L_{N_1},P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_{N_1},P} + \mathcal{R}_{L_{N_2},P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_{N_2},P} \\
&\qquad + \mathcal{R}_{L_F,P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_F,P}.
\end{aligned} \tag{35}$$

First of all, we analyze $\mathcal{R}_{L_{N_1},P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_{N_1},P}$. Note that

$$\mathcal{R}_{L_{N_1},P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) = \sum_{j\in J_{N_1}} \mathcal{R}_{L_j,P}(f_{D_1,\lambda_{D_2,j},\gamma_{D_2,j}}).$$

According to (Steinwart and Christmann, 2008, Theorem 8.24) we have on the set N_1 variance bound $\theta = \frac{q}{q+1}$ with constant $V := 6c_{\mathrm{NE}}^{\frac{q}{q+1}}$. Then, for fixed data set D_1 and

$$\tau_n^{N_1} := \tau + \ln(1 + |\Lambda_n \times \Gamma_n|^{|J_{N_1}|}),$$

as well as $n - l \geq n/4$ for $n \geq 4$, and $l := \lfloor \frac{n}{2} \rfloor + 1$, we find by (Steinwart and Christmann, 2008, Theorem 7.2) with probability P^{n-l} not less than $1 - e^{-\tau}$ that

$$\begin{aligned}
&\mathcal{R}_{L_{N_1},P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_{N_1},P} \\
&\leq 6\left(\inf_{(\boldsymbol{\lambda},\boldsymbol{\gamma})\in(\Lambda_n\times\Gamma_n)^{|J_{N_1}|}} \mathcal{R}_{L_{N_1},P}(f_{D_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_{N_1},P} \right) \\
&\qquad + 4\left(\frac{48 c_{\mathrm{NE}}^{\frac{q}{q+1}} (\tau + \ln(1 + |\Lambda_n \times \Gamma_n|^{|J_{N_1}|})}{n-l} \right)^{\frac{q+1}{q+2}} \\
&\leq 6\left(\inf_{(\boldsymbol{\lambda},\boldsymbol{\gamma})\in(\Lambda_n\times\Gamma_n)^{|J_{N_1}|}} \mathcal{R}_{L_{N_1},P}(f_{D_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_{N_1},P} \right) + c_q \left(\frac{\tau_n^{N_1}}{n} \right)^{\frac{q+1}{q+2}}.
\end{aligned} \tag{36}$$

By Theorem 17 for $p \in (0, \frac{1}{2})$ we obtain with probability P^l not less than $1 - 3|\Lambda_n \times \Gamma_n|^{|J_{N_1}|} e^{-\tau}$ that

$$\begin{aligned}
&\mathcal{R}_{L_{N_1},P}(f_{D_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_{N_1},P} \\
&\leq c_1 \Bigg(\sum_{j\in J_{N_1}} \frac{\lambda_j r^d}{\gamma_j^d} + \max_{j\in J_{N_1}} \gamma_j^\beta + \left(\frac{r_n}{n}\right)^{\frac{q+1}{q+2-p}} \left(\sum_{j\in J_{N_1}} \lambda_j^{-1} \gamma_j^{-\frac{d}{p}} P_X(A_j) \right)^{\frac{p(q+1)}{q+2-p}} \\
&\qquad + \left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}} \Bigg)
\end{aligned}$$

holds for all $(\boldsymbol{\lambda},\boldsymbol{\gamma}) \in \Lambda_n^{|J_{N_1}|} \times \Gamma_n^{|J_{N_1}|}$ simultaneously and some constant $c_1 > 0$ depending on d, β, p, q. Then, Lemma 28 i) for all $\varepsilon_1 > 0$ yields

$$
\begin{aligned}
&\inf_{(\boldsymbol{\lambda},\boldsymbol{\gamma})\in(\Lambda_n\times\Gamma_n)^{|J_{N_1}|}} \mathcal{R}_{L_{N_1},P}(f_{D_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_{N_1},P} \qquad (37)\\
&\leq \inf_{(\boldsymbol{\lambda},\boldsymbol{\gamma})\in(\Lambda_n\times\Gamma_n)^{|J_{N_1}|}} c_1 \Bigg(\sum_{j\in J_{N_1}} \frac{\lambda_j r^d}{\gamma_j^d} + \max_{j\in J_{N_1}} \gamma_j^\beta \\
&\qquad + \left(\left(\frac{r_n}{n}\right)^{\frac{1}{p}} \sum_{j\in J_{N_1}} \lambda_j^{-1} \gamma_j^{-\frac{d}{p}} P_X(A_j) \right)^{\frac{p(q+1)}{q+2-p}} + \left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}} \Bigg) \\
&\leq c_2 \left(n^{-\beta\kappa(\nu+1)+\varepsilon_1} + \tau^{\frac{q+1}{q+2}} n^{-\frac{q+1}{q+2}} \right)
\end{aligned}
$$

where $c_2 > 0$ is a constant depending on d, β, q and ε_1. By inserting (37) into (36) we have with probability P^n not less than $1 - (1 + 3|\Lambda_n \times \Gamma_n|^{|J_{N_1}|})e^{-\tau}$ that

$$
\begin{aligned}
&\mathcal{R}_{L_{N_1},P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_{N_1},P} \\
&\leq 6 \left(\inf_{(\boldsymbol{\lambda},\boldsymbol{\gamma})\in(\Lambda_n\times\Gamma_n)^{|J_{N_1}|}} \mathcal{R}_{L_{N_1},P}(f_{D_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_{N_1},P} \right) + c_q \left(\frac{\tau_n^{N_1}}{n} \right)^{\frac{q+1}{q+2}} \\
&\leq c_2 \left(n^{-\beta\kappa(\nu+1)+\varepsilon_1} + \tau^{\frac{q+1}{q+2}} n^{-\frac{q+1}{q+2}} \right) + c_q \left(\frac{\tau_n^{N_1}}{n} \right)^{\frac{q+1}{q+2}} .
\end{aligned}
$$

Analogously to the calculations at the beginning of the poof of Theorem 17, we find by Lemma 26 for $t = 2r_n$ that $|J_{N_1}| \leq c_d r_n r_n^{-d}$. Thus, we obtain with the inequalities (31)

resp. (34) that

$$
\begin{aligned}
&\mathcal{R}_{L_{N_1},P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_{N_1},P} \qquad (38)\\
&\leq c_2 \left(n^{-\beta\kappa(\nu+1)+\varepsilon_1} + \tau^{\frac{q+1}{q+2}} n^{-\frac{q+1}{q+2}} \right) + c_q \left(\frac{\tau + \ln(1+|\Lambda_n \times \Gamma_n|^{|J_{N_1}|})}{n} \right)^{\frac{q+1}{q+2}} \\
&\leq c_3 \Bigg(n^{-\beta\kappa(\nu+1)+\varepsilon_1} + \tau^{\frac{q+1}{q+2}} n^{-\frac{q+1}{q+2}} + \left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}} \\
&\qquad + \left(\frac{\ln(1+|\Lambda_n \times \Gamma_n|^{|J_{N_1}|})}{n} \right)^{\frac{q+1}{q+2}} \Bigg) \\
&\leq c_3 \left(n^{-\beta\kappa(\nu+1)+\varepsilon_1} + 2\tau^{\frac{q+1}{q+2}} n^{-\frac{q+1}{q+2}} + \left(\frac{|J_{N_1}| \ln(2|\Lambda_n \times \Gamma_n|)}{n} \right)^{\frac{q+1}{q+2}} \right) \\
&\leq c_3 \left(n^{-\beta\kappa(\nu+1)+\varepsilon_1} + 2\tau^{\frac{q+1}{q+2}} n^{-\frac{q+1}{q+2}} + \left(\frac{c_d r_n \ln(2|\Lambda_n \times \Gamma_n|)}{r_n^d n} \right)^{\frac{q+1}{q+2}} \right) \\
&= c_3 \left(n^{-\beta\kappa(\nu+1)+\varepsilon_1} + 2\tau^{\frac{q+1}{q+2}} n^{-\frac{q+1}{q+2}} + \left(\frac{c_d \ln(2|\Lambda_n \times \Gamma_n|)}{n^{1-\nu(d-1)}} \right)^{\frac{q+1}{q+2}} \right) \\
&\leq c_4 \left(2n^{-\beta\kappa(\nu+1)+\hat{\varepsilon}_1} + 2\tau^{\frac{q+1}{q+2}} n^{-\frac{q+1}{q+2}} \right),
\end{aligned}
$$

where $\hat{\varepsilon}_1 > 0$ and where $c_3, c_4 > 0$ are constants depending on $d, \beta, q, \varepsilon_1$ resp. $d, \beta, q, \hat{\varepsilon}_1$. A variable transformation in τ together with Lemma 26 yields

$$
\mathcal{R}_{L_{N_1},P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_{N_1},P} \leq c_5 \tau^{\frac{q+1}{q+2}} \cdot n^{-\beta\kappa(\nu+1)+\tilde{\varepsilon}_1} \qquad (39)
$$

for some $\tilde{\varepsilon}_1 > 0$ with probability P^n not less than $1 - e^{-\tau}$, where $c_5 > 0$ is a constant depending on $d, \beta, q, \tilde{\varepsilon}_1$.

Next, we analyze $\mathcal{R}_{L_{N_2},P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_{N_2},P}$ by the same procedure. According to (Steinwart and Christmann, 2008, Theorem 8.24) we have on the set N_2 variance bound $\theta = \frac{q}{q+1}$ with constant $V := 6c_{\mathrm{NE}}^{\frac{q}{q+1}}$. Then, for fixed data set D_1 and

$$
\tau_n^{N_2} := \tau + \ln(1 + |\Lambda_n \times \Gamma_n|^{|J_{N_2}|}),
$$

as well as $n - l \geq n/4$ for $n \geq 4$, and $l := \lfloor \frac{n}{2} \rfloor + 1$, we find by (Steinwart and Christmann, 2008, Theorem 7.2) with probability P^{n-l} not less than $1 - e^{-\tau}$ that

$$
\begin{aligned}
&\mathcal{R}_{L_{N_2},P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_{N_2},P} \qquad (40)\\
&\leq 6 \left(\inf_{(\boldsymbol{\lambda},\boldsymbol{\gamma}) \in (\Lambda_n \times \Gamma_n)^{|J_{N_2}|}} \mathcal{R}_{L_{N_2},P}(f_{D_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_{N_2},P} \right) + c_q \left(\frac{\tau_n^{N_2}}{n} \right)^{\frac{q+1}{q+2}}.
\end{aligned}
$$

By Theorem 19 for $s_n = r_n$, $p \in (0, \frac{1}{2})$ and $\hat{\varepsilon} > 0$ we obtain with probability P^n not less than $1 - (1 + 3|\Lambda_n \times \Gamma_n|^{|J_{N_2}|})e^{-\tau}$ that

$$\begin{aligned}
&\mathcal{R}_{L_{N_2},P}(f_{D_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_{N_2},P} \\
&\leq c_6 \Bigg(\left(\frac{r_n}{\min_{j \in J_{N_2}} \gamma_j} \right)^d \sum_{j \in J_{N_2}} \lambda_j n^{\hat{\varepsilon}} \\
&\qquad + \left(\frac{r_n}{n} \right)^{\frac{q+1}{q+2-p}} \left(\sum_{j \in J_{N_2}} \lambda_j^{-1} \gamma_j^{-\frac{d}{p}} P_X(A_j) \right)^{\frac{p(q+1)}{q+2-p}} + \left(\frac{\tau}{n} \right)^{\frac{q+1}{q+2}} \Bigg)
\end{aligned}$$

holds for all $(\boldsymbol{\lambda}, \boldsymbol{\gamma}) \in \Lambda_n^{|J_{N_2}|} \times \Gamma_n^{|J_{N_2}|}$ simultaneously and some constant $c_6 > 0$ depending on d, β, p, q and $\hat{\varepsilon}$. Then, Lemma 28 ii) for all $\varepsilon_2 > 0$ yields

$$\begin{aligned}
&\inf_{(\boldsymbol{\lambda},\boldsymbol{\gamma}) \in (\Lambda_n \times \Gamma_n)^{|J_{N_2}|}} \mathcal{R}_{L_{N_2},P}(f_{D_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_{N_2},P} \qquad (41) \\
&\leq \inf_{(\boldsymbol{\lambda},\boldsymbol{\gamma}) \in (\Lambda_n \times \Gamma_n)^{|J_{N_2}|}} c_6 \Bigg(\left(\frac{r_n}{\min_{j \in J_{N_2}} \gamma_j} \right)^d \sum_{j \in J_{N_2}} \lambda_j n^{\tilde{\varepsilon}} \\
&\qquad + \left(\left(\frac{r_n}{n} \right)^{\frac{1}{p}} \sum_{j \in J_{N_2}} \lambda_j^{-1} \gamma_j^{-\frac{d}{p}} P_X(A_j) \right)^{\frac{p(q+1)}{q+2-p}} + \left(\frac{\tau}{n} \right)^{\frac{q+1}{q+2}} \Bigg) \\
&\leq c_7 \left(n^{\varepsilon_2} \left(r_n^{d-1} n \right)^{-\frac{q+1}{q+2}} + \tau^{\frac{q+1}{q+2}} \cdot n^{-\frac{q+1}{q+2}} \right),
\end{aligned}$$

where $c_7 > 0$ is a constant depending on d, β, q and ε_2. We insert (41) into (40) and obtain

$$\begin{aligned}
&\mathcal{R}_{L_{N_2},P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_{N_2},P} \\
&\leq c_7 \left(n^{\varepsilon_2} \left(r_n^{d-1} n \right)^{-\frac{q+1}{q+2}} + \tau^{\frac{q+1}{q+2}} \cdot n^{-\frac{q+1}{q+2}} \right) + c_q \left(\frac{\tau_n^{N_2}}{n} \right)^{\frac{q+1}{q+2}}.
\end{aligned}$$

with probability P^n not less than $1 - (1 + 3|\Lambda_n \times \Gamma_n|^{|J_{N_2}|})e^{-\tau}$. An analogous calculation as in (38) yields

$$\begin{aligned}
&\mathcal{R}_{L_{N_2},P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_{N_2},P} \\
&\leq c_8 \left(n^{\varepsilon_2} \left(r_n^{d-1} n \right)^{-\frac{q+1}{q+2}} + 2\tau^{\frac{q+1}{q+2}} \cdot n^{-\frac{q+1}{q+2}} + \left(\frac{c_d \ln(2|\Lambda_n \times \Gamma_n|)}{n^{1-\nu(d-1)}} \right)^{\frac{q+1}{q+2}} \right) \\
&\leq c_8 \left(n^{\varepsilon_2} \left(n^{1-\nu(d-1)} \right)^{-\frac{q+1}{q+2}} + 2\tau^{\frac{q+1}{q+2}} \cdot n^{-\frac{q+1}{q+2}} + \left(\frac{c_d \ln(2|\Lambda_n \times \Gamma_n|)}{n^{1-\nu(d-1)}} \right)^{\frac{q+1}{q+2}} \right) \\
&\leq c_9 \left(n^{\hat{\varepsilon}_2} \left(n^{1-\nu(d-1)} \right)^{-\frac{q+1}{q+2}} + \tau^{\frac{q+1}{q+2}} \cdot n^{-\frac{q+1}{q+2}} \right),
\end{aligned}$$

where $\hat{\varepsilon}_2 > 0$ and where $c_8, c_9 > 0$ are constants depending on $d, \beta, q, \varepsilon_2$ resp. $d, \beta, q, \hat{\varepsilon}_2$. A variable transformation in τ together with Lemma 26 then yields

$$\mathcal{R}_{L_{N_2},P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_{N_2},P} \leq c_{10}\tau^{\frac{q+1}{q+2}} \cdot n^{\tilde{\varepsilon}_2} \left(n^{1-\nu(d-1)}\right)^{-\frac{q+1}{q+2}} \tag{42}$$

with probability P^n not less than $1-e^{-\tau}$, where $c_{10} > 0$ is a constant depending on $d, \beta, q, \tilde{\varepsilon}_2$.

Next, we analyze $\mathcal{R}_{L_F,P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_F,P}$. According to Theorem 5 we have on the set F the best possible variance bound $\theta = 1$ with constant $V := 2c_{\mathrm{LC}}r_n^{-\zeta}$. Then, for fixed data set D_1 and

$$\tau_n^F := \tau + \ln(1 + |\Lambda_n \times \Gamma_n|^{|J_F|}),$$

as well as $n - l \geq n/4$ for $n \geq 4$, and $l := \lfloor \frac{n}{2} \rfloor + 1$, we find by (Steinwart and Christmann, 2008, Theorem 7.2) with probability P^{n-l} not less than $1 - e^{-\tau}$ that

$$\begin{aligned} &\mathcal{R}_{L_F,P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_F,P} \\ &\leq 6\left(\inf_{(\boldsymbol{\lambda},\boldsymbol{\gamma})\in(\Lambda_n\times\Gamma_n)^{|J_F|}} \mathcal{R}_{L_F,P}(f_{D_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_F,P}\right) + \frac{2c_{\mathrm{LC}}\tau_n^F}{r_n^\zeta n}. \end{aligned} \tag{43}$$

Then, Theorem 21 for $s_n = r_n$, $p \in (0, \frac{1}{2})$ and $\tilde{\varepsilon} > 0$ yields

$$\begin{aligned} &\mathcal{R}_{L_F,P}(f_{D_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_F,P} \\ &\leq c_{10}\left(\left(\frac{r_n}{\min_{j\in J_F}\gamma_j}\right)^d \sum_{j\in J_F} \lambda_j n^{\tilde{\varepsilon}} + \left(\sum_{j\in J_F} \lambda_j^{-1}\gamma_j^{-\frac{d}{p}} P_X(A_j)\right)^p n^{-1} + \frac{\tau}{r^\zeta n}\right), \end{aligned}$$

with probability P^l not less than $1 - 3|\Lambda_n \times \Gamma_n|^{|J_F|}e^{-\tau}$ and for all $(\boldsymbol{\lambda},\boldsymbol{\gamma}) \in \Lambda_n^{|J_F|} \times \Gamma_n^{|J_F|}$ simultaneously and some constant $c_{10} > 0$ depending on d, p and $\tilde{\varepsilon}$. Again, Lemma 28 iii) for all $\varepsilon_3 > 0$ yields

$$\begin{aligned} &\inf_{(\boldsymbol{\lambda},\boldsymbol{\gamma})\in(\Lambda_n\times\Gamma_n)^{|J_F|}} \mathcal{R}_{L_F,P}(f_{D_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_F,P} \\ &\leq \inf_{(\boldsymbol{\lambda},\boldsymbol{\gamma})\in(\Lambda_n\times\Gamma_n)^{|J_F|}} c_{10}\left(\left(\frac{r_n}{\min_{j\in J_F}\gamma_j}\right)^d \sum_{j\in J_F} \lambda_j n^{\tilde{\varepsilon}}\right. \\ &\qquad \left. + \left(\sum_{j\in J_F} \lambda_j^{-1}\gamma_j^{-\frac{d}{p}} P_X(A_j)\right)^p n^{-1} + \frac{\tau}{r_n^\zeta n}\right) \\ &\leq c_{11}\left(\max\{r_n^{-d}, r_n^{-\zeta}\} \cdot n^{-1+\varepsilon_3} + \tau \cdot r_n^{-\zeta} n^{-1}\right), \end{aligned} \tag{44}$$

where $c_{11} > 0$ is a constant depending on d and ε_3. By inserting (44) into (43) we find

$$\begin{aligned}
&\mathcal{R}_{L_F,P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_F,P} \qquad (45)\\
&\leq 6\left(\inf_{(\boldsymbol{\lambda},\boldsymbol{\gamma})\in(\Lambda_n\times\Gamma_n)^{|J_F|}} \mathcal{R}_{L_F,P}(f_{D_1,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_F,P}\right) + \frac{2c_{\mathrm{LC}}\tau_n^F}{r_n^\zeta n}\\
&\leq c_{12}\left(\max\{r_n^{-d}, r_n^{-\zeta}\}\cdot n^{-1+\varepsilon_3} + \tau\cdot r_n^{-\zeta}n^{-1} + \frac{\tau + \ln(1+|\Lambda_n\times\Gamma_n|^{|J_F|})}{r_n^\zeta n}\right)\\
&\leq c_{12}\left(\max\{r_n^{-d}, r_n^{-\zeta}\}\cdot n^{-1+\varepsilon_3} + 2\tau\cdot r_n^{-\zeta}n^{-1} + \frac{|J_F|\ln(2|\Lambda_n\times\Gamma_n|)}{r_n^\zeta n}\right)\\
&\leq c_{12}\left(\max\{r_n^{-d}, r_n^{-\zeta}\}\cdot n^{-1+\varepsilon_3} + 2\tau\cdot r_n^{-\zeta}n^{-1} + \frac{m_n\ln(2|\Lambda_n\times\Gamma_n|)}{r_n^\zeta n}\right)\\
&\leq c_{13}\left(\max\{r_n^{-d}, r_n^{-\zeta}\}\cdot n^{-1+\varepsilon_3} + 2\tau\cdot r_n^{-\zeta}n^{-1} + \frac{\ln(2|\Lambda_n\times\Gamma_n|)}{r_n^d r_n^\zeta n}\right)\\
&\leq c_{14}\left(\max\{r_n^{-d}, r_n^{-\zeta}\}\cdot n^{-1+\hat{\varepsilon}_3} + 2\tau\cdot r_n^{-\zeta}n^{-1}\right),
\end{aligned}$$

where $\hat{\varepsilon}_3 > 0$ and where $c_{12}, c_{13}, c_{14} > 0$ are constants depending on d, ε_3 resp. $d, \hat{\varepsilon}_3$. A variable transformation in τ then yields

$$\mathcal{R}_{L_F,P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_F,P} \leq c_{15}\tau\cdot\max\{r_n^{-d}, r_n^{-\zeta}\}\cdot n^{-1+\tilde{\varepsilon}_3} \qquad (46)$$

with probability P^n not less than $1 - e^{-\tau}$, where $c_{15} > 0$ is a constant depending on d and $\tilde{\varepsilon}_3$.

Finally, we compose (39), (42), (46) and insert these inequalities into (35). We obtain with probability P^n not less than $1 - 3e^{-\tau}$ that

$$\begin{aligned}
&\mathcal{R}_{L_J,P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_J,P}\\
&\leq \mathcal{R}_{L_{N_1},P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_{N_1},P} + \mathcal{R}_{L_{N_2},P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_{N_2},P}\\
&\qquad + \mathcal{R}_{L_F,P}(f_{D_1,\boldsymbol{\lambda}_{D_2},\boldsymbol{\gamma}_{D_2}}) - \mathcal{R}^*_{L_F,P}\\
&\leq c_{16}\tau\left(n^{-\beta\kappa(\nu+1)+\tilde{\varepsilon}_1} + n^{\tilde{\varepsilon}_2}\left(n^{1-\nu(d-1)}\right)^{-\frac{q+1}{q+2}} + \max\{r_n^{-d}, r_n^{-\zeta}\}\cdot n^{-1+\tilde{\varepsilon}_3}\right)\\
&\leq c_{16}\tau n^{\varepsilon}\left(2n^{-\beta\kappa(\nu+1)} + \max\{r_n^{-d}, r_n^{-\zeta}\}\cdot n^{-1}\right)\\
&\leq c_{17}\tau n^{\varepsilon}\cdot n^{-\beta\kappa(\nu+1)},
\end{aligned}$$

where in the last step we applied $\beta\kappa(\nu+1) \leq 1 - \nu\max\{d,\zeta\}$ analogously to the calculations in the proof of Theorem 2, where $\varepsilon := \max\{\tilde{\varepsilon}_1, \tilde{\varepsilon}_2, \tilde{\varepsilon}_3\}$ and where $c_{16}, c_{17} > 0$ are constants depending on d, β, q and ε. ∎

4.4 Oracle Inequalities and Learning rates on predefined sets

In this subsection, we state the theorems leading to the proof of our main result in Theorem 2. They show the individual oracle inequalities and learning rates on the sets defined in

(10) resp. (15). We present first the general oracle inequality for localized SVMs on that all results are based on and discuss some necessary results concerning entropy numbers of localized Gaussian kernels. After that we decompose our analysis in the following way. We derive in Section 4.4.1 bounds on the approximation error on our predefined sets. Then, in Sections 4.4.2 and 4.4.3 we present the oracle inequalities and learning rates on the sets N_1 resp. N_2 and F.

Note that in this section, for some measure μ we denote by $L_2(\mu)$ the Lebesgue spaces of order 2. We write $\mathrm{D_X}$ for the empirical measure w.r.t. the x-samples of D and we write $P_{X|A}$ for restriction of the marginal distribution P_X onto some set $A \subset X$.

Before we state a more general oracle inequality in the next theorem, we recall the definition of so-called entropy numbers, see (Carl and Stephani, 1990) or (Steinwart and Christmann, 2008, Definition A.5.26), which are necessary to measure the capacity of the underlying RKHS. For normed spaces $(E, \|\cdot\|_E)$ and $(F, \|\cdot\|_F)$, as well as an integer $i \geq 1$, the i-th (dyadic) entropy number of a bounded, linear operator $S : E \to F$ is defined by

$$\begin{aligned} e_i(S : E \to F) &:= e_i(SB_E, \|\cdot\|_F) \\ &:= \inf\left\{\varepsilon > 0 : \exists s_1, \ldots, s_{2^{i-1}} \in SB_E \text{ such that } SB_E \subset \bigcup_{j=1}^{2^{i-1}} (s_j + \varepsilon B_F)\right\}, \end{aligned}$$

where we use the convention $\inf \emptyset := \infty$, and B_E as well as B_F denote the closed unit balls in E and F, respectively.

Theorem 9 (Oracle Inequality for Localized SVMs) *Let $L : X \times Y \times \mathbb{R} \to [0, \infty)$ be the hinge loss. Based on a partition $(A_j)_{j=1,\ldots,m}$ of $B_{\ell_2^d}$, where $\mathring{A}_j \neq \emptyset$ for every $j \in \{1, \ldots, m\}$, we assume **(H)**. Furthermore, for an arbitrary index set $J \subset \{1, \ldots, m\}$, we assume that for $\theta \in [0, 1]$ to be the exponent of the variance bound (14) w.r.t. the loss L_J. Assume that for fixed $n \geq 1$ there exist constants $p \in (0, 1)$ and $a_J > 0$ such that*

$$\mathbb{E}_{D_X \sim P_X^n} e_i(\mathrm{id} : H_J \to L_2(\mathrm{D_X})) \leq a_J \, i^{-\frac{1}{2p}}, \qquad i \geq 1. \tag{47}$$

Finally, fix an $f_0 \in H_J$ with $\|f_0\|_\infty \leq 1$. Then, for all fixed $\tau > 0$, $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_m) > 0$, and $a := \max\{a_J, 2\}$ the localized SVM predictor given by (2) using $\hat{H}_1, \ldots, \hat{H}_m$ and L_J satisfies

$$\begin{aligned} &\sum_{j \in J} \lambda_j \|\widehat{f}_{D_j, \lambda_j}\|_{\hat{H}_j}^2 + \mathcal{R}_{L_J, P}(\widehat{f}_{D, \boldsymbol{\lambda}}) - \mathcal{R}^*_{L_J, P} \\ &\leq 9 \left(\sum_{j \in J} \lambda_j \|\mathbf{1}_{A_j} f_0\|_{\hat{H}_j}^2 + \mathcal{R}_{L_J, P}(f_0) - \mathcal{R}^*_{L_J, P} \right) + C \left(\frac{a^{2p}}{n} \right)^{\frac{1}{2-p-\theta+\theta p}} + 3 \left(\frac{72 V \tau}{n} \right)^{\frac{1}{2-\theta}} + \frac{30\tau}{n} \end{aligned}$$

with probability P^n not less than $1 - 3e^{-\tau}$, where $C > 0$ is a constant only depending on p, V, θ.

Proof We apply (Eberts and Steinwart, 2015, Theorem 5). The hinge loss is Lipschitz continuous and can be clipped at $M = 1$. Since $\|f_0\|_\infty \leq 1$ we have $\|L \circ f_0\|_\infty \leq 2$ such

that $B_0 = 2$. A look into the proof of (Eberts and Steinwart, 2015, Theorem 5) shows that two things can be slightly modified. First, it suffices to assume to have average entropy numbers of the form in (47). Second, it suffices to consider the individual RKHS-norms on the local set $J \subset \{1, \ldots, m\}$ instead of the whole set $J = \{1, \ldots, m\}$. By combining these observations yields the result. ∎

We remark that the constant $C > 0$ in Theorem 9 is exactly the constant from (Steinwart and Christmann, 2008, Theorem 7.23). As the following two lemmata shows, we obtain a bound of the form (47).

Lemma 10 *Let $A \subset B_{\ell_2^d}$ be such that $\mathring{A} \neq \emptyset$ and $A \subset B_r(z)$ with $r > 0, z \in X$. Let $H_\gamma(A)$ be the RKHS of the Gaussian kernel k_γ over A. Then, for all $p \in (0, \frac{1}{2})$ there exists a constant $c_{d,p} > 0$ such that for all $\gamma \leq r$ and $i \geq 1$ we have*

$$e_i(\mathrm{id} : H_\gamma(A) \to L_2(P_{X|A})) \leq c_{d,p} \sqrt{P_X(A)} \cdot r^{\frac{d}{2p}} \gamma^{-\frac{d}{2p}} i^{-\frac{1}{2p}},$$

where $c_{d,p} := (3c_d)^{\frac{1}{2p}} \left(\frac{d+1}{2ep}\right)^{\frac{d+1}{2p}}$.

Proof Following the lines of (Meister and Steinwart, 2016, Theorem 12) we consider the commutative diagram

$$\begin{array}{ccc} H_\gamma(A) & \xrightarrow{\quad \mathrm{id} \quad} & L_2(P_{X|A}) \\ {\scriptstyle I_{B_r}^{-1} \circ I_A} \downarrow & & \uparrow {\scriptstyle \mathrm{id}} \\ H_\gamma(B_r) & \xrightarrow[\quad \mathrm{id} \quad]{} & \ell_\infty(B_r) \end{array}$$

where the extension operator $I_A : H_\gamma(A) \to H_\gamma(\mathbb{R}^d)$ and the restriction operator $I_{B_r}^{-1} : H_\gamma(\mathbb{R}^d) \to H_\gamma(B_r)$, defined in (Steinwart and Christmann, 2008, Theorem 4.37), are isometric isomorphisms such that $\|I_{B_r}^{-1} \circ I_A : H_\gamma(A) \to H_\gamma(B_r)\| = 1$. According to (Steinwart and Christmann, 2008, (A.38) and (A.39)) we then have

$$\begin{aligned} &e_i(\mathrm{id} : H_\gamma(A) \to L_2(P_{X|A})) \\ &\leq \|I_{B_r}^{-1} \circ I_A : H_\gamma(A) \to H_\gamma(B_r)\| \cdot e_i(\mathrm{id} : H_\gamma(B_r) \to \ell_\infty(B_r)) \cdot \|\mathrm{id} : \ell_\infty(B_r) \to L_2(P_{X|A})\|, \end{aligned} \tag{48}$$

where we find for $f \in \ell_\infty(B_r)$ that

$$\|\mathrm{id} : \ell_\infty(B_r) \to L_2(P_{X|A})\| \leq \|f\|_\infty \sqrt{P_X(A)} \tag{49}$$

since

$$\|f\|_{L_2(P_{X|A})} = \left(\int_X \mathbf{1}_A(x) |f(x)|^2 dP_X(x)\right)^{\frac{1}{2}} \leq \|f\|_\infty \cdot \left(\int_X \mathbf{1}_A(x) dP_X(x)\right)^{\frac{1}{2}} \leq \|f\|_\infty \sqrt{P_X(A)}.$$

Furthermore, by (Steinwart and Christmann, 2008, (A.38) and (A.39)) and (Farooq and Steinwart, 2019, Theorem 5) we obtain

$$e_i(\mathrm{id} : H_\gamma(B_r) \to \ell_\infty(B_r)) \leq e_i(\mathrm{id} : H_{\frac{\gamma}{r}}(r^{-1}B) \to \ell_\infty(r^{-1}B)) \leq c_{d,p} \cdot r^{\frac{d}{2p}} \gamma^{-\frac{d}{2p}} i^{-\frac{1}{2p}}, \quad (50)$$

where $c_{d,p} := (3c_d)^{\frac{1}{2p}} \left(\frac{d+1}{2ep}\right)^{\frac{d+1}{2p}}$. Plugging (49) and (50) into (48) yields

$$e_i(\mathrm{id} : H_\gamma(A) \to L_2(P_{X|A})) \leq c_{d,p}\sqrt{P_X(A)} \cdot r^{\frac{d}{2p}} \gamma^{-\frac{d}{2p}} i^{-\frac{1}{2p}}.$$

■

Lemma 11 *Based on a partition $(A_j)_{j=1,\ldots,m}$ of $B_{\ell_2^d}$, where $\mathring{A}_j \neq \emptyset$ and $A_j \subset B_r(z_j)$ for $r > 0, z_j \in B_{\ell_2^d}$ for every $j \in \{1,\ldots,m\}$, we assume **(H)**. Then, for all $p \in (0, \frac{1}{2})$ there exists a constant $\tilde{c}_{d,p} > 0$ such that for all $\gamma_j \leq r$ and $i \geq 1$ we have*

$$e_i(\mathrm{id} : H_J \to L_2(\mathrm{D_X})) \leq \tilde{c}_{d,p}|J|^{\frac{1}{2p}} r^{\frac{d}{2p}} \left(\sum_{j\in J} \lambda_j^{-1} \gamma_j^{-\frac{d}{p}} \mathrm{D_X}(A_j)\right)^{\frac{1}{2}} i^{\frac{1}{2p}}, \qquad i \geq 1,$$

and, for the average entropy numbers we have

$$\mathbb{E}_{D_X \sim P_X^n} e_i(\mathrm{id} : H_J \to L_2(\mathrm{D_X})) \leq \tilde{c}_{d,p}|J|^{\frac{1}{2p}} r^{\frac{d}{2p}} \left(\sum_{j\in J} \lambda_j^{-1} \gamma_j^{-\frac{d}{p}} P_X(A_j)\right)^{\frac{1}{2}} i^{\frac{1}{2p}}, \qquad i \geq 1.$$

The proof shows that the constant is given by $\tilde{c}_{d,p} := 2\,(9\ln(4)c_d)^{\frac{1}{2p}} \left(\frac{d+1}{2ep}\right)^{\frac{d+1}{2p}}$.

Proof We define $a_j := c_{d,p}\sqrt{D_X(A_j)} \cdot r^{\frac{d}{2p}} \gamma_j^{-\frac{d}{2p}}$. By Lemma 10 we have

$$e_i(\mathrm{id} : H_{\gamma_j}(A_j) \to L_2(\mathrm{D_{X|A_j}})) \leq a_j i^{-\frac{1}{2p}}$$

for $j \in J, i \geq 1$. Following the lines of the proof of (Meister and Steinwart, 2016, Theorem 11) we find that

$$e_i(\mathrm{id} : H_J \to L_2(\mathrm{D_X})) \leq 2|J|^{\frac{1}{2}} \left(3\ln(4)\sum_{j\in J} \lambda_j^{-p} a_j^{2p}\right)^{\frac{1}{2p}} i^{\frac{1}{2p}}.$$

By inserting a_j and by applying $\|\cdot\|_{\ell_p^{|J|}}^p \leq |J|^{1-p}\|\cdot\|_{\ell_1^{|J|}}^p$ we obtain

$$\begin{aligned}
e_i(\mathrm{id} : H_J \to L_2(\mathrm{D_X})) &\leq 2|J|^{\frac{1}{2}} \left(3\ln(4)\sum_{j\in J}\lambda_j^{-p}a_j^{2p}\right)^{\frac{1}{2p}} i^{\frac{1}{2p}} \\
&= 2|J|^{\frac{1}{2}}(3\ln(4))^{\frac{1}{2p}} \left(\sum_{j\in J}\lambda_j^{-p}\left(c_{d,p}\sqrt{\mathrm{D_X}(A_j)}\cdot r^{\frac{d}{2p}}\gamma_j^{-\frac{d}{2p}}\right)^{2p}\right)^{\frac{1}{2p}} i^{\frac{1}{2p}} \\
&= c_{d,p}2(3\ln(4))^{\frac{1}{2p}}|J|^{\frac{1}{2}}r^{\frac{d}{2p}} \left(\sum_{j\in J}\left(\lambda_j^{-1}\gamma_j^{-\frac{d}{p}}\mathrm{D_X}(A_j)\right)^p\right)^{\frac{1}{2p}} i^{\frac{1}{2p}} \\
&\leq \tilde{c}_{d,p}|J|^{\frac{1}{2}}r^{\frac{d}{2p}}|J|^{\frac{1-p}{2p}} \left(\sum_{j\in J}\lambda_j^{-1}\gamma_j^{-\frac{d}{p}}\mathrm{D_X}(A_j)\right)^{\frac{1}{2}} i^{\frac{1}{2p}} \\
&= \tilde{c}_{d,p}|J|^{\frac{1}{2p}}r^{\frac{d}{2p}} \left(\sum_{j\in J}\lambda_j^{-1}\gamma_j^{-\frac{d}{p}}\mathrm{D_X}(A_j)\right)^{\frac{1}{2}} i^{\frac{1}{2p}},
\end{aligned}$$

where $\tilde{c}_{d,p} := c_{d,p}2(3\ln(4))^{\frac{1}{2p}}$ and $c_{d,p}$ is the constant from Lemma 10. Finally, by considering the above inequality in expectation yields

$$\begin{aligned}
\mathbb{E}_{D_X\sim P_X^n} e_i(\mathrm{id} : H_J \to L_2(\mathrm{D_X})) &\leq \tilde{c}_{d,p}|J|^{\frac{1}{2p}}r^{\frac{d}{2p}} \left(\sum_{j\in J}\lambda_j^{-1}\gamma_j^{-\frac{d}{p}}\mathbb{E}_{D_X\sim P_X^n}\mathrm{D_X}(A_j)\right)^{\frac{1}{2}} i^{\frac{1}{2p}} \\
&\leq \tilde{c}_{d,p}|J|^{\frac{1}{2p}}r^{\frac{d}{2p}} \left(\sum_{j\in J}\lambda_j^{-1}\gamma_j^{-\frac{d}{p}}P_X(A_j)\right)^{\frac{1}{2}} i^{\frac{1}{2p}}.
\end{aligned}$$

■

4.4.1 Bounds on Approximation Error

We define for an $f_0 : X \to \mathbb{R}$ the function

$$A_J^{(\gamma)}(\boldsymbol{\lambda}) := \sum_{j\in J}\lambda_j\|\mathbf{1}_{A_j}f_0\|_{\mathring{H}_j}^2 + \mathcal{R}_{L_J,P}(f_0) - \mathcal{R}_{L_J,P}^*. \tag{51}$$

Recall that we aim to find an $f_0 \in H_J$ such that both, the norm and the approximation error in $A_J^{(\gamma)}(\boldsymbol{\lambda})$ are small. We show in the following that a suitable choice for f_0 is a function that is constructed by convolutions of some $f \in L^2(\mathbb{R}^d)$ with the function $K_\gamma : \mathbb{R}^d \to \mathbb{R}$, defined by

$$K_\gamma(x) = \left(\frac{2}{\pi^{1/2}\gamma}\right)^{d/2} e^{-2\gamma^{-2}\|x-\cdot\|_2^2}. \tag{52}$$

Note that $K_\gamma * f(x) = \langle \Phi_\gamma(x), f\rangle_{L_2(\mathbb{R}^d)}$ for $x \in \mathbb{R}^d$, where Φ_γ is a feature map of a Gaussian kernel (see Steinwart and Christmann, 2008, Lemma 4.45). The following lemma shows that a restriction of the convolution is contained in a local RKHS and that we control the individual RKHS norms in (51).

Lemma 12 (Convolution) *Let $A \subset B_r(z)$ for some $r > 0, z \in B_{\ell_2^d}$. Furthermore, let $H_\gamma(A)$ be the RKHS of the Gaussian kernel k_γ over A with $\gamma > 0$ and let the function $K_\gamma : \mathbb{R}^d \to \mathbb{R}$ by defined as in (52). Moreover, for $\rho \geq r$ define the function $f_\gamma^\rho : \mathbb{R}^d \to \mathbb{R}$ by*

$$f_\gamma^\rho(x) := (\pi\gamma^2)^{-d/4} \cdot \mathbf{1}_{B_\rho(z)}(x) \cdot \tilde{f}(x),$$

where $\tilde{f} : \mathbb{R}^d \to \mathbb{R}$ is some function with $\|\tilde{f}\|_\infty \leq 1$. Then, we have $\|K_\gamma * f_\gamma^\rho\|_\infty \leq 1$ and $\mathbf{1}_A(K_\gamma * f_\gamma^\rho) \in \hat{H}_\gamma(A)$ with

$$\|\mathbf{1}_A(K_\gamma * f_\gamma^\rho)\|^2_{\hat{H}_\gamma(A)} \leq \left(\frac{\rho^2}{\pi\gamma^2}\right)^{d/2} \mathrm{vol}_d(B).$$

Proof Obviously, $f_\gamma^\rho \in L_2(\mathbb{R}^d)$ such that we find

$$\begin{aligned}
\|f_\gamma^\rho\|^2_{L_2(\mathbb{R}^d)} &= \int_{\mathbb{R}^d} |(\pi\gamma^2)^{-d/4} \cdot \mathbf{1}_{B_\rho(z)}(x)\tilde{f}(x)|^2\, dx \\
&\leq (\pi\gamma^2)^{-d/2} \int_{\mathbb{R}^d} |\mathbf{1}_{B_\rho(z)}(x)|^2\, dx \\
&= (\pi\gamma^2)^{-d/2} \int_{B_\rho(z)} 1\, dx \\
&= (\pi\gamma^2)^{-d/2}\, \mathrm{vol}_d(B_\rho(z)) \\
&= \left(\frac{\rho^2}{\pi\gamma^2}\right)^{d/2} \mathrm{vol}_d(B).
\end{aligned} \tag{53}$$

Since the map $K_\gamma * \cdot : L_2(\mathbb{R}^d) \to H_\gamma(A)$ given by

$$K_\gamma * g(x) := \left(\frac{2}{\pi^{1/2}\gamma}\right)^{d/2} \int_{\mathbb{R}^d} e^{-2\gamma^{-2}\|x-y\|_2^2} \cdot g(y)\, dy, \qquad g \in L_2(\mathbb{R}^d), x \in A$$

is a metric surjection (see Steinwart and Christmann, 2008, Proposition 4.46), we find

$$\|(K_\gamma * f_\gamma^\rho)_{|A}\|^2_{H_\gamma(A)} \leq \|f_\gamma^\rho\|^2_{L_2(\mathbb{R}^d)}. \tag{54}$$

Next, Young's inequality (see Steinwart and Christmann, 2008, Theorem A.5.23) yields

$$\|K_\gamma * f_\gamma^\rho\|_\infty = \left(\frac{2}{\pi\gamma^2}\right)^{d/2} \|k_\gamma * (\mathbf{1}_{B_\rho(z)}\tilde{f})\|_\infty \leq \left(\frac{2}{\pi\gamma^2}\right)^{d/2} \|k_\gamma\|_1 \|\tilde{f}\|_\infty \leq 1. \tag{55}$$

Hence, with (54) and (53) we find

$$\|\mathbf{1}_A(K_\gamma * f_\gamma^\rho)\|^2_{\hat{H}_\gamma(A)} = \|(K_\gamma * f_\gamma^\rho)_{|A}\|^2_{H_\gamma(A)} \leq \|f_\gamma^\rho\|^2_{L_2(\mathbb{R}^d)} \leq \left(\frac{\rho^2}{\pi\gamma^2}\right)^{d/2} \mathrm{vol}_d(B).$$

■

In order to bound the excess risks in (51) over the sets N_1, N_2 and F we apply Zhang's equality given by

$$\mathcal{R}_{L_J,P}(f_0) - \mathcal{R}^*_{L_J,P} = \int_{\bigcup_{j\in J} A_j} |f_0 - f^*_{L_{\text{class}},P}||2\eta - 1| dP_X \tag{56}$$

(see Steinwart and Christmann, 2008, Theorem 2.31). We begin with an analysis on the set N_1, whose cells have no intersection with the decision boundary. For such cells the subsequent lemma presents a suitable function f_0 and its difference to $f^*_{L_{\text{class}},P}$ that occurs in (56). In particular, the function f_0 is a convolution of K_γ with $2\eta - 1$ since we have $f^*_{L_{\text{class}},P}(x) = 2\eta(x) - 1$ for $x \in A_j$ with $j \in J^s_{N_1}$ as mentioned in Section 3.2.

Lemma 13 (Convolution on N_1 and its difference to $f^*_{L_{\text{class}},P}$) *Let the assumptions of Lemma 12 be satisfied with $A \cap X_1 \neq \emptyset$ and $A \cap X_{-1} \neq \emptyset$. We define the function $f^{3r}_\gamma : \mathbb{R}^d \to \mathbb{R}$ by*

$$f^{3r}_\gamma(x) := (\pi\gamma^2)^{-d/4} \cdot \mathbf{1}_{B_{3r}(z)\cap(X_1\cup X_{-1})}(x)\,\text{sign}(2\eta(x) - 1). \tag{57}$$

Then, we find for all $x \in A$ that

$$|K_\gamma * f^{3r}_\gamma(x) - f^*_{L_{\text{class}},P}(x)| \leq \frac{2}{\Gamma(d/2)} \int_{2\Delta^2_\eta(x)\gamma^{-2}}^\infty e^{-t} t^{d/2-1}\, dt,$$

where K_γ is the function defined in (52).

Proof Let us consider w.l.o.g. $x \in A \cap X_1$. Then,

$$\Delta_\eta(x) = \inf_{\bar{x}\in X_{-1}} \|x - \bar{x}\|_2 \leq \text{diam}(B_r(z)) = 2r. \tag{58}$$

Next, we denote by $\mathring{B}$ the open ball and show that $\mathring{B}_{\Delta_\eta(x)}(x) \subset B_{3r}(z) \cap X_1$. For $x' \in \mathring{B}_{\Delta_\eta(x)}(x)$ we have $\|x' - x\|_2 < \Delta_\eta(x)$ such that $x' \in X_1$. Furthermore, (58) yields $\|x' - z_j\|_2 \leq \|x' - x\|_2 + \|x - z_j\|_2 < \Delta_\eta(x) + r \leq 2r + r = 3r$ and hence $x' \in B_{3r}(z)$. We find

$$\begin{aligned}
K_\gamma * f^{3r}_\gamma(x) &= \left(\frac{2}{\pi^{1/2}\gamma}\right)^{d/2} \int_{\mathbb{R}^d} e^{-2\gamma^{-2}\|x-y\|_2^2} (\pi\gamma^2)^{-d/4} \mathbf{1}_{B_{3r}(z)\cap(X_1\cup X_{-1})}(y) \text{sign}\,(2\eta(y) - 1)\, dy \\
&= \left(\frac{2}{\pi\gamma^2}\right)^{d/2} \int_{\mathbb{R}^d} e^{-2\gamma^{-2}\|x-y\|_2^2} \cdot \mathbf{1}_{B_{3r}(z)\cap(X_1\cup X_{-1})}(y) \text{sign}\,(2\eta(y) - 1)\, dy \\
&= \left(\frac{2}{\pi\gamma^2}\right)^{d/2} \left(\int_{B_{3r}(z)\cap X_1} e^{-2\gamma^{-2}\|x-y\|_2^2} dy - \int_{B_{3r}(z)\cap X_{-1}} e^{-2\gamma^{-2}\|x-y\|_2^2} dy \right) \\
&\geq \left(\frac{2}{\pi\gamma^2}\right)^{d/2} \left(\int_{\mathring{B}_{\Delta_\eta(x)}(x)} e^{-2\gamma^{-2}\|x-y\|_2^2} dy - \int_{\mathbb{R}^d\setminus\mathring{B}_{\Delta_\eta(x)}(x)} e^{-2\gamma^{-2}\|x-y\|_2^2} dy \right) \\
&= 2\left(\frac{2}{\pi\gamma^2}\right)^{d/2} \int_{\mathring{B}_{\Delta_\eta(x)}(x)} e^{-2\gamma^{-2}\|x-y\|_2^2} dy - 1.
\end{aligned}$$

Since $f^*_{L_{\text{class}},P}(x) = 1$ we obtain by Lemma 12 for $\rho = 3r$ and $\tilde{f} = \mathbf{1}_{X_1 \cup X_{-1}} \operatorname{sign}(2\eta - 1)$, and by Lemma 27 that

$$\begin{aligned}
|K_\gamma * f^{3r}_\gamma(x) - f^*_{L_{\text{class}},P}(x)| &= |K_\gamma * f^{3r}_\gamma(x) - 1| \\
&= 1 - K_\gamma * f^{3r}_\gamma(x) \\
&\leq 2 - 2\left(\frac{2}{\pi\gamma^2}\right)^{d/2} \int_{\mathring{B}_{\Delta_\eta(x)}(x)} e^{-2\gamma^{-2}\|x-y\|_2^2}\, dy \\
&= 2 - \frac{2}{\Gamma(d/2)} \int_0^{2\Delta^2_\eta(x)\gamma^{-2}} e^{-t} t^{d/2-1}\, dt \qquad (59) \\
&= \frac{2}{\Gamma(d/2)} \left(\int_0^\infty e^{-t} t^{d/2-1}\, dt - \int_0^{2\Delta^2_\eta(x)\gamma^{-2}} e^{-t} t^{d/2-1}\, dt \right) \\
&= \frac{2}{\Gamma(d/2)} \int_{2\Delta^2_\eta(x)\gamma^{-2}}^\infty e^{-t} t^{d/2-1}\, dt.
\end{aligned}$$

The case $x \in A \cap X_0$ is clear and for $x \in A \cap X_{-1}$ the calculation yields the same inequality, hence (59) holds for all $x \in A$. ∎

Under the assumption that P has some MNE β we immediately obtain in the next theorem a bound on the approximation error on the set N_1.

Theorem 14 (Approximation Error on N_1) *Let **(A)** and **(H)** be satisfied and let P have MNE $\beta \in (0, \infty]$. Define the set of indices*

$$J := \{\, j \in \{1, \dots m\} \,|\, A_j \cap X_1 \neq \emptyset \text{ and } A_j \cap X_{-1} \neq \emptyset \,\}.$$

and the function $f_0 : X \to \mathbb{R}$ by

$$f_0 := \sum_{j \in J} \mathbf{1}_{A_j} \left(K_{\gamma_j} * f^{3r}_{\gamma_j} \right),$$

where the functions K_γ and $f^{3r}_{\gamma_j}$ are defined in (52) and (57). Then, $f_0 \in H_J$ and $\|f_0\|_\infty \leq 1$. Moreover, there exist constants $c_d, c_{d,\beta} > 0$ such that

$$A^{(\boldsymbol{\gamma})}_J(\boldsymbol{\lambda}) \leq c_d \cdot \sum_{j \in J} \frac{\lambda_j r^d}{\gamma_j^d} + c_{d,\beta} \cdot \max_{j \in J} \gamma_j^\beta.$$

Proof By Lemma 12 for $\rho = 3r$ and $\tilde{f} = \mathbf{1}_{X_1 \cup X_{-1}} \operatorname{sign}(2\eta - 1)$ we have immediately that $f_0 \in H_J$ as well as $\|f_0\|_\infty = \|K_\gamma * f^{3r}_\gamma\|_\infty \leq 1$. Moreover, Lemma 12 yields

$$\sum_{j \in J} \lambda_j \|\mathbf{1}_{A_j} f_0\|^2_{\hat{H}_j} = \sum_{j \in J} \lambda_j \|\mathbf{1}_{A_j} (K_{\gamma_j} * f^{3r}_{\gamma_j})\|^2_{\hat{H}_j} \leq c_d \cdot \sum_{j \in J} \frac{\lambda_j r^d}{\gamma_j^d}$$

for some constant $c_d > 0$. Next, we bound the excess risk of f_0. To this end, we fix w.l.o.g. an $x \in A_j \cap X_1$ and find

$$\Delta_\eta(x) = \inf_{\bar{x} \in X_{-1}} \|x - \bar{x}\|_2 \leq \operatorname{diam}(B_r(z)) = 2r$$

such that $A_j \subset \{\Delta_\eta(x) \leq 2r\}$ for every $j \in J$. Together with Zhang's equality (see Steinwart and Christmann, 2008, Theorem 2.31), and Lemma 13 we then obtain

$$
\begin{aligned}
&\mathcal{R}_{L_J,P}(f_0) - \mathcal{R}^*_{L_J,P} \\
&= \sum_{j\in J} \int_{A_j} |(K_{\gamma_j} * f^{3r}_{\gamma_j})(x) - f^*_{L_{\text{class}},P}(x)||2\eta(x) - 1|\, dP_X(x) \\
&\leq \frac{2}{\Gamma(d/2)} \sum_{j\in J} \int_{A_j} \int_{2\Delta^2_\eta(x)\gamma_j^{-2}}^{\infty} e^{-t} t^{d/2-1}\, dt |2\eta(x) - 1|\, dP_X(x) \\
&= \frac{2}{\Gamma(d/2)} \sum_{j\in J} \int_{A_j} \int_0^{\infty} \mathbf{1}_{[2\Delta^2_\eta(x)\gamma_j^{-2},\infty)}(t) e^{-t} t^{d/2-1}\, dt |2\eta(x) - 1|\, dP_X(x) \\
&= \frac{2}{\Gamma(d/2)} \sum_{j\in J} \int_{A_j} \int_0^{\infty} \mathbf{1}_{[0,\sqrt{t/2}\gamma_j)}(\Delta_\eta(x)) e^{-t} t^{d/2-1}\, dt |2\eta(x) - 1|\, dP_X(x) \\
&= \frac{2}{\Gamma(d/2)} \int_0^{\infty} \sum_{j\in J} \int_{A_j} \mathbf{1}_{[0,\sqrt{t/2}\gamma_j)}(\Delta_\eta(x)) |2\eta(x) - 1|\, dP_X(x) e^{-t} t^{d/2-1}\, dt \\
&\leq \frac{2}{\Gamma(d/2)} \int_0^{\infty} \int_{\{\Delta_\eta(x)\leq 2r\}} \mathbf{1}_{[0,\sqrt{t/2}\gamma_{\max})}(\Delta_\eta(x)) |2\eta(x) - 1|\, dP_X(x) e^{-t} t^{d/2-1}\, dt \\
&\leq \frac{2}{\Gamma(d/2)} \int_0^{\infty} \int_{\{\Delta_\eta(x)\leq \min\{2r,\sqrt{t/2}\gamma_{\max}\}\}} |2\eta(x) - 1|\, dP_X(x) e^{-t} t^{d/2-1}\, dt.
\end{aligned}
$$

Next, a simple calculation shows that

$$
\min\left\{2r, \sqrt{t/2}\gamma_{\max}\right\} = \begin{cases} 2r, & \text{if } t \geq 8r^2\gamma_{\max}^{-2}, \\ \sqrt{t/2}\gamma_{\max}, & \text{if else.} \end{cases}
$$

and that $1 \leq \left(\frac{t\gamma_{\max}^2}{8r^2}\right)^{\beta/2}$ for $t \geq 8r^2\gamma_{\max}^{-2}$. Finally, the definition of the margin-noise exponent β yields

$$
\begin{aligned}
&\mathcal{R}_{L_J,P}(f_0) - \mathcal{R}^*_{L_J,P} \\
&\leq \frac{2}{\Gamma(d/2)} \int_0^\infty \int_{\{\Delta_\eta(x) \leq \min\{2r, \sqrt{t/2}\gamma_{\max}\}\}} |2\eta - 1| \, dP_X e^{-t} t^{d/2-1} \, dt \\
&\leq \frac{2c_{\text{MNE}}^\beta}{\Gamma(d/2)} \int_0^\infty \left(\min\{2r, \sqrt{t/2}\gamma_{\max}\}\right)^\beta e^{-t} t^{d/2-1} \, dt \\
&= \frac{2c_{\text{MNE}}^\beta}{\Gamma(d/2)} \left(\int_0^{8r^2\gamma_{\max}^{-2}} \gamma_{\max}^\beta \left(\frac{t}{2}\right)^{\beta/2} e^{-t} t^{d/2-1} \, dt + \int_{8r^2\gamma_{\max}^{-2}}^\infty (2r)^\beta e^{-t} t^{d/2-1} \, dt \right) \\
&\leq \frac{2c_{\text{MNE}}^\beta}{\Gamma(d/2)} \left(\frac{\gamma_{\max}^\beta}{2^{\beta/2}} \int_0^{8r^2\gamma_{\max}^{-2}} e^{-t} t^{(d+\beta)/2-1} \, dt + (2r)^\beta \int_{8r^2\gamma_{\max}^{-2}}^\infty e^{-t} t^{d/2-1} \, dt \right) \\
&\leq \frac{2c_{\text{MNE}}^\beta}{\Gamma(d/2)} \left(\frac{\gamma_{\max}^\beta}{2^{\beta/2}} \int_0^{8r^2\gamma_{\max}^{-2}} e^{-t} t^{(d+\beta)/2-1} \, dt + (2r)^\beta \left(\frac{\gamma_{\max}^2}{8r^2}\right)^{\beta/2} \int_{8r^2\gamma_{\max}^{-2}}^\infty e^{-t} t^{(d+\beta)/2-1} \, dt \right) \\
&= \frac{2c_{\text{MNE}}^\beta}{\Gamma(d/2)} \frac{\gamma_{\max}^\beta}{2^{\beta/2}} \left(\int_0^{8r^2\gamma_{\max}^{-2}} e^{-t} t^{(d+\beta)/2-1} \, dt + \int_{8r^2\gamma_{\max}^{-2}}^\infty e^{-t} t^{(d+\beta)/2-1} \, dt \right) \\
&= \frac{2^{1-\beta/2} c_{\text{MNE}}^\beta \Gamma((d+\beta)/2)}{\Gamma(d/2)} \gamma_{\max}^\beta .
\end{aligned}
$$

■

In the next step we develop bounds on the approximation error on sets that have no intersection with the decision boundary, that is, N_2 and F. Recall that we apply (56) and again, the subsequent lemma presents a suitable function f_0 and its difference to $f^*_{L_{\text{class}},P}$ that occurs in (56). Note that on those sets we have $f^*_{L_{\text{class}},P}(x) = 1$ for $x \in A_j$ with $j \in J^s_{N_2}$ or $j \in J^s_F$ and hence, we we choose a function $f_0 \in H_J$ that is a convolution of K_γ, defined in (52), with a constant function that we have to cut off to ensure that it is an element of $L_2(\mathbb{R}^d)$. Unfortunately, we will always make an error on such cells since (Steinwart and Christmann, 2008, Corollary 4.44) shows that Gaussian RKHSs on an open ball do not contain constant functions. In order to make the convoluted function as flat as possible on a cell, we choose the radius ω_+ of the ball on which f is a constant arbitrary large, that is $\omega_+ > r$. We remark that although the radius is arbitrary large we receive by convolution a function that is still contained in a local RKHS over a cell A_j.

Lemma 15 (Difference to $f^*_{L_{\text{class}},P}$ on cells in N_2 or F) *Let the assumptions of Lemma 12 be satisfied with $A \cap X_1 = \emptyset$ or $A \cap X_{-1} = \emptyset$. For $\omega_- > 0$ we define $\omega_+ := \omega_- + r$ and the function $f_\gamma^{\omega_+} : \mathbb{R}^d \to \mathbb{R}$ by*

$$
f_\gamma^{\omega_+}(x) := \begin{cases} (\pi\gamma^2)^{-d/4} \cdot \mathbf{1}_{B_{\omega_+}(z) \cap (X_1 \cup X_0)}(x), & \text{if } x \in A \cap (X_1 \cup X_0), \\ (-1) \cdot (\pi\gamma^2)^{-d/4} \cdot \mathbf{1}_{B_{\omega_+}(z) \cap X_{-1}}(x), & \text{else.} \end{cases} \tag{60}
$$

Then, we find for all $x \in A$ that

$$|(K_\gamma * f_\gamma^{\omega_+})(x) - f^*_{L_{\text{class}},P}(x)| \leq \frac{1}{\Gamma(d/2)} \int_{(\omega_-)^2 2\gamma^{-2}}^{\infty} e^{-t} t^{d/2-1}\, dt,$$

where K_γ is the function defined in (52)*.*

Proof We assume w.l.o.g. that $x \in A \cap (X_1 \cup X_0)$ and show in a first step that $\mathring{B}_{\omega_-}(x) \subset B_{\omega_+}(z)$. To this end, consider an $x' \in \mathring{B}_{\omega_-}(x)$. Since $A \subset B_r(z)$ we find

$$\|x' - z\|_2 \leq \|x' - x\|_2 + \|x - z\|_2 < \omega_- + r = \omega_+$$

and hence, $x' \in B_{\omega_+}(z)$. Next, we obtain with Lemma 27 that

$$\begin{aligned}
K_\gamma * f_\gamma^{\omega_+}(x) &= \left(\frac{2}{\pi^{1/2}\gamma}\right)^{d/2} \int_{\mathbb{R}^d} e^{-2\gamma^{-2}\|x-y\|_2^2} (\pi\gamma^2)^{-d/4} \cdot \mathbf{1}_{B_{\omega_+}(z)}\, dy \\
&= \left(\frac{2}{\pi\gamma^2}\right)^{d/2} \int_{B_{\omega_+}(z)} e^{-2\gamma^{-2}\|x-y\|_2^2}\, dy \\
&\geq \left(\frac{2}{\pi\gamma^2}\right)^{d/2} \int_{\mathring{B}_{\omega_-}(x)} e^{-2\gamma^{-2}\|x-y\|_2^2}\, dy \\
&= \frac{1}{\Gamma(d/2)} \int_0^{2(\omega_-)^2\gamma^{-2}} e^{-t} t^{d/2-1}\, dt.
\end{aligned}$$

Since $f^*_{L_{\text{class}},P}(x) = 1$, we finally obtain with Lemma 12 for $\rho = \omega_+$ and $\tilde{f} := \mathbf{1}_{X_1 \cup X_0}$ that

$$\begin{aligned}
|(K_\gamma * f_\gamma^{\omega_+})(x) - f^*_{L_{\text{class}},P}(x)| &= |(K_\gamma * f_\gamma^{\omega_+})(x) - 1| \\
&= 1 - (K_\gamma * f_\gamma^{\omega_+})(x) \\
&\leq 1 - \frac{1}{\Gamma(d/2)} \int_0^{2(\omega_-)^2\gamma^{-2}} e^{-t} t^{d/2-1}\, dt \\
&= \frac{1}{\Gamma(d/2)} \int_{2(\omega_-)^2\gamma^{-2}}^{\infty} e^{-t} t^{d/2-1}\, dt.
\end{aligned}$$

For $x \in A \cap X_{-1}$ the latter calculations yields with $\tilde{f} := \mathbf{1}_{X_{-1}}$ the same results and hence, the latter inequality holds for all $x \in A$. ■

In the next theorem we state bounds on the approximation error over the sets F and N_2. We obtain directly a bound for the set F, however, to obtain a bound on N_2 we need the additional assumption that P has MNE β.

Theorem 16 (Approximation Error on F and N_2) *Let **(A)** and **(H)** be satisfied and define the set of indices*

$$J := \{\, j \in \{1, \dots m\} \,|\, A_j \cap X_1 = \emptyset \text{ or } A_j \cap X_{-1} = \emptyset \,\}.$$

For some $\omega_- > 0$ define $\omega_+ := \omega_- + r > 0$ and let the function $f_{\gamma_j}^{\omega_+}$ for every $j \in J$ be defined as in (60)*. Moreover, define the function $f_0 : X \to \mathbb{R}$ by*

$$f_0 := \bigcup_{j\in J} \mathbf{1}_{A_j} \left(K_{\gamma_j} * f_{\gamma_j}^{\omega_+} \right).$$

Then, $f_0 \in H_J$ and $\|f_0\|_\infty \leq 1$. Furthermore, for all $\xi > 0$ there exist constants $c_d, c_{d,\xi} > 0$ such that

$$A_J^{(\gamma)}(\boldsymbol{\lambda}) \leq c_d \cdot \sum_{j\in J} \lambda_j \left(\frac{\omega_+}{\gamma_j} \right)^d + c_{d,\xi} \left(\frac{\max_{j\in J} \gamma_j}{\omega_-} \right)^{2\xi} \sum_{j\in J} P_X(A_j).$$

In addition, if P has MNE $\beta \in (0, \infty]$ and we have $A_j \subset \{\Delta_\eta(x) \leq s\}$ for every $j \in J$, then

$$A_J^{(\gamma)}(\boldsymbol{\lambda}) \leq c_d \cdot \sum_{j\in J} \lambda_j \left(\frac{\omega_+}{\gamma_j} \right)^d + c_{d,\xi} \left(\frac{\max_{j\in J} \gamma_j}{\omega_-} \right)^{2\xi} (c_{\mathrm{MNE}} \cdot s)^\beta$$

Proof By Lemma 12 with $\rho = \omega_+$ and $\tilde{f} := \mathbf{1}_{X_1 \cup X_0}$ resp. $\tilde{f} := \mathbf{1}_{X_{-1}}$ we have immediately $f_0 \in H_J$ and $\|f_0\|_\infty = \|K_{\gamma_j} * f_{\gamma_j}^{\omega_+}\|_\infty \leq 1$. Moreover, Lemma 12 yields

$$\sum_{j\in J} \lambda_j \|\mathbf{1}_{A_j} f_0\|_{\hat{H}_j}^2 = \sum_{j\in J} \lambda_j \|\mathbf{1}_{A_j} (K_{\gamma_j} * f_{\gamma_j}^{\omega_+})\|_{\hat{H}_j}^2 \leq c_d \cdot \sum_{j\in J} \lambda_j \left(\frac{\omega_+}{\gamma_j} \right)^d$$

for some constant $c_d > 0$. Next, we bound the excess risk of f_0. We find by applying Zhang's equality (e.g., Steinwart and Christmann, 2008, Theorem 2.31), Lemma 15 and (Steinwart and Christmann, 2008, Lemma A.1.1) for some arbitrary $\xi > 0$ that

$$\begin{aligned}
\mathcal{R}_{L_J,P}(f_0) - \mathcal{R}_{L_J,P}^* &= \sum_{j\in J} \int_{A_j} |(K_{\gamma_j} * f_{\gamma_j}^{\omega_+}) - f_{L_{\mathrm{class}},P}^*| |2\eta - 1| \, dP_X \\
&\leq \sum_{j\in J} \int_{A_j} \frac{1}{\Gamma(d/2)} \int_{(\omega_-)^2 2\gamma_j^{-2}}^\infty e^{-t} t^{d/2-1} \, dt |2\eta - 1| \, dP_X \\
&\leq \frac{1}{\Gamma(d/2)} \int_{2(\omega_-)^2\gamma_{\max}^{-2}}^\infty e^{-t} t^{d/2-1} \, dt \sum_{j\in J} \int_{A_j} |2\eta - 1| \, dP_X \\
&\leq \frac{\Gamma\left(d/2, 2(\omega_-)^2\gamma_{\max}^{-2}\right)}{\Gamma(d/2)} \sum_{j\in J} \int_{A_j} |2\eta - 1| \, dP_X \\
&\leq \frac{2^{-\xi}\Gamma(d/2+\xi)}{\Gamma(d/2)} \left(\frac{\gamma_{\max}}{\omega_-} \right)^{2\xi} \cdot \sum_{j\in J} P_X(A_j).
\end{aligned}$$

If in addition P has MNE β and $A_j \subset \{\Delta_\eta \leq s\}$ for every $j \in J$ we modify the previous calculation of the excess risk. Then, we obtain again with Zhang's equality, Lemma 15 and

(Steinwart and Christmann, 2008, Lemma A.1.1) for some arbitrary $\xi > 0$ that

$$\begin{aligned}
\mathcal{R}_{L_J,P}(f_0) - \mathcal{R}^*_{L_J,P} &\leq \sum_{j\in J} \int_{A_j} \frac{1}{\Gamma(d/2)} \int_{(\omega_-)^2 2\gamma_j^{-2}}^{\infty} e^{-t} t^{d/2-1}\, dt |2\eta - 1|\, dP_X \\
&\leq \frac{1}{\Gamma(d/2)} \int_{(\omega_-)^2 2\gamma_{\max}^{-2}}^{\infty} e^{-t} t^{d/2-1}\, dt \int_{\{\Delta_\eta \leq s\}} |2\eta - 1|\, dP_X \\
&\leq \frac{c_{\mathrm{MNE}}^{\beta} \Gamma(d/2, (\omega_-)^2 2\gamma_{\max}^{-2})}{\Gamma(d/2)} \cdot s^{\beta} \\
&\leq \frac{c_{\mathrm{MNE}}^{\beta} 2^{-\xi} \Gamma(d/2+\xi)}{\Gamma(d/2)} \left(\frac{\gamma_{\max}}{\omega_-}\right)^{2\xi} s^{\beta}.
\end{aligned}$$

By combining the results for the norm and the excess risk yields finally the bounds on the respective approximation error. ∎

Both bounds in the theorem above depend on the parameter $\xi > 0$. However, we will observe in the theorems in Section 4.4.3, which state the corresponding oracle inequalities, that by setting ω_- appropriately this ξ will not have an influence any more.

4.4.2 Oracle inequalities and learning rates on N_1

Based on the the general oracle inequality in Section 9 and the results from the previous section we establish in this section an oracle inequality on the set N_1 and derive learning rates.

Theorem 17 (Oracle Inequality on N_1) *Let P have MNE $\beta \in (0,\infty]$ and NE $q \in [0,\infty]$ and let **(G)** and **(H)** be satisfied. Moreover, let **(A)** be satisfied for some $r := n^{-\nu}$ with $\nu > 0$ and define the set of indices*

$$J := \{\, j \in \{1,\dots,m\} \,|\, \forall x \in A_j :\, P_X(A_j \cap X_1) > 0 \text{ and } P_X(A_j \cap X_{-1}) > 0 \,\}.$$

Let $\tau \geq 1$ be fixed and define $n^* := \left(\frac{4}{\delta^*}\right)^{\frac{1}{\nu}}$. Then, for all $p \in (0, \frac{1}{2})$, $n \geq n^*$, $\boldsymbol{\lambda} := (\lambda_1,\dots,\lambda_m) \in (0,\infty)^m$ and $\boldsymbol{\gamma} := (\gamma_1,\dots,\gamma_m) \in (0,r]^m$ the SVM given in (2) satisfies

$$\begin{aligned}
&\mathcal{R}_{L_J,P}(\widehat{f}_{D,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_J,P} \\
&\leq 9 c_{d,\beta} \left(\sum_{j\in J} \frac{\lambda_j r^d}{\gamma_j^d} + \max_{j\in J} \gamma_j^{\beta} \right) + c_{d,p,q} \left(\frac{r}{n}\right)^{\frac{q+1}{q+2-p}} \left(\sum_{j\in J} \lambda_j^{-1} \gamma_j^{-\frac{d}{p}} P_X(A_j) \right)^{\frac{p(q+1)}{q+2-p}} + \tilde{c}_{p,q} \left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}}
\end{aligned} \tag{61}$$

with probability P^n not less than $1 - 3e^{-\tau}$ and for some constants $c_{d,\beta}, c_{d,p,q} > 0$ and $\tilde{c}_{p,q} > 0$ only depending on d, β, p, q.

Proof We apply the generic oracle inequality given in Theorem 9 and bound first of all the contained constant a^{2p} . To this end, we remark that an analogous calculation as in the

proof of Theorem 14 shows that $A_j \subset \{\Delta_\eta \leq 2r\}$ for every $j \in J$. Since

$$n \geq \left(\frac{4}{\delta^*}\right)^{\frac{1}{\nu}} \Leftrightarrow 4r \leq \delta^*$$

we obtain by Lemma 26 for $t = 2r$ that

$$|J| \leq c_1 r^{-d+1}, \tag{62}$$

where c_1 is a positive constant only depending on d. Together with Lemma 11 we then find that

$$\begin{aligned}
a^{2p} &= \max\left\{\tilde{c}_{d,p}|J|^{\frac{1}{2p}} r^{\frac{d}{2p}} \left(\sum_{j\in J} \lambda_j^{-1}\gamma_j^{-\frac{d}{p}} P_X(A_j)\right)^{\frac{1}{2}}, 2\right\}^{2p} \\
&\leq \tilde{c}_{d,p}^{2p}|J|r^d \left(\sum_{j\in J} \lambda_j^{-1}\gamma_j^{-\frac{d}{p}} P_X(A_j)\right)^{p} + 4^p \\
&\leq c_1\tilde{c}_{d,p}^{2p} \cdot r \left(\sum_{j\in J} \lambda_j^{-1}\gamma_j^{-\frac{d}{p}} P_X(A_j)\right)^{p} + 4^p,
\end{aligned}$$

where $c_{d,p} := 2c_1\left(9\ln(4)c_d\right)^{\frac{1}{2p}}\left(\frac{d+1}{2ep}\right)^{\frac{d+1}{2p}}$. Moreover, (Steinwart and Christmann, 2008, Lemma 8.24) delivers a variance bound for $\theta = \frac{q}{q+1}$ and constant $V := 6c_{NE}^{\frac{q}{q+1}}$. We denote by $A_J^{(\gamma)}(\boldsymbol{\lambda})$ the approximation error, defined in (51). Then, we have by Theorem 9 with $\tau \geq 1$ that

$$\begin{aligned}
&\mathcal{R}_{L_J,P}(\widehat{f}_{D,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_J,P} \\
&\leq 9A_J^{(\gamma)}(\boldsymbol{\lambda}) + c_{p,q}\left(\frac{a^{2p}}{n}\right)^{\frac{q+1}{q+2-p}} + 3c_{NE}^{\frac{q}{q+2}}\left(\frac{432\tau}{n}\right)^{\frac{q+1}{q+2}} + \frac{30\tau}{n} \\
&\leq 9A_J^{(\gamma)}(\boldsymbol{\lambda}) + c_{p,q}\left[c_1\tilde{c}_{d,p}^{2p} \cdot r\left(\sum_{j\in J} \lambda_j^{-1}\gamma_j^{-\frac{d}{p}} P_X(A_j)\right)^{p} + 4^p\right]^{\frac{q+1}{q+2-p}} \cdot n^{-\frac{q+1}{q+2-p}} + c_q\left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}} \\
&\leq 9A_J^{(\gamma)}(\boldsymbol{\lambda}) + c_{d,p,q}\left(\frac{r}{n}\right)^{\frac{q+1}{q+2-p}}\left(\sum_{j\in J} \lambda_j^{-1}\gamma_j^{-\frac{d}{p}} P_X(A_j)\right)^{\frac{p(q+1)}{q+2-p}} + c_{p,q}4^{\frac{p(q+1)}{q+2-p}} n^{-\frac{q+1}{q+2}} + c_q\left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}} \\
&\leq 9A_J^{(\gamma)}(\boldsymbol{\lambda}) + c_{d,p,q}\left(\frac{r}{n}\right)^{\frac{q+1}{q+2-p}}\left(\sum_{j\in J} \lambda_j^{-1}\gamma_j^{-\frac{d}{p}} P_X(A_j)\right)^{\frac{p(q+1)}{q+2-p}} + \tilde{c}_{p,q}\left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}}
\end{aligned}$$

holds with probability P^n not less than $1 - 3e^{-\tau}$ and with positive constants $c_{d,p,q} := c_{p,q}\left(c_1\tilde{c}_{d,p}^{2p}\right)^{\frac{q+1}{q+2-p}}$, $c_q := 2\max\left\{3c_{NE}^{\frac{q}{q+2}}432^{\frac{q+1}{q+2}}, 30\right\}$ and $\tilde{c}_{p,q} := 2\max\left\{c_{p,q}4^{\frac{p(q+1)}{q+2-p}}, c_q\right\}$. Fi-

nally, Theorem 14 yields for the approximation error the bound

$$A_J^{(\gamma)}(\boldsymbol{\lambda}) \leq c_2 \left(\sum_{j \in J} \frac{\lambda_j r^d}{\gamma_j^d} + \max_{j \in J} \gamma_j^\beta \right),$$

where $c_2 > 0$ is a constant depending on d and β. By plugging this into the oracle inequality above yields the result. ∎

Theorem 18 (Learning Rates on N_1) *Let the assumptions of Theorem 17 be satisfied with m_n and with*

$$\begin{aligned} r_n &\simeq n^{-\nu}, \\ \gamma_{n,j} &\simeq r_n^\kappa n^{-\kappa}, \\ \lambda_{n,j} &\simeq n^{-\sigma}, \end{aligned} \tag{63}$$

for all $j \in \{1, \ldots, m_n\}$. Moreover, define $\kappa := \frac{q+1}{\beta(q+2)+d(q+1)}$ and let

$$\nu \leq \frac{\kappa}{1-\kappa} \tag{64}$$

and $\sigma \geq 1$ be satisfied. Then, for all $\varepsilon > 0$ there exists a constant $c_{\beta,d,\varepsilon,\sigma,q} > 0$ such that for $\boldsymbol{\lambda}_n := (\lambda_{n,1}, \ldots, \lambda_{n,m_n}) \in (0,\infty)^m$, and $\boldsymbol{\gamma}_n := (\gamma_{n,1}, \ldots, \gamma_{n,m_n}) \in (0, r_n]^{m_n}$, and all n sufficiently large we have with probability P^n not less than $1 - 3e^{-\tau}$ that

$$\mathcal{R}_{L_J,P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_J,P} \leq c_{\beta,d,\varepsilon,\sigma,q} \cdot \tau^{\frac{q+1}{q+2}} \cdot r_n^{\beta\kappa} n^{-\beta\kappa+\varepsilon}.$$

In particular, the proof shows that one can even choose $\sigma \geq \kappa(\beta+d)(\nu+1) - \nu > 0$.

Proof We write $\lambda_n := n^{-\sigma}$ and $\gamma_n := r_n^\kappa n^{-\kappa}$. As in the proof of Theorem 17 we find

$$|J| \leq c_d r_n^{-d+1}$$

for some constant $c_d > 0$. Together with Theorem 17 we then obtain that

$$\begin{aligned} &\mathcal{R}_{L_J,P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_J,P} \\ &\leq c_1 \left(\sum_{j \in J} \frac{\lambda_{n,j} r^d}{\gamma_{n,j}^d} + \max_{j \in J} \gamma_{n,j}^\beta + \left(\frac{r}{n}\right)^{\frac{q+1}{q+2-p}} \left(\sum_{j \in J} \lambda_{n,j}^{-1} \gamma_{n,j}^{-\frac{d}{p}} P_X(A_j) \right)^{\frac{p(q+1)}{q+2-p}} + \left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}} \right) \\ &\leq c_2 \left(|J| \frac{\lambda_n r_n^d}{\gamma_n^d} + \gamma_n^\beta + \left(\frac{r_n}{n}\right)^{\frac{q+1}{q+2-p}} \left(\lambda_n^{-1} \gamma_n^{-\frac{d}{p}} \sum_{j \in J} P_X(A_j) \right)^{\frac{p(q+1)}{q+2-p}} + \left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}} \right) \\ &\leq c_2 \left(\frac{\lambda_n r_n}{\gamma_n^d} + \gamma_n^\beta + \left(\frac{r_n \lambda_n^{-p} \gamma_n^{-d}}{n} \right)^{\frac{q+1}{q+2-p}} + \left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}} \right) \end{aligned}$$

holds with probability P^n not less than $1-3e^{-\tau}$ and for some positive constant c_1, c_2 depending on β, d, p and q. Moreover, with (63), $\sigma \geq \kappa(\beta+d)(\nu+1)-\nu$ and $\frac{(1-d\kappa)(q+1)}{q+2} = \left(\frac{\beta(q+2)}{\beta(q+2)+d(q+1)}\right)\frac{q+1}{q+2} = \beta\kappa$ we find

$$
\begin{aligned}
&\mathcal{R}_{L_J,P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_J,P} \\
&\leq c_2 \left(\frac{\lambda_n r_n}{\gamma_n^d} + \gamma_n^\beta + \left(\frac{r_n \lambda_n^{-p} \gamma_n^{-d}}{n} \right)^{\frac{q+1}{q+2-p}} + \left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}} \right) \\
&= c_2 \left(\frac{r_n}{n^{\kappa(\beta+d)(\nu+1)-\nu}\gamma_n^d} + r_n^{\beta\kappa} n^{-\beta\kappa} + \left(\frac{r_n^{1-d\kappa}}{n^{1-d\kappa}} \right)^{\frac{q+1}{q+2-p}} \left(n^{-\sigma}\right)^{\frac{p(q+1)}{q+2-p}} + \left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}} \right) \\
&\leq c_3 \left(\frac{n^{-\nu}}{n^{\kappa(\beta+d)(\nu+1)-\nu} n^{-\nu d\kappa} n^{-d\kappa}} + r_n^{\beta\kappa} n^{-\beta\kappa} + \left(\frac{r_n}{n}\right)^{\frac{(1-d\kappa)(q+1)}{q+2-p}} n^{\varepsilon} + \left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}} \right) \\
&\leq c_3 \left(n^{-\nu\beta\kappa} n^{-\beta\kappa} + r_n^{\beta\kappa} n^{-\beta\kappa} + \left(\frac{r_n}{n}\right)^{\frac{(1-d\kappa)(q+1)}{q+2}} n^{\varepsilon} + \left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}} \right) \\
&\leq c_4 \left(r_n^{\beta\kappa} n^{-\beta\kappa+\varepsilon} + \tau^{\frac{q+1}{q+2}} n^{-\frac{q+1}{q+2}} \right) \\
&\leq c_5 \tau^{\frac{q+1}{q+2}} \cdot r_n^{\beta\kappa} n^{-\beta\kappa+\varepsilon},
\end{aligned}
$$

where p is chosen sufficiently small such that $\varepsilon \geq \frac{p\sigma(q+1)}{q+2} \geq 0$ and where the constants $c_3, c_4, c_5 > 0$ depend on $\beta, d, \varepsilon, \sigma$ and q. ∎

4.4.3 Oracle inequalities and learning rates on N_2, F

Based on the the general oracle inequality in Section 9 and the results from the previous section we establish in this section an oracle inequality on the set N_2 and F. Moreover, we derive learning rates.

Theorem 19 (Oracle inequality on N_2) *Let P have MNE $\beta \in (0,\infty]$ and NE $q \in [0,\infty]$ and let **(G)** and **(H)** be satisfied. Moreover, let **(A)** be satisfied for some $r := n^{-\nu}$ with $\nu > 0$. Define for $s := n^{-\alpha}$ with $\alpha > 0$ and $\alpha \leq \nu$ the set of indices*

$$J := \{\, j \in \{1,\dots,m\} \,|\, \forall x \in A_j : \Delta_\eta(x) \leq 3s \text{ and } P_X(A_j \cap X_1) = 0 \text{ or } P_X(A_j \cap X_{-1}) = 0 \,\}.$$

Let $\tau \geq 1$ be fixed and define $n^* := \left(4^{-1}\delta^*\right)^{-\frac{1}{\alpha}}$. Then, for all $\varepsilon > 0$, $p \in (0, \frac{1}{2})$, $n \geq n^*$, $\boldsymbol{\lambda} := (\lambda_1,\dots,\lambda_m) \in (0,\infty)^m$, and $\boldsymbol{\gamma} := (\gamma_1,\dots,\gamma_m) \in (0,r]^m$ the SVM given in (2) satisfies

$$
\begin{aligned}
&\mathcal{R}_{L_J,P}(\widehat{f}_{D,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_J,P} \\
&\leq \left(\frac{c_{d,\beta,\varepsilon,q} \cdot r}{\min_{j\in J}\gamma_j} \right)^d n^{\varepsilon} \sum_{j\in J} \lambda_j + c_{d,p,q} \left(\frac{s}{n}\right)^{\frac{q+1}{q+2-p}} \left(\sum_{j\in J} \lambda_j^{-1} \gamma_j^{-\frac{d}{p}} P_X(A_j) \right)^{\frac{p(q+1)}{q+2-p}} + c_{d,\beta,\varepsilon,p,q} \left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}}
\end{aligned} \tag{65}
$$

with probability P^n not less than $1-3e^{-\tau}$ and with constants $c_{d,\beta,\varepsilon,q}, c_{d,p,q} > 0$ and $c_{d,\beta,\varepsilon,p,q} > 0$.

Proof We apply the generic oracle inequality given in Theorem 9 and bound first of all the contained constant a^{2p} .To this end, we remark that we find with $\alpha \leq \nu$ that

$$\left(4^{-1}\delta^*\right)^{-\frac{1}{\alpha}} \leq n \qquad \Rightarrow \qquad \delta^* \geq 4n^{-\alpha} \geq 3n^{-\alpha} + n^{-\nu} = 3s + r$$

such that we obtain by Lemma 26 for $t = 3s$ that

$$|J| \leq c_1 \cdot sr^{-d}, \tag{66}$$

where c_1 is a positive constant only depending on d. Together with Lemma 11 we then find for the constant a^{2p} from Theorem 9 that

$$\begin{aligned}
a^{2p} &= \max\left\{ \tilde{c}_{d,p}|J|^{\frac{1}{2p}} r^{\frac{d}{2p}} \left(\sum_{j\in J} \lambda_j^{-1} \gamma_j^{-\frac{d}{p}} P_X(A_j) \right)^{\frac{1}{2}}, 2 \right\}^{2p} \\
&\leq \tilde{c}_{d,p}^{2p} \cdot |J| r^d \left(\sum_{j\in J} \lambda_j^{-1} \gamma_j^{-\frac{d}{p}} P_X(A_j) \right)^{p} + 4^p \\
&\leq c_1 \tilde{c}_{d,p}^{2p} \cdot s \left(\sum_{j\in J} \lambda_j^{-1} \gamma_j^{-\frac{d}{p}} P_X(A_j) \right)^{p} + 4^p,
\end{aligned}$$

where $c_{d,p} := 2c_1 \left(9\ln(4)c_d\right)^{\frac{1}{2p}} \left(\frac{d+1}{2ep}\right)^{\frac{d+1}{2p}}$. Again, (Steinwart and Christmann, 2008, Lemma 8.24) yields a variance bound for $\theta = \frac{q}{q+1}$ and constant $V := 6c_{NE}^{\frac{q}{q+1}}$. We denote by $A_J^{(\gamma)}(\boldsymbol{\lambda})$ the approximation error, defined in (51), and find by Theorem 9 with $\tau \geq 1$ that

$$\begin{aligned}
&\mathcal{R}_{L_J,P}(\widehat{f}_{D,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_J,P} \\
&\leq 9A_J^{(\gamma)}(\boldsymbol{\lambda}) + c_{p,q}\left(\frac{a^{2p}}{n}\right)^{\frac{q+1}{q+2-p}} + 3c_{NE}^{\frac{q}{q+2}}\left(\frac{432\tau}{n}\right)^{\frac{q+1}{q+2}} + \frac{30\tau}{n} \\
&\leq 9A_J^{(\gamma)}(\boldsymbol{\lambda}) + c_{p,q}\left[c_1\tilde{c}_{d,p}^{2p} \cdot s\left(\sum_{j\in J} \lambda_j^{-1}\gamma_j^{-\frac{d}{p}} P_X(A_j)\right)^{p} + 4^p\right]^{\frac{q+1}{q+2-p}} n^{-\frac{q+1}{q+2-p}} + c_q\left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}} \\
&\leq 9A_J^{(\gamma)}(\boldsymbol{\lambda}) + c_{d,p,q}\left(\frac{s}{n}\right)^{\frac{q+1}{q+2-p}} \left(\sum_{j\in J} \lambda_j^{-1}\gamma_j^{-\frac{d}{p}} P_X(A_j)\right)^{\frac{p(q+1)}{q+2-p}} + c_{p,q}4^{\frac{p(q+1)}{q+2-p}} \cdot n^{-\frac{q+1}{q+2}} + c_q\left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}} \\
&\leq 9A_J^{(\gamma)}(\boldsymbol{\lambda}) + c_{d,p,q}\left(\frac{s}{n}\right)^{\frac{q+1}{q+2-p}} \left(\sum_{j\in J} \lambda_j^{-1}\gamma_j^{-\frac{d}{p}} P_X(A_j)\right)^{\frac{p(q+1)}{q+2-p}} + \tilde{c}_{p,q}\left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}}
\end{aligned} \tag{67}$$

holds with probability P^n not less than $1-3e^{-\tau}$ and with positive constants $c_{d,p,q} := c_{p,q}\left(c_1\tilde{c}_{d,p}^{2p}\right)^{\frac{q+1}{q+2-p}}$, $c_q := 2\max\left\{3c_{NE}^{\frac{q}{q+2}}432^{\frac{q+1}{q+2}}, 30\right\}$ and $\tilde{c}_{p,q} := 2\max\left\{c_{p,q}4^{\frac{p(q+1)}{q+2-p}}, c_q\right\}$. Finally, Theorem 16 for $\omega_- := \gamma_{\max}n^{\frac{q+1}{2\xi(q+2)}}$, where $\xi > 0$, and $\omega_+ := \omega_- + r$, yields

$$
\begin{aligned}
A_J^{(\boldsymbol{\gamma})}(\boldsymbol{\lambda}) &\le c_2\left(\sum_{j\in J}\lambda_j\left(\frac{\omega_+}{\gamma_j}\right)^d + \left(\frac{\gamma_{\max}}{\omega_-}\right)^{2\xi}s^\beta\right)\\
&\le c_2\left(\left(\frac{\omega_+}{\gamma_{\min}}\right)^d\sum_{j\in J}\lambda_j + \left(\frac{\gamma_{\max}}{\omega_-}\right)^{2\xi}s^\beta\right)\\
&= c_2\left(\left(\frac{\gamma_{\max}n^{\frac{q+1}{2\xi(q+2)}}+r}{\gamma_{\min}}\right)^d\sum_{j\in J}\lambda_j + n^{-\frac{q+1}{q+2}}s^\beta\right) \qquad (68)\\
&\le c_3\left(n^{\frac{d(q+1)}{2\xi(q+2)}}\left(\frac{\gamma_{\max}+r}{\gamma_{\min}}\right)^d\sum_{j\in J}\lambda_j + n^{-\frac{q+1}{q+2}}s^\beta\right)\\
&\le c_4\left(n^\varepsilon\left(\frac{r}{\gamma_{\min}}\right)^d\sum_{j\in J}\lambda_j + n^{-\frac{q+1}{q+2}}s^\beta\right),
\end{aligned}
$$

where in the last step that we applied $\gamma_{\max}\le r$, and where we picked an arbitrary $\varepsilon > 0$ and chose ξ sufficiently large such that $\varepsilon \ge \frac{d(q+1)}{2\xi(q+2)} > 0$. The constants $c_2, c_3 > 0$ only depend on d, β and ξ, whereas $c_4 > 0$ depends only on d, β, ε and q. By plugging this into the oracle inequality above yields

$$
\begin{aligned}
&\mathcal{R}_{L_J,P}(\widehat{f}_{D,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_J,P}\\
&\le 9c_4\left(n^\varepsilon\left(\frac{r}{\gamma_{\min}}\right)^d\sum_{j\in J}\lambda_j + n^{-\frac{q+1}{q+2}}s^\beta\right)\\
&\qquad + c_{d,p,q}\left(\frac{s}{n}\right)^{\frac{q+1}{q+2-p}}\left(\sum_{j\in J}\lambda_j^{-1}\gamma_j^{-\frac{d}{p}}P_X(A_j)\right)^{\frac{p(q+1)}{q+2-p}} + \tilde{c}_{p,q}\left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}}\\
&\le 9c_4n^\varepsilon\left(\frac{r}{\gamma_{\min}}\right)^d\sum_{j\in J}\lambda_j + c_{d,p,q}\left(\frac{s}{n}\right)^{\frac{q+1}{q+2-p}}\left(\sum_{j\in J}\lambda_j^{-1}\gamma_j^{-\frac{d}{p}}P_X(A_j)\right)^{\frac{p(q+1)}{q+2-p}} + c_{d,\beta,\varepsilon,p,q}\left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}}.
\end{aligned}
$$

■

Theorem 20 (Learning Rates on N_2) *Let the assumption of Theorem 19 be satisfied for m_n, $s \simeq s_n$ and*

$$\begin{aligned} r_n &\simeq n^{-\nu}, \\ \gamma_{n,j} &\simeq r_n, \\ \lambda_{n,j} &\simeq n^{-\sigma} \end{aligned} \tag{69}$$

with some $\sigma \geq 1$ and $1 + \alpha - \nu d > 0$, and for all $j \in \{1, \ldots, m_n\}$. Then, for all $\varepsilon > 0$ there exists a constant $c_{\beta,d,\varepsilon,\sigma,q} > 0$ such that for $\boldsymbol{\lambda}_n := (\lambda_{n,1}, \ldots, \lambda_{n,m_n}) \in (0,\infty)^m$, and $\boldsymbol{\gamma}_n := (\gamma_{n,1}, \ldots, \gamma_{n,m_n}) \in (0, r_n]^{m_n}$, and all n sufficiently large we have with probability P^n not less than $1 - 3e^{-\tau}$ that

$$\mathcal{R}_{L_J,P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_J,P} \leq c_{\beta,d,\varepsilon,\sigma,q} \tau^{\frac{q+1}{q+2}} \cdot \left(\frac{s_n}{r_n^d}\right)^{\frac{q+1}{q+2}} n^{-\frac{q+1}{q+2}+\varepsilon}.$$

Proof We write $\lambda_n := n^{-\sigma}$ and $\gamma_n := r_n$. By Theorem 19, Lemma 26 and (69) we find with probability P^n not less than $1 - 3e^{-\tau}$ that

$$\begin{aligned} &\mathcal{R}_{L_J,P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_J,P} \\ &\leq c_1 \left(\left(\frac{r}{\gamma_{\min}}\right)^d \sum_{j\in J} \lambda_{n,j} n^{\varepsilon} + \left(\frac{s_n}{n}\right)^{\frac{q+1}{q+2-p}} \left(\sum_{j\in J} \lambda_{n,j}^{-1} \gamma_{n,j}^{-\frac{d}{p}} P_X(A_j) \right)^{\frac{p(q+1)}{q+2-p}} + \left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}} \right) \\ &\leq c_2 \left(|J| \lambda_n n^{\varepsilon} + \left(\frac{s_n}{\gamma_n^d n}\right)^{\frac{q+1}{q+2-p}} \left(\lambda_n^{-1} \sum_{j\in J} P_X(A_j) \right)^{\frac{p(q+1)}{q+2-p}} + \left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}} \right) \\ &\leq c_2 \left(\frac{s_n \lambda_n n^{\varepsilon}}{r_n^d} + \left(\frac{s_n}{r_n^d n}\right)^{\frac{q+1}{q+2-p}} \lambda_n^{-\frac{p(q+1)}{q+2-p}} + \left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}} \right) \\ &\leq c_2 \left(\frac{s_n n^{\varepsilon}}{r_n^d n^{\sigma}} + \left(\frac{s_n}{r_n^d n}\right)^{\frac{q+1}{q+2}} n^{\frac{p\sigma(q+1)}{q+2-p}} + \left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}} \right) \\ &\leq c_2 \left(\frac{s_n n^{\varepsilon}}{r_n^d n} + \left(\frac{s_n}{r_n^d n}\right)^{\frac{q+1}{q+2}} n^{\hat{\varepsilon}} + \left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}} \right) \\ &\leq c_3 \left(\frac{s_n n^{\varepsilon}}{r_n^d n} + \left(\frac{s_n}{r_n^d n}\right)^{\frac{q+1}{q+2}} n^{\hat{\varepsilon}} + \left(\frac{\tau}{n}\right)^{\frac{q+1}{q+2}} \right) \\ &\leq c_4 \tau^{\frac{q+1}{q+2}} n^{\varepsilon} \left(\left(\frac{s_n}{r_n^d n}\right)^{\frac{q+1}{q+2}} + n^{-\frac{q+1}{q+2}} \right) \\ &\leq c_5 \tau^{\frac{q+1}{q+2}} \left(\frac{s_n}{r_n^d}\right)^{\frac{q+1}{q+2}} n^{-\frac{q+1}{q+2}+\varepsilon}, \end{aligned}$$

where we chose p sufficiently small such that $\varepsilon \geq \frac{p\sigma(q+1)}{q+2-p} > 0$. The constants $c_1, c_2 > 0$ depend only on d, β, ε, p and q, whereas the constants $c_3, c_4, c_5 > 0$ depend on $d, \beta, \varepsilon, \sigma$ and q. ■

Theorem 21 (Oracle Inequality on F) *Let P have LC $\zeta \in [0,\infty)$ and let **(G)** and **(H)** be satisfied. Moreover, let **(A)** be satisfied for some $r := n^{-\nu}$ with $\nu > 0$. Define for $s := n^{-\alpha}$ with $\alpha > 0$ and $\alpha \leq \nu$ the set of indices*

$$J := \{\, j \in \{1,\dots,m\} \,|\, \forall x \in A_j : \Delta_\eta(x) \geq s \,\}.$$

Furthermore, let $\tau \geq 1$ be fixed. Then, for all $\varepsilon > 0$, $p \in (0, \frac{1}{2})$, $n \geq 1$, $\boldsymbol{\lambda} := (\lambda_1,\dots,\lambda_m) \in (0,\infty)^m$, and $\boldsymbol{\gamma} := (\gamma_1,\dots,\gamma_m) \in (0,r]^m$ the SVM given in (2) satisfies

$$\begin{aligned}
&\mathcal{R}_{L_J,P}(\widehat{f}_{D,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_J,P} \\
&\leq \left(\frac{c_{d,\varepsilon}\cdot r}{\min_{j\in J}\gamma_j}\right)^d n^\varepsilon \sum_{j\in J}\lambda_j + c_{d,p}\left(\sum_{j\in J}\lambda_j^{-1}\gamma_j^{-\frac{d}{p}}P_X(A_j)\right)^p n^{-1} + c_{d,\varepsilon,p}\cdot\frac{\tau}{s^\zeta n}
\end{aligned} \tag{70}$$

with probability P^n not less than $1-3e^{-\tau}$ and some constants $c_{d,\varepsilon}, c_{p,q}, c_{d,\varepsilon,p} > 0$.

Proof We apply the generic oracle inequality given in Theorem 9. To this end, we find for the contained constant a^{2p} with Lemma 10 and (4) that

$$\begin{aligned}
a^{2p} &= \max\left\{\tilde{c}_{d,p}|J|^{\frac{1}{2p}} r^{\frac{d}{2p}}\left(\sum_{j\in J}\lambda_j^{-1}\gamma_j^{-\frac{d}{p}}P_X(A_j)\right)^{\frac{1}{2}}, 2\right\}^{2p} \\
&\leq \tilde{c}_{d,p}^{2p}|J|r^d\left(\sum_{j\in J}\lambda_j^{-1}\gamma_j^{-\frac{d}{p}}P_X(A_j)\right)^p + 4^p \\
&\leq c_1\tilde{c}_{d,p}^{2p}\left(\sum_{j\in J}\lambda_j^{-1}\gamma_j^{-\frac{d}{p}}P_X(A_j)\right)^p + 4^p
\end{aligned} \tag{71}$$

where $c_1 > 0$ is a constant depending on d. According to Lemma 5 we have variance bound $\theta = 1$ and constant $V := 2c_{\mathrm{LC}}s^{-\zeta}$. We denote by $A^{(\gamma)}(\boldsymbol{\lambda})$ the approximation error, defined in (51), and obtain by Theorem 9 together with (71) with probability P^n not less than $1-3e^{-\tau}$ that

$$\begin{aligned}
&\mathcal{R}_{L_J,P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_J,P} \\
&\leq 9A_J^{(\gamma)}(\boldsymbol{\lambda}) + \frac{c_p\cdot a^{2p}}{n} + \frac{432c_{\mathrm{LC}}\tau}{s^\zeta n} + \frac{30\tau}{n} \\
&\leq 9A_J^{(\gamma)}(\boldsymbol{\lambda}) + c_pc_1\tilde{c}_{d,p}^{2p}\left(\sum_{j\in J}\lambda_j^{-1}\gamma_j^{-\frac{d}{p}}P_X(A_j)\right)^p n^{-1} + \frac{c_p4^p}{n} + \frac{432c_{\mathrm{LC}}\tau}{s^\zeta n} + \frac{30\tau}{n} \\
&\leq 9A_J^{(\gamma)}(\boldsymbol{\lambda}) + c_{d,p}\left(\sum_{j\in J}\lambda_j^{-1}\gamma_j^{-\frac{d}{p}}P_X(A_j)\right)^p n^{-1} + \hat{c}_p\frac{\tau}{s^\zeta n}
\end{aligned} \tag{72}$$

for some constants $c_{d,p}, \hat{c}_p > 0$. For the approximation error Theorem 16 with $\omega_- := \gamma_{\max} n^{\frac{1}{2\xi}}$, where $\xi > 0$, and $\omega_+ := \omega_- + r$, yields

$$\begin{aligned} A_J^{(\gamma)}(\boldsymbol{\lambda}) &\leq c_2 \left(\cdot \sum_{j \in J} \lambda_j \left(\frac{\omega_+}{\gamma_j} \right)^d + c_4 \cdot \left(\frac{\gamma_{\max}}{\omega_-} \right)^{2\xi} P_X(F) \right) \\ &\leq c_2 \left(\left(\frac{\omega_+}{\gamma_{\min}} \right)^d \sum_{j \in J} \lambda_j + c_4 \cdot \left(\frac{\gamma_{\max}}{\omega_-} \right)^{2\xi} \right) \\ &= c_2 \left(\left(\frac{\gamma_{\max} n^{\frac{1}{2\xi}}}{\gamma_{\min}} + \frac{r}{\gamma_{\min}} \right)^d \sum_{j \in J} \lambda_j + n^{-1} \right) \\ &= c_3 \left(n^{\varepsilon} \left(\frac{r}{\gamma_{\min}} \right)^d \sum_{j \in J} \lambda_j + n^{-1} \right), \end{aligned}$$

where we applied in the last step that $\gamma_{\max} \leq r$ and where we fixed an ε and chose ξ sufficiently large such that $\varepsilon \geq \frac{d}{2\xi} > 0$. The constants $c_2 > 0$ and $c_3 > 0$ only depend on d, ξ resp. d, ε. By combining the results above we have

$$\begin{aligned} &\mathcal{R}_{L_J,P}(f_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_J,P} \\ &\leq 9c_3 \left(n^{\varepsilon} \left(\frac{r}{\gamma_{\min}} \right)^d \sum_{j \in J} \lambda_j + n^{-1} \right) + c_{d,p} \left(\sum_{j \in J} \lambda_j^{-1} \gamma_j^{-\frac{d}{p}} P_X(A_j) \right)^p n^{-1} + \hat{c}_p \frac{\tau}{s^\zeta n} \\ &\leq 9c_3 n^{\varepsilon} \left(\frac{r}{\gamma_{\min}} \right)^d \sum_{j \in J} \lambda_j + c_{d,p} \left(\sum_{j \in J} \lambda_j^{-1} \gamma_j^{-\frac{d}{p}} P_X(A_j) \right)^p n^{-1} + c_4 \frac{\tau}{s^\zeta n} \end{aligned}$$

for some constant $c_4 > 0$ depending on d, ε and p. ∎

Theorem 22 (Learning Rate on F) *Let the assumptions of Theorem 21 be satisfied for m_n, $s \simeq s_n$ and with*

$$\begin{aligned} r_n &\simeq n^{-\nu}, \\ \gamma_{n,j} &\simeq r_n \\ \lambda_{n,j} &\simeq n^{-\sigma}, \end{aligned} \tag{73}$$

for all $j \in \{1, \ldots, m_n\}$ and with $\max\{\nu d, \alpha\zeta\} < 1$ and $\sigma \geq 1$. Then, for all $\varepsilon > 0$ there exists a constant $c_{d,\varepsilon,\sigma} > 0$ such that for $\boldsymbol{\lambda}_n := (\lambda_{n,1}, \ldots, \lambda_{n,m_n}) > 0$, and $\boldsymbol{\gamma}_n := (\gamma_{n,1}, \ldots, \gamma_{n,m_n}) \in (0, r_n]^{m_n}$, and all $n \geq 1$ we have with probability P^n not less than $1 - 3e^{-\tau}$ that

$$\mathcal{R}_{L_J,P}(\widehat{f}_{D,\boldsymbol{\lambda}_n,\boldsymbol{\gamma}_n}) - \mathcal{R}^*_{L_J,P} \leq c_{d,\varepsilon,\sigma} \tau \cdot \max\{r_n^{-d}, s_n^{-\zeta}\} n^{-1+\varepsilon}.$$

Proof We write $\lambda_n := n^{-\sigma}$ and $\gamma_n := r_n$. Then, we obtain by Theorem 21 and (73) with probability P^n not less than $1 - 3e^{-\tau}$ that

$$\begin{aligned}
\mathcal{R}_{L_J,P}(\widehat{f}_{D,\boldsymbol{\lambda},\boldsymbol{\gamma}}) - \mathcal{R}^*_{L_J,P} &\leq c_1 \left(n^{\varepsilon} \left(\frac{r_n}{\min_{j\in J} \gamma_j} \right)^d \sum_{j\in J} \lambda_{n,j} + \left(\sum_{j\in J} \lambda_{n,j}^{-1} \gamma_{n,j}^{-\frac{d}{p}} P_X(A_j) \right)^p n^{-1} + \frac{\tau}{s_n^{\zeta} n} \right) \\
&\leq c_2 \left(n^{\varepsilon} |J| \lambda_n + \lambda_n^{-p} r_n^{-d} \left(\sum_{j\in J} P_X(A_j) \right)^p n^{-1} + \frac{\tau}{s_n^{\zeta} n} \right) \\
&\leq c_2 \tau \left(r_n^{-d} n^{-\sigma+\varepsilon} + n^{\sigma p} r_n^{-d} n^{-1} + s_n^{-\zeta} n^{-1} \right) \\
&\leq c_3 \tau \left(r_n^{-d} n^{-1+\varepsilon} + n^{\varepsilon} r_n^{-d} n^{-1} + s_n^{-\zeta} n^{-1} \right) \\
&\leq c_3 \tau \left(2 r_n^{-d} n^{-1+\varepsilon} + s_n^{-\zeta} n^{-1} \right) \\
&\leq c_4 \tau \cdot \max\{r_n^{-d}, s_n^{-\zeta}\} n^{-1+\varepsilon}
\end{aligned}$$

where p is chosen sufficiently small such that $\varepsilon \geq p\sigma > 0$ and where the constants $c_1, c_2 > 0$ depend only on d, ε, p and the constants $c_3, c_4 > 0$ only on d, ε, σ. ■

Appendix A.

In this appendix we state some results on margin conditions.

Lemma 23 (Reverse Hölder yields lower control) *Let (X, d) be a metric space 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 $X_0 = \partial_X X_1 = \partial_X X_{-1}$. If η is reverse Hölder-continuous with exponent $\rho \in (0, 1]$, that is, if there exists a constant $c > 0$ such that*

$$|\eta(x) - \eta(x')| \geq c \cdot d(x, x')^\rho, \qquad x, x' \in X,$$

then, Δ_η controls the noise from below by the exponent ρ.

Proof We fix w.l.o.g. an $x \in X_{-1}$. By the reverse Hölder continuity we obtain

$$c\Delta_\eta^\rho(x) = c \inf_{\tilde{x} \in X_1} (d(x, \tilde{x}))^\rho \leq \inf_{\tilde{x} \in X_1} |\eta(x) - \eta(\tilde{x})| \leq \inf_{\tilde{x} \in X_1} \eta(\tilde{x}) - \eta(x).$$

Since $\eta(\tilde{x}) > 1/2$ for all $\tilde{x} \in X_1$, we find by continuity of η and $\partial X_1 = X_0$ that $\inf_{\tilde{x} \in X_1} \eta(\tilde{x}) = 1/2$. Thus,

$$\Delta_\eta^\rho(x) \leq (2c)^{-1}(1 - 2\eta(x)).$$

Obviously, the last inequality is immediately satisfied for $x \in X_0$ and for $x \in X_1$ the calculation is similar. Hence, Δ_η controls the noise by the exponent ρ from below, that is,

$$\Delta_\eta^\rho(x) \leq c_{\mathrm{LC}}|2\eta(x) - 1|, \qquad x \in X,$$

where $c_{\mathrm{LC}} := (2c)^{-1}$. ∎

Lemma 24 (LC and ME yield NE) *Let (X, d) be a metric space and let P be a probability measure on $X \times \{-1, 1\}$ that has ME $\alpha \in [0, \infty)$ for the version η of its posterior probability. Assume that the associated distance to the decision boundary controls the noise from below by the exponent $\zeta \in [0, \infty)$. Then, P has NE $q = \frac{\alpha}{\zeta}$.*

Proof Since P has ME $\alpha \in [0, \infty)$, we find for some $t > 0$ that

$$\frac{\Delta_\eta^\zeta(x)}{c_{\mathrm{LC}}} \leq |2\eta(x) - 1| < t, \qquad x \in X,$$

and we follow that $\Delta_\eta(x) \leq (c_{\mathrm{LC}}t)^{\frac{1}{\zeta}}$. Consequently, the definition of the noise exponent yields

$$\begin{aligned} P_X\left(\{x \in X : |2\eta(x) - 1| < t\}\right) &\leq P_X\left(\{x \in X : \Delta_\eta(x) \leq (c_{\mathrm{LC}}t)^{\frac{1}{\zeta}}\}\right) \\ &\leq c_{\mathrm{ME}}^\alpha (c_{\mathrm{LC}}t)^{\frac{\alpha}{\zeta}}. \end{aligned}$$

∎

Remark 25 *i) One can show by using similar arguments as in (Blaschzyk and Steinwart, 2018, Lemma 2.1) together with (Steinwart, 2015, Lemma A.10.4(i)) that there exists a $\delta^* > 0$ such that the lower bound*

$$\lambda^d(\{x \in X | \Delta_\eta(x) \leq \delta\}) \geq c_d \cdot \delta \qquad \text{for all } \delta \in (0, \delta^*]$$

and some $c_d > 0$ is satisfied.

ii) Assume that η is Hölder-smooth with exponent ρ, that P has NE q and that P_X has a density w.r.t. the Lebesgue measure that is bounded away from zero. Then, part i) together with (Blaschzyk and Steinwart, 2018, Lemma A.2) yields

$$ct^{\frac{1}{\rho}} \leq P_X(\{\Delta_\eta(x) \leq t^{\frac{1}{\rho}}\}) \leq P_X\left(\{x \in X : |2\eta(x) - 1| < t\}\right) \leq c_{\mathrm{NE}} t^q$$

for some constant $c > 0$. Thus, $\rho q > 1$ can never be satisfied.

Appendix B.

In this appendix we state some technical lemmata.

Lemma 26 (Number of cells) *Let assumptions **(A)** and **(G)** be satisfied. Let $t \geq r$ such that $t + r \leq \delta^*$, where $\delta^* > 0$ is the constant from (9), and define*

$$J := \{\, j \in J \,|\, \forall\, x \in A_j : \Delta_\eta(x) \leq t \,\}.$$

Then, there exists a constant $c_d > 0$ such that

$$|J| \leq c_d \cdot t r^{-d}.$$

Proof We define $T := \bigcup_{j \in J} A_j$ and $\tilde{T} := \bigcup_{j \in J} B_r(z_j)$. Obviously, $T \subset \tilde{T}$ since $A_j \subset B_r(z_j)$ for all $j \in J$. Furthermore, we have for all $x \in \tilde{T}$ that $\Delta_\eta(x) \leq \tilde{t}$, where $\tilde{t} := t + r$. Then, we obtain with (Blaschzyk and Steinwart, 2018, Lemma 2.1) that

$$\lambda^d(\tilde{T}) \leq \lambda^d\left(\left\{\Delta(x) \leq \tilde{t}\right\}\right) \leq 4\mathcal{H}^{d-1}(X_0) \cdot \tilde{t}. \tag{74}$$

Moreover,

$$\lambda^d(\tilde{T}) = \lambda^d\left(\bigcup_{j \in J} B_r(z_j)\right) \geq \lambda^d\left(\bigcup_{j \in J} B_{\frac{r}{4}}(z_j)\right) = |J| \lambda^d\left(B_{\frac{r}{4}}(z)\right) = |J| \left(\frac{r}{4}\right)^d \lambda^d(B), \tag{75}$$

since $B_{\frac{r}{4}}(z_i) \cap B_{\frac{r}{4}}(z_j) = \emptyset$ for $i \neq j$. To see the latter, assume that we have an $x \in B_{\frac{r}{4}}(z_i) \cap B_{\frac{r}{4}}(z_j)$. But then, $\|z_i - z_j\|_2 \leq \|x - z_j\|_2 + \|x - z_i\|_2 \leq \frac{r}{4} + \frac{r}{4} \leq \frac{r}{2}$, which is not true, since we assumed $\|z_i - z_j\|_2 > \frac{r}{2}$ for all $i \neq j$. Hence, the balls with radius $\frac{r}{4}$ are disjoint. Finally, by (74) together with (75) and $t \geq r$ we find

$$|J| \leq \frac{4^d \lambda^d(\tilde{T})}{r^d \lambda^d(B)} \leq \frac{2^{2d+2} \mathcal{H}^{d-1}(\{x \in X | \eta = 1/2\}) \cdot \tilde{t}}{r^d \lambda^d(B)} \leq \frac{2^{2d+3} \mathcal{H}^{d-1}(\{x \in X | \eta = 1/2\}) \cdot t}{r^d \lambda^d(B)}.$$

■

Lemma 27 *Let $X \subset \mathbb{R}^d$ and $\gamma, \rho > 0$. Then, we have*

$$\left(\frac{2}{\pi\gamma^2}\right)^{d/2} \int_{B_\rho(x)} e^{-2\gamma^{-2}\|x-y\|_2^2}\, dy = \frac{1}{\Gamma(d/2)} \int_0^{2\rho^2\gamma^{-2}} e^{-t} t^{d/2-1}\, dt.$$

Proof For $\rho > 0$ we find that

$$\begin{aligned}
&\left(\frac{2}{\pi\gamma^2}\right)^{d/2} \int_{B_\rho(x)} e^{-2\gamma^{-2}\|x-y\|_2^2}\, dy \\
&= \left(\frac{2}{\pi\gamma^2}\right)^{d/2} \int_{B_\rho(0)} e^{-2\gamma^{-2}\|y\|_2^2}\, dy \\
&= \left(\frac{2}{\pi\gamma^2}\right)^{d/2} \frac{\pi^{d/2}}{\Gamma(d/2+1)} \int_0^\rho e^{-2\gamma^{-2}t^2} d \cdot t^{d-1}\, dt \\
&= \left(\frac{2}{\pi\gamma^2}\right)^{d/2} \frac{2\pi^{d/2}}{d\Gamma(d/2)} \int_0^{\sqrt{2}\rho\gamma^{-1}} e^{-t^2} d \cdot t^{d-1} \cdot \frac{1}{\sqrt{2}\gamma^{-1}} \left(\frac{\gamma}{\sqrt{2}}\right)^{d-1} dt \\
&= \frac{2}{\Gamma(d/2)} \int_0^{2\rho^2\gamma^{-2}} e^{-t} t^{d/2-1} \cdot \frac{1}{2}\, dt \\
&= \frac{1}{\Gamma(d/2)} \int_0^{2\rho^2\gamma^{-2}} e^{-t} t^{d/2-1}\, dt.
\end{aligned}$$

■

Lemma 28 *Let $(A_j)_{j=1,\ldots,m}$ be a partition of $B_{\ell_2^d}$. Let $d \geq 1, p \in (0, \frac{1}{2})$ and let $r_n \in (0,1]$. For $\rho_n \leq n^{-2}$ and $\delta_n \leq n^{-1}$ fix a finite ρ_n-net $\Lambda_n \subset (0, n^{-1}]$ and a finite $\delta_n r_n$-net $\Gamma_n \subset (0, r_n]$. Let $J \subset \{1, \ldots, m_n\}$ be an index set and for all $j \in J$ let $\gamma_j \in (0, r_n]$, $\lambda_j > 0$. Define $\gamma_{\max} := \max_{j\in J}$, as well as analogously $\gamma_{\min}$.*

i) Let $\beta \in (0,1], q \in [0,\infty)$ and let $|J| \leq c_d r_n^{-d+1}$ for some constant $c_d > 0$. Then, for all $\varepsilon_1 > 0$ there exists a constant $\tilde{c}_1 > 0$ such that

$$\begin{aligned}
&\inf_{(\boldsymbol{\lambda},\boldsymbol{\gamma})\in(\Lambda_n\times\Gamma_n)^{|J|}} \left(\sum_{j\in J} \frac{\lambda_j r_n^d}{\gamma_j^d} + \gamma_{\max}^\beta \right. \\
&\qquad \left. + \left(\frac{r_n}{n}\right)^{\frac{q+1}{q+2-p}} \left(\sum_{j\in J} \lambda_j^{-1} \gamma_j^{-\frac{d}{p}} P_X(A_j) \right)^{\frac{p(q+1)}{q+2-p}} \right) \\
&\leq \tilde{c}_1 \cdot n^{-\beta\kappa(\nu+1)+\varepsilon_1}.
\end{aligned}$$

ii) Let $\beta \in (0,1], q \in [0,\infty)$ and let $|J| \leq c_d r_n^{-d+1}$ for some constant $c_d > 0$. Then, for all $\tilde{\varepsilon}, \varepsilon_2 > 0$ there exists a constant $\tilde{c}_2 > 0$ such that

$$\begin{aligned}\inf_{(\boldsymbol{\lambda},\boldsymbol{\gamma})\in(\Lambda_n\times\Gamma_n)^{|J|}} \Bigg(& \left(\frac{r_n}{\gamma_{\min}}\right)^d \sum_{j\in J} \lambda_j n^{\tilde{\varepsilon}} \\ & + \left(\frac{r_n}{n}\right)^{\frac{q+1}{q+2-p}} \left(\sum_{j\in J} \lambda_j^{-1} \gamma_j^{-\frac{d}{p}} P_X(A_j)\right)^{\frac{p(q+1)}{q+2-p}} \Bigg) \\ & \leq \tilde{c}_2 \cdot n^{\varepsilon_2} \left(r_n^{d-1} n\right)^{-\frac{q+1}{q+2}}.\end{aligned}$$

iii) Let $|J| \leq c_d r_n^{-d}$. Then, for all $\tilde{\varepsilon}, \varepsilon_3 > 0$ there exists a constant $\tilde{c}_3 > 0$ such that

$$\begin{aligned}&\inf_{(\boldsymbol{\lambda},\boldsymbol{\gamma})\in(\Lambda_n\times\Gamma_n)^{|J|}} \left(\left(\frac{r_n}{\gamma_{\min}}\right)^d \sum_{j\in J} \lambda_j n^{\tilde{\varepsilon}} + \left(\sum_{j\in J} \lambda_j^{-1} \gamma_j^{-\frac{d}{p}} P_X(A_j)\right)^p n^{-1} \right) \\ &\leq \tilde{c}_3 \cdot r_n^{-d} n^{-1+\varepsilon_3}.\end{aligned}$$

Proof We follow the lines of the proof of (Meister and Steinwart, 2016, Lemma 14). Let us assume that $\Lambda_n := \{\lambda^{(1)}, \dots \lambda^{(u)}\}$ and $\Gamma_n := \{\gamma^{(1)}, \dots \gamma^{(v)}\}$ such that $\lambda^{(i-1)} < \lambda^{(i)}$ and $\gamma^{(l-1)} < \gamma^{(l)}$ for all $i = 2, \dots, u$ and $l = 2, \dots, v$. Furthermore, let $\gamma^{(0)} = \lambda^{(0)} := 0$ and $\lambda^{(u)} := n^{-1}, \gamma^{(v)} := r_n$. Then, fix a pair $(\lambda^*, \gamma^*) \in [0, n^{-1}] \times [0, r_n]$. Following the lines of the proof of (Steinwart and Christmann, 2008, Lemma 6.30) there exist indices $i \in \{1, \dots, u\}$ and $l \in \{1, \dots, v\}$ such that

$$\begin{aligned}\lambda^* &\leq \lambda^{(i)} \leq \lambda^* + 2\rho_n, \\ \gamma^* &\leq \gamma^{(l)} \leq \gamma^* + 2\delta_n r_n.\end{aligned} \tag{76}$$

i) With (76) we find

$$\begin{aligned}&\inf_{(\boldsymbol{\lambda},\boldsymbol{\gamma})\in(\Lambda_n\times\Gamma_n)^{|J|}} \left(\sum_{j\in J} \frac{\lambda_j r_n^d}{\gamma_j^d} + \gamma_{\max}^\beta + \left(\frac{r_n}{n}\right)^{\frac{q+1}{q+2-p}} \left(\sum_{j\in J} \lambda_j^{-1} \gamma_j^{-\frac{d}{p}} P_X(A_j)\right)^{\frac{p(q+1)}{q+2-p}} \right) \\ &\leq \sum_{j\in J} \frac{\lambda^{(i)} r_n^d}{\left(\gamma^{(l)}\right)^d} + \left(\gamma^{(l)}\right)^\beta + \left(\frac{r}{n}\right)^{\frac{q+1}{q+2-p}} \left(\sum_{j\in J} \left(\lambda^{(i)}\right)^{-1} \left(\gamma^{(l)}\right)^{-\frac{d}{p}} P_X(A_j)\right)^{\frac{p(q+1)}{q+2-p}} \\ &\leq |J| \frac{\lambda^{(i)} r_n^d}{\left(\gamma^{(l)}\right)^d} + \left(\gamma^{(l)}\right)^\beta + \left(\frac{r_n \left(\lambda^{(i)}\right)^{-p} \left(\gamma^{(l)}\right)^{-d}}{n}\right)^{\frac{q+1}{q+2-p}} \left(\sum_{j\in J} P_X(A_j)\right)^{\frac{p(q+1)}{q+2-p}} \\ &\leq \frac{(\lambda^* + 2\rho_n) r_n}{(\gamma^*)^d} + (\gamma^* + 2\delta_n r_n)^\beta + \left(\frac{r_n (\lambda^*)^{-p}}{(\gamma^*)^d n}\right)^{\frac{q+1}{q+2-p}} \\ &\leq c_1 \left(\lambda^* r_n (\gamma^*)^{-d} + (\gamma^*)^\beta + \left(\frac{r_n (\lambda^*)^{-p}}{(\gamma^*)^d n}\right)^{\frac{q+1}{q+2-p}} + \rho_n r_n (\gamma^*)^{-d} + (\delta_n r_n)^\beta \right)\end{aligned}$$

for some $c_1 > 0$. We define $\lambda^* := n^{-\sigma}$ for some $\sigma \in [1,2]$ and $\gamma^* := r_n^{\kappa} n^{-\kappa}$. Obviously, $\lambda^* \in [0, n^{-1}]$. Moreover, we have $\gamma^* \in [0, r_n]$ since $\nu \leq \frac{\kappa}{1-\kappa}$. Then, we obtain with $\rho_n \leq n^{-2}$ and $\delta_n \leq n^{-1}$, and together with $1 \geq \kappa(\beta+d)(\nu+1) - \nu > 0$ and $\frac{(1-d\kappa)(q+1)}{q+2-p} > \frac{(1-d\kappa)(q+1)}{q+2} = \left(\frac{\beta(q+2)}{\beta(q+2)+d(q+1)}\right)\frac{q+1}{q+2} = \beta\kappa$ that

$$
\begin{aligned}
&c_1 \left(\lambda^* r_n (\gamma^*)^{-d} + (\gamma^*)^{\beta} + \left(\frac{r_n (\lambda^*)^{-p}}{(\gamma^*)^d n} \right)^{\frac{q+1}{q+2-p}} + \rho_n r_n (\gamma^*)^{-d} + (\delta_n r_n)^{\beta} \right) \\
&\leq c_1 \left(n^{-\sigma} r_n r_n^{-d\kappa} n^{d\kappa} + r_n^{\beta\kappa} n^{-\beta\kappa} + \left(\frac{r_n^{1-d\kappa} (\lambda^*)^{-p}}{n^{1-d\kappa}} \right)^{\frac{q+1}{q+2-p}} + n^{-2} r_n (\gamma^*)^{-d} + \left(r_n n^{-1}\right)^{\beta} \right) \\
&\leq c_2 \left(r_n^{-1+(\beta+d)\kappa} n^{-(\beta+d)\kappa} r_n r_n^{-d\kappa} n^{d\kappa} + r_n^{\beta\kappa} n^{-\beta\kappa} + \left(r_n n^{-1}\right)^{\frac{(1-d\kappa)(q+1)}{q+2-p}} n^{\frac{p\sigma(q+1)}{q+2-p}} + \left(r_n n^{-1}\right)^{\beta} \right) \\
&\leq c_2 \left(r_n^{\beta\kappa} n^{-\beta\kappa} + r_n^{\beta\kappa} n^{-\beta\kappa} n^{\varepsilon_1} + \left(r_n n^{-1}\right)^{\beta} \right) \\
&\leq c_3 \cdot n^{-\beta\kappa(\nu+1)+\varepsilon_1}
\end{aligned}
$$

holds for some constants $c_2, c_3 > 0$ and where p is chosen sufficiently small such that $\varepsilon_1 \geq \frac{p\sigma(q+1)}{q+2-p} > 0$.

ii) With (76) we find

$$
\begin{aligned}
&\inf_{(\boldsymbol{\lambda},\boldsymbol{\gamma}) \in (\Lambda_n \times \Gamma_n)^{|J|}} \left(\left(\frac{r_n}{\gamma_{\min}}\right)^d \sum_{j \in J} \lambda_j n^{\tilde{\varepsilon}} + \left(\frac{r_n}{n}\right)^{\frac{q+1}{q+2-p}} \left(\sum_{j \in J} \lambda_j^{-1} \gamma_j^{-\frac{d}{p}} P_X(A_j) \right)^{\frac{p(q+1)}{q+2-p}} \right) \\
&\leq \left(\frac{r_n}{\gamma^{(l)}}\right)^d \sum_{j \in J} \lambda^{(i)} n^{\tilde{\varepsilon}} + \left(\frac{r_n}{n}\right)^{\frac{q+1}{q+2-p}} \left(\sum_{j \in J} \left(\lambda^{(i)}\right)^{-1} \left(\gamma^{(l)}\right)^{-\frac{d}{p}} P_X(A_j) \right)^{\frac{p(q+1)}{q+2-p}} \\
&\leq \left(\frac{r_n}{\gamma^{(l)}}\right)^d |J| \lambda^{(i)} n^{\tilde{\varepsilon}} + \left(\frac{r_n}{\left(\gamma^{(l)}\right)^d n} \right)^{\frac{q+1}{q+2-p}} \left(\lambda^{(i)}\right)^{-\frac{p(q+1)}{q+2-p}} \left(\sum_{j \in J} P_X(A_j) \right)^{\frac{p(q+1)}{q+2-p}} \\
&\leq c_4 \frac{r_n \lambda^{(i)} n^{\tilde{\varepsilon}}}{\left(\gamma^{(l)}\right)^d} + \left(\frac{r_n}{\left(\gamma^{(l)}\right)^d n} \right)^{\frac{q+1}{q+2-p}} \left(\lambda^{(i)}\right)^{-\frac{p(q+1)}{q+2-p}} \\
&\leq c_4 \frac{r_n (\lambda^* + 2\rho_n) n^{\tilde{\varepsilon}}}{(\gamma^*)^d} + \left(\frac{r_n}{(\gamma^*)^d n} \right)^{\frac{q+1}{q+2-p}} (\lambda^*)^{-\frac{p(q+1)}{q+2-p}} \\
&\leq c_4 \frac{r_n \lambda^* n^{\tilde{\varepsilon}}}{(\gamma^*)^d} + \left(\frac{r_n}{(\gamma^*)^d n} \right)^{\frac{q+1}{q+2-p}} (\lambda^*)^{-\frac{p(q+1)}{q+2-p}} + 2c_4 \frac{\rho_n r_n n^{\tilde{\varepsilon}}}{(\gamma^*)^d}
\end{aligned}
$$

for some constant $c_4 > 0$ depending on d. We define $\gamma^* := r_n$ and $\lambda^* := n^{-\sigma}$ for some $\sigma \in [1,2]$. Then, we obtain with $\rho_n \leq n^{-2}$ that

$$\begin{aligned}
&c_4 \frac{r_n \lambda^* n^{\tilde{\varepsilon}}}{(\gamma^*)^d} + \left(\frac{r_n}{(\gamma^*)^d n}\right)^{\frac{q+1}{q+2-p}} (\lambda^*)^{-\frac{p(q+1)}{q+2-p}} + 2c_4 \frac{r_n \rho_n n^{\tilde{\varepsilon}}}{(\gamma^*)^d} \\
&= c_4 \frac{n^{-\sigma} n^{\tilde{\varepsilon}}}{r_n^{d-1}} + \left(\frac{1}{r_n^{d-1} n}\right)^{\frac{q+1}{q+2-p}} n^{\frac{p\sigma(q+1)}{q+2-p}} + 2c_4 \frac{\rho_n n^{\tilde{\varepsilon}}}{r_n^{d-1}} \\
&\leq c_4 \frac{n^{-1} n^{\tilde{\varepsilon}}}{r_n^{d-1}} + \left(r_n^{d-1} n\right)^{-\frac{q+1}{q+2}} n^{\hat{\varepsilon}} + 2c_4 \frac{\rho_n n^{\tilde{\varepsilon}}}{r_n^{d-1}} \\
&\leq c_5 n^{\varepsilon_2} \left(\left(r_n^{d-1} n\right)^{-1} + \left(r_n^{d-1} n\right)^{-\frac{q+1}{q+2}} + n^{-2} \left(r_n^{d-1}\right)^{-1} \right) \\
&\leq c_6 n^{\varepsilon_2} \left(r_n^{d-1} n\right)^{-\frac{q+1}{q+2}},
\end{aligned}$$

where $c_5, c_6 > 0$ are constants depending on d and where p is chosen sufficiently small such that $\hat{\varepsilon} \geq \frac{p\sigma(q+1)}{q+2-p}$ and $\varepsilon_2 := \max\{\tilde{\varepsilon}, \hat{\varepsilon}\}$.

iii) We find with (76) that

$$\begin{aligned}
&\inf_{(\boldsymbol{\lambda},\boldsymbol{\gamma}) \in (\Lambda_n \times \Gamma_n)^{|J|}} \left(n^{\tilde{\varepsilon}} \left(\frac{r_n}{\gamma_{\min}}\right)^d \sum_{j \in J} \lambda_j + \left(\sum_{j \in J} \lambda_j^{-1} \gamma_j^{-\frac{d}{p}} P_X(A_j) \right)^p n^{-1} \right) \\
&\leq n^{\tilde{\varepsilon}} \left(\frac{r_n}{\gamma_{\min}}\right)^d \sum_{j \in J} \lambda^{(i)} + \left(\sum_{j \in J} \left(\lambda^{(i)}\right)^{-1} \left(\gamma^{(l)}\right)^{-\frac{d}{p}} P_X(A_j) \right)^p n^{-1} \\
&\leq n^{\tilde{\varepsilon}} \left(\frac{r_n}{\gamma_{\min}}\right)^d |J| \lambda^{(i)} + \left(\lambda^{(i)}\right)^{-p} \left(\gamma^{(l)}\right)^{-d} \left(\sum_{j \in J} P_X(A_j) \right)^p n^{-1} \\
&\leq c_7 n^{\tilde{\varepsilon}} \left(\gamma^{(l)}\right)^{-d} \lambda^{(i)} + \left(\lambda^{(i)}\right)^{-p} \left(\gamma^{(l)}\right)^{-d} n^{-1} \\
&\leq c_7 n^{\tilde{\varepsilon}} (\gamma^*)^{-d} (\lambda^* + 2\rho_n) + (\lambda^*)^{-p} (\gamma^*)^{-d} n^{-1} \\
&= c_7 n^{\tilde{\varepsilon}} (\gamma^*)^{-d} \lambda^* + (\lambda^*)^{-p} (\gamma^*)^{-d} n^{-1} + 2\rho_n c_7 n^{\tilde{\varepsilon}} (\gamma^*)^{-d}
\end{aligned}$$

holds for some constant $c_7 > 0$ depending on d. We define $\gamma^* := r_n$ and $\lambda^* := n^{-\sigma}$ for some $\sigma \in [1,2]$. Then, we obtain with $\rho_n \leq n^{-2}$

$$\begin{aligned}
&c_7 n^{\tilde{\varepsilon}} (\gamma^*)^{-d} \lambda^* + (\lambda^*)^{-p} (\gamma^*)^{-d} n^{-1} + 2\rho_n c_7 n^{\tilde{\varepsilon}} (\gamma^*)^{-d} \\
&\leq c_7 n^{\tilde{\varepsilon}} r_n^{-d} n^{-\sigma} + n^{p\sigma} r_n^{-d} n^{-1} + 2c_7 n^{\tilde{\varepsilon}} r_n^{-d} n^{-2} \\
&\leq c_7 n^{\tilde{\varepsilon}} r_n^{-d} n^{-1} + n^{\hat{\varepsilon}} r_n^{-d} n^{-1} + 2c_7 n^{\tilde{\varepsilon}} r_n^{-d} n^{-2} \\
&\leq c_8 \cdot r_n^{-d} n^{-1+\varepsilon_3}
\end{aligned}$$

for some constant $c_8 > 0$ depending on d and where p is chosen sufficiently small such that $\hat{\varepsilon} \geq p\sigma > 0$. Here, $\varepsilon_3 := \max\{\tilde{\varepsilon}, \hat{\varepsilon}\}$.

■

References

J.-Y. Audibert and A. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35:608–633, 2007.

M. Belkin, D. J. Hsu, and P. Mitra. Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. *Advances in Neural Information Processing Systems (NIPS)*, 31:2306–2317, 2018.

K. P. Bennett and J. A. Blue. A support vector machine approach to decision trees. *IEEE International Joint Conference on Neural Networks Proceedings*, 3:2396–2401, 1998.

P. Binev, A. Cohen, W. Dahmen, and R. DeVore. Classification algorithms using adaptive partitioning. *Ann. Statist.*, 42:2141–2163, 2014.

I. Blaschzyk and I. Steinwart. Improved Classification Rates under Refined Margin Conditions. *Electron. J. Stat.*, 12:793–823, 2018.

I. K. Blaschzyk. *Improved Classification Rates for Localized Algorithms under Margin Conditions*. Springer Spektrum, 2020.

L. Bottou and V. Vapnik. Local learning algorithms. *Neural Comput.*, 4:888–900, 1992.

B. Carl and I. Stephani. *Entropy, Compactness and the Approximation of Operators*. Cambridge University Press, Cambridge, 1990.

H. Cheng, P.-N. Tan, and R. Jin. Localized support vector machine and its efficient algorithm. In *SIAM International Conference on Data Mining*, pages 461–466, 2007.

M. Döring, L. Györfi, and H. Walk. *Exact rate of convergence of kernel-based classification rule*, volume 605, pages 71–91. Springer International Publishing, 7 2015.

F. Dumpert and A. Christmann. Universal consistency and robustness of localized support vector machines. *Neurocomputing*, 315:96–106, 2018.

M. Eberts and I. Steinwart. Optimal regression rates for SVMs using Gaussian kernels. *Electron. J. Statist.*, 7:1–42, 2013.

M. Eberts and I. Steinwart. Optimal Learning Rates for Localized SVMs. *ArXiv e-prints 1507.06615*, 2015.

M. Farooq and I. Steinwart. Learning rates for kernel-based expectile regression. *Mach. Learn.*, 108(2):203–227, 2019.

H. Federer. *Geometric Measure Theory*. Springer, Berlin, 1969.

M. Fernandez-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *J. Mach. Learn. Res.*, 15: 3133–3181, 2014.

R. Hable. Universal consistency of localized versions of regularized kernel methods. *J. Mach. Learn. Res.*, 14(1):153–186, 2013.

T. Hamm and I. Steinwart. Adaptive Learning Rates for Support Vector Machines Working on Data with Low Intrinsic Dimension. *Ann. Statist.*, 49:3153–3180, 2021.

G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-normalizing neural networks. *Advances in Neural Information Processing Systems (NIPS)*, 30:971–980, 2017.

M. Kohler and A. Krzyzak. On the rate of convergence of local averaging plug-in classification rules under a margin condition. *IEEE Trans. Inf. Theor.*, 53(5):1735–1742, 2007.

J. Lin and L. Rosasco. Optimal Rates for Multi-pass Stochastic Gradient Methods. *Journal of Machine Learning Research*, 18(97):1–47, 2017.

S. Lin, X. Guo, and D.-X. Zhou. Distributed Learning with Regularized Least Squares. *J. Mach. Learn. Res.*, 18(92):1–31, 2017a.

S. Lin, J. Zeng, and X. Chang. Learning rates for classification with Gaussian kernels. *Neural Comput.*, 29(12):3353–3380, 2017b.

E. Mammen and A. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 12 1999.

M. Meister and I. Steinwart. Optimal Learning Rates for Localized SVMs. *J. Mach. Learn. Res.*, 17(194):1–44, 2016.

N. Mücke. Reducing training time by efficient localized kernel regression. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 89:2603–2610, 2019.

N. Mücke and G. Blanchard. Parallelizing spectrally regularized kernel algorithms. *J. Mach. Learn. Res.*, 19(1):1069–1097, 2018.

L. Pillaud-Vivien, A. Rudi, and F. Bach. Statistical Optimality of Stochastic Gradient Descent on Hard Learning Problems through Multiple Passes. *Advances in Neural Information Processing Systems (NIPS)*, 31:8114–8124, 2018.

A. Rahimi and B. Recht. Random Features for Large-Scale Kernel Machines. *Advances in Neural Information Processing Systems (NIPS)*, 20:1177–1184, 2008.

A. Rudi and L. Rosasco. Generalization Properties of Learning with Random Features. *Advances in Neural Information Processing Systems (NIPS)*, 30:3215–3225, 2017.

A. Rudi, R. Camoriano, and L. Rosasco. Less is More: Nyström Computational Regularization. *Advances in Neural Information Processing Systems (NIPS)*, 28:1657–1665, 2015.

I. Steinwart. Fully adaptive density-based clustering. *Ann. Statist.*, 43:2132–2167, 2015.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.

I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.*, 35(2):575–607, 04 2007.

I. Steinwart and P. Thomann. liquidSVM: A fast and versatile SVM package. *ArXiv e-prints 1702.06899*, February 2017.

P. Thomann, I. Blaschzyk, M. Meister, and I. Steinwart. Spatial Decompositions for Large Scale SVMs. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 54:1329–1337, 2017.

C. K. I. Williams and M. Seeger. Using the Nyström Method to Speed Up Kernel Machines. *Advances in Neural Information Processing Systems (NIPS)*, 13:682–688, 2001.

H. Zhang, A. C. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, 2:2126–2136, 2006.

Y. Zhang, J. Duchi, and M. Wainwright. Divide and Conquer Kernel Ridge Regression: A Distributed Algorithm with Minimax Optimal Rates. *J. Mach. Learn. Res.*, 16(102): 3299–3340, 2015.